Latent State-Predictive Exploration for Deep Reinforcement Learning

Yiming Wang[1], Kaiyan Zhao[2], Borong Zhang[1], Yan Li[3], 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]Undergraduate School of Artificial Intelligence, Shenzhen Polytechnic University, Shenzhen, China
{wang.yiming,mc45324,yb57411}@connect.um.edu.mo, zhao.kaiyan@whu.edu.cn, ryanlhu@um.edu.mo

Abstract

Reinforcement learning (RL) has achieved promising results in continuous control tasks, where efficient exploration of the state space is crucial for success. However, many recent RL approaches still struggle with sample inefficiency and insufficient exploration for long-horizon tasks, particularly in environments characterized by high-dimensional and complex state spaces. To address these challenges, we propose a novel exploration framework, Latent State Predictive Exploration (LSPE). The core idea behind LSPE is to endow the agent with a form of "foresight" to enhance exploration in long-horizon settings. Specifically, LSPE employs a state encoder to learn compact latent representations from high-dimensional visual observations, effectively filtering out irrelevant or noisy information. To further enrich and stabilize these representations, we incorporate a diffusion-based self-predictive module that enforces temporal consistency by predicting future states, thereby improving both exploration and downstream predictive control. Additionally, we introduce an Exploration Reward Function (ERF) that explicitly encourages the agent to visit novel latent states. This reward signal promotes more efficient and scalable exploration in complex environments. We evaluate LSPE across a diverse set of challenging long-horizon navigation and manipulation tasks, spanning simulation environments such as Habitat and Robosuite, as well as deployment on a real robot in a *physical indoor environment*. Experimental results show that LSPE substantially enhances exploration efficiency and scales effectively to complex, high-dimensional tasks.

1 Introduction

Reinforcement learning (RL) has achieved remarkable success in continuous control domains such as locomotion, manipulation, and navigation. However, these achievements often rely on well-shaped reward signals and relatively short-horizon tasks. In more realistic scenarios, where agents must make decisions over extended time scales and receive sparse or delayed feedback, standard RL methods struggle to scale effectively. These *long-horizon* tasks introduce fundamental challenges, particularly in terms of sample efficiency, learning effectiveness and most critically, exploration. A central bottleneck in long-horizon RL is the agent's ability to explore complex, high-dimensional state spaces effectively, especially when observations are visual and redundant. Naive or short-sighted exploration strategies often fail, as rewards may only be attained after long sequences of purposeful actions. As illustrated in Fig.1, effective exploration is crucial for solving such tasks. Consider a navigation task in a realistic indoor environment, where the agent must reach a distant goal based solely on high-dimensional visual observations. We compare our method to two representative baselines: the exploration bonus–based method RND (Burda et al. 2019a) and the action noise–based method NoisyNet (Fortunato et al. 2018). While both baselines exhibit limited coverage and tend to get stuck in local regions, our method drives more thorough exploration, leading to significantly higher success rates.

(a) (b)

Figure 1: (a) The navigation task within the realistic indoor scene with visual observation. The goal for the agent is to navigate to the designated position. (b) Comparison of task success rates and explored rate (state space coverage) among our method (LSPE), the exploration bonus–based baseline RND, and the action noise–based baseline NoisyNet.

Classical exploration strategies such as ϵ-greedy or Boltzmann exploration select random actions with non-zero probability (Mnih et al. 2015), but these are inadequate for deep exploration where agents must discover distant rewards. More recent approaches introduce intrinsic rewards or exploration bonuses to promote novelty-seeking behavior. For instance, Bellemare et al. (Bellemare et al. 2016a) use pseudo-counts to reward infrequently visited states. Curiosity-based methods (Pathak et al. 2017a; Burda et al. 2019b; Pathak et al. 2017b; Burda et al. 2019a) lever-

*Corresponding author.

age prediction error in learned dynamics models to define exploration bonuses. Others (Raileanu and Rocktäschel 2020; Zhang et al. 2021b; Wang et al. 2023) employ state-difference metrics as intrinsic motivation signals. While these techniques have achieved success in specific settings, they often suffer from poor scalability in visually complex, high-dimensional environments. Methods based on action noise (Fortunato et al. 2018) or logic-guided policies (Zhang et al. 2024) attempt to diversify behavior but still fall short in long-horizon control tasks. Furthermore, pixel-level novelty signals are noisy and unstable, making common intrinsic rewards such as curiosity or prediction error unreliable in high-dimensional visual tasks.

To address these limitations, we propose **Latent State Predictive Exploration (LSPE)**[1], a novel framework that equips agents with *latent foresight* to improve exploration in long-horizon tasks. LSPE is composed of three key components: (1) a compact state encoder that extracts informative latent features from high-dimensional visual observations, (2) a diffusion-based self-predictive module that models long-term temporal dynamics by forecasting future states, and (3) a latent-space exploration reward function (ERF) that drives the agent toward novel and uncertain directions in the latent space. By jointly leveraging these components, LSPE enables temporal-aware representation learning, structured exploration, and improved sample efficiency in complex, long-horizon environments.

Our main contributions can be summarized as follows:

- We propose LSPE, a novel framework that equips agents with *latent foresight* and introduces a directional-variance exploration reward to facilitate deep and efficient exploration in long-horizon tasks.
- We develop a temporal-aware bisimulation-based representation learning approach by integrating temporally consistent predictions from a diffusion-based self-predictive module, resulting in more robust and stable latent representations.
- We conduct extensive experiments on challenging long-horizon navigation and manipulation tasks across Habitat, Robosuite, and **real-world** indoor environments. LSPE consistently outperforms other baselines in terms of learning efficiency, task success rate, and scalability.

2 Preliminaries

Reinforcement Learning. We model the decision-making environment as 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 the transition probability from state $s \in \mathcal{S}$ to $s' \in \mathcal{S}$ under action $a \in \mathcal{A}$, r is the reward function, and $\gamma \in [0, 1)$ is the discount factor. In long-horizon tasks, the agent interacts with the environment over extended trajectories $(s_0, a_0, s_1, a_1, \ldots)$, where meaningful rewards may occur only after a long sequence of actions. The agent's objective is to learn a policy $\pi : \mathcal{S} \to \mathcal{A}$ that maximizes the expected cumulative discounted return: $\mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)]$. This is typically achieved by learning a value function $V^\pi(s) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s\right]$, which estimates the utility of each state under policy π.

Latent Exploration Reward. To tackle the exploration challenge in long-horizon, sparse-reward settings, we augment the task reward r with an exploration reward F, resulting in a combined reward:

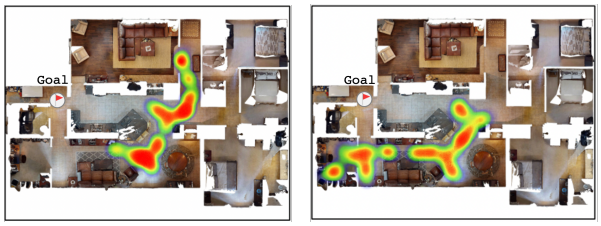
$$r' = r + F \tag{1}$$

Prior approaches define F using pseudo-counts (Bellemare et al. 2016a), prediction error (Burda et al. 2019a), or state differences (Wang et al. 2024). However, these methods often fail in high-dimensional visual environments due to noisy and ambiguous signals in observation space.

To address this, we introduce a state encoder $\phi : \mathcal{S} \to \mathcal{Z}$ that maps raw observations to a compact latent space $\mathcal{Z}$. This latent space filters out irrelevant details and supports scalable, semantically grounded exploration. We then define the exploration reward directly in $\mathcal{Z}$, guiding the agent to seek novel and informative latent states. The specific formulation is described next.

3 Problem Statement

Exploration in long-horizon reinforcement learning (RL) remains a fundamental challenge, especially in realistic environments with *sparse rewards* and *high-dimensional observations*. To motivate our proposed framework, we begin with a concrete example that illustrates the limitations of standard exploration techniques.



(a) (b)

Figure 2: Trajectory heatmaps of policies trained with classical exploration strategies in a high-dimensional indoor navigation task: (a) Action noise (NoisyNet) (Fortunato et al. 2018), (b) Exploration reward (RND) (Burda et al. 2019a).

Illustrative Example. Consider a long-horizon navigation task, such as the realistic indoor scenario shown in Fig. 1. The agent navigate to a designated goal position and state is observed as an RGB image captured by the agent's front-facing camera. A sparse reward of $r = 1$ is given only upon successful task completion; otherwise, the agent receives $r = 0$. This setup exemplifies the *sparse reward* problem, where informative feedback is rare, making it hard for the agent to learn effective policy through random exploration.

Challenge: Ineffective Exploration in Realistic Settings. We evaluate two representative baselines: NoisyNet (Fortunato et al. 2018), which adds stochasticity to the policy, and RND (Burda et al. 2019a), which uses prediction error as intrinsic reward. As shown in Fig. 2, both methods fail to reach

[1]Project: https://sites.google.com/view/lspe-robotic-control/

the goal, with trajectories confined to a small region of the room. This failure stems from two compounding challenges:

- **High-dimensional observations** introduce redundancy and noise, expanding the exploration space and weakens the agent's ability to detect novel states.
- **Sparse rewards** provide little guidance, making undirected exploration (e.g., via random action noise) ineffective in discovering meaningful trajectories.

So the need for exploration strategies can be summarized as:

(N1): Abstract away irrelevant visual details through compact state encoding.

(N2): Reason about long-term novelty and uncertainty to guide exploration more effectively.

Our method addresses both needs by integrating temporally consistent latent representations (sec 4.1 and sec 4.2) and directional uncertainty-aware exploration (sec 4.3).

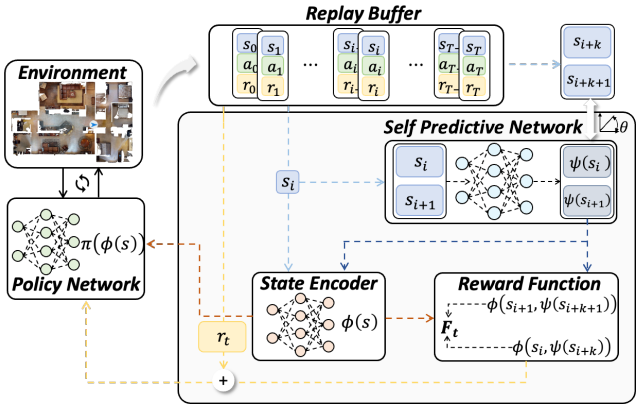


Figure 3: Overview of the LSPE framework. The agent encodes high-dimensional observations via a state encoder ϕ, and predicts future states using a diffusion-based self-predictive network ψ. A latent-space exploration reward F_t is computed based on directional variance in latent predictions. All components are jointly trained to enable temporally aware and structured exploration.

4 Methodology

Our objective is to develop an effective exploration strategy for deep reinforcement learning that generalizes across diverse long-horizon tasks in realistic environments with high-dimensional state spaces.

4.1 Bisimulation-based State Encoder

To achieve **(N1)**, we begin by learning a state encoder $\phi : \mathcal{S} \to \mathcal{Z}$ that maps high-dimensional observations to a compact latent space, thereby reducing the effective exploration space. The representation function ϕ is designed to satisfy two key desiderata: (1) compressing the raw input into a low-dimensional and compact representation, and (2) filtering out task-irrelevant noise while preserving the essential information necessary for decision-making.

Motivated by recent advances in representation learning in RL (Zhang et al. 2021a; Castro et al. 2021), we adopt a bisimulation-based objective to learn the latent space. Bisimulation metrics group behaviorally similar states by quantifying dissimilarity in terms of both reward differences and the divergence in next-state distributions. For example, MICo (Castro et al. 2021) defines the following metric:

Definition 1 (MICo Bisimulation Metric)**.** *Let $\mathbb{M}$ be the space of distance functions $d : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$. The MICo behavioral metric function $\mathcal{F}^\pi_M : \mathbb{M} \to \mathbb{M}$ is defined as:*

$$\mathcal{F}^\pi_M(d)(s_i, s_j) = \mathbb{E}_{a_{i,j}\sim\pi}|r_i - r_j| + \gamma\mathbb{E}_{s'_{i,j}\sim P} d(s'_i, s'_j) \quad (2)$$

where $r_i = r(s_i, a_i)$ denotes the reward and P denotes the transition probability function.

Bisimulation-based representation learning typically minimizes the following objective:

$$\mathcal{L}(\phi) = \mathbb{E}_{s_i,s_j\sim\mathcal{D},s'_i\sim P^\pi_{s_i},s'_j\sim P^\pi_{s_j}}[d_\phi(s_i, s_j) - |r_i - r_j| - \gamma d_\phi(s_{i+1}, s_{j+1})] \quad (3)$$

Here, d_ϕ denotes the latent distance induced by encoder ϕ, and $\mathcal{D}$ represents the replay buffer.

Limitations in Long-Horizon RL. Despite its theoretical appeal, the bisimulation objective in Eq. (3) faces critical limitations in long-horizon settings. First, it relies on single-step temporal difference updates, making it well-suited for capturing short-term dynamics but ill-equipped to encode long-term temporal dependencies. This issue becomes especially pronounced in tasks with sparse rewards, where the term $|r_i - r_j|$ is often zero for most transitions, rendering the learning signal ineffective. Consequently, the encoder is not incentivized to differentiate states based on long-term behavioral outcomes. Furthermore, encoding temporal information into latent representations while maintaining training stability remains a non-trivial challenge. Capturing long-term dependencies without destabilizing the encoder requires the agent to reason about future latent dynamics.

To address these limitations, we introduce a self-predictive network that endows the agent with latent-level *foresight*, enabling temporally consistent and enriched representation learning (detailed next).

4.2 Self-Predictive Representation Learning

Motivation. While bisimulation-based encoders offer a principled way to abstract away irrelevant details and group behaviorally similar states, they typically rely on short-term transitions and single-step dynamics. This limits their effectiveness in long-horizon tasks, especially under sparse rewards and high-dimensional observations. To address this, we equip the agent with predictive "foresight" through a self-supervised mechanism that anticipates future states based on its current observation.

Prior self-predictive methods, such as contrastive learning via cosine similarity (Oord, Li, and Vinyals 2018) or deterministic regression of the next frame as in World Models (Ha and Schmidhuber 2018) and PlaNet (Hafner et al. 2019b), are inherently limited. These approaches collapse

diverse futures into a single prediction, blurring semantically distinct outcomes, and suffer from error accumulation over long horizons. As a result, they fail to provide calibrated and meaningful structure in temporally extended visual tasks.

Diffusion-Based State Prediction. To address these limitations, we propose a Diffusion-Based Self-Predictive Network (**D-SPN**), which models the conditional distribution of future states using a de-noising diffusion process. Unlike previous predictive methods, D-SPN generates future observations via a stochastic generative process, preserving multimodal futures and providing temporally consistent, uncertainty-aware predictions.

Formally, given an input observation $s_t \in \mathcal{S}$, we train D-SPN to model the conditional distribution using a denoising diffusion probabilistic model (DDPM) (Ho, Jain, and Abbeel 2020; Nichol and Dhariwal 2021). The training process involves two components:

(1) Forward process: We corrupt the future observation s_{t+k} with Gaussian noise over N steps to obtain $s_{t+k}^{(n)}$,

$$q(s_{t+k}^{(n)}|s_{t+k}) = \mathcal{N}(s_{t+k}^{(n)}; \sqrt{\alpha_n}s_{t+k}, (1-\alpha_n)\mathbf{I}) \quad (4)$$

where α_n denotes the variance schedule at step n, and $\mathbf{I}$ is the identity covariance, implying isotropic noise.

(2) Reverse process: We train a denoising model ϵ_ψ to recover noise added during the forward process, conditioned on s_t,

$$\mathcal{L}_{\text{D-SPN}}(\psi) = \mathbb{E}_{s_{t+k},n,\epsilon}[\|\epsilon_\psi(s_{t+k}^{(n)}, s_t, n) - \epsilon\|^2] \quad (5)$$

At inference time, we sample $s_{t+k}^{(N)} \sim \mathcal{N}(0, I)$ and iteratively apply the learned reverse denoising steps to obtain $\hat{s}_{t+k} = \psi(s_{t+k})$, the predicted future temporal embedding conditioned on s_t. This diffusion-based prediction strategy effectively maintains temporal consistency, captures multimodal future behaviors, and produces uncertainty estimates essential for exploration under long-horizon settings.

Temporal-Aware Bisimulation. To integrate the long-range temporal consistency learned by D-SPN into the encoder training, we extend the bisimulation metric to account for multi-step temporal structure. Specifically, we define a temporal-aware bisimulation metric as follows:

Definition 2 (Temporal-Aware Bisimulation Metric)**.** *We define a temporal-aware bisimulation-based distance function $d_\phi^{TM} : (\mathcal{S} \times \mathcal{S}) \times (\mathcal{S} \times \mathcal{S}) \to \mathbb{R}$ which compares state pairs (s_i, s_{i+k}) and (s_j, s_{j+k}) within their respective trajectories as:*

$$\begin{aligned} d_\phi^{TM}(s_i, s_{i+k}, s_j, s_{j+k}) &= |r(s_i, a_i) - r(s_j, a_j)| \\ &+ \gamma\, \mathbb{E}[d_\phi^{TM}(s_{i+1}, s_{i+k}, s_{j+1}, s_{j+k})] \end{aligned} \quad (6)$$

where $k \in 1, ..., K$ is the prediction horizon.

We jointly train the encoder ϕ and D-SPN (ψ) by minimizing a consistency objective between the temporal-aware bisimulation metric and the predicted future states from D-SPN. To stabilize joint training and prevent representational drift, we adopt an EMA (Exponential Moving Average) target encoder ϕ_{target}, which evolves slowly and provides a fixed reference for the temporal metric:

$$\begin{aligned} \mathcal{L} = \mathbb{E}[&d_\phi^{\text{TM}}((s_i, \psi(s_{i+k})), (s_j, \psi(s_{j+k})) - |r_i - r_j| \\ &- \gamma d_{\phi_{\text{target}}}^{\text{TM}}((s_{i+1}, \psi(s_{i+k}), (s_{j+1}, \psi(s_{j+k}))]^2 \end{aligned} \quad (7)$$

where $\psi(s_{t+k}) = \hat{s}_{t+k}$ is the predicted future state and ϕ_{target} is updated via $\phi_{\text{target}} \leftarrow \tau\phi_{\text{target}} + (1-\tau)\phi$ with a decay factor $\tau = 0.99$. This formulation aligns the learned latent representation with temporally grounded dynamics captured by the diffusion process, enabling richer abstraction and better generalization in long-horizon settings.

Theorem 1 (Fixed-Point Convergence)**.** *Given a policy π, with the convergence of the diffusion-based state predictive model (D-SPN), the temporal-aware bisimulation metric operator $\mathcal{F}$ has a fixed-point d_ϕ^{TM}.*

Convergence Guarantee. Proof in Appendix A. Based on Theorem 1, our extended formulation retains convergence convergence under the bisimulation framework.

4.3 Exploration Reward Function

Motivation. To achieve **(N2)**: incentivizing exploration in high-dimensional, long-horizon environments, we propose a *directional-variance reward* that leverages two complementary signals: (i) the latent displacement between adjacent states in temporal-aware bisimulation metric space, and (ii) the predictive uncertainty from the diffusion-based D-SPN model. Traditional exploration bonuses, such as prediction error, pseudo-counts, or ℓ_p-norm state differences, either scale poorly with high-dimensional inputs or fail to capture state uncertainty in long-horizon setting. Our approach encourages the agent to move in directions that are both novel and uncertain, thereby promoting effective exploration. Given two consecutive states (s_t, s_{t+1}), we define the latent displacement as:

$$\Delta_t = \phi(s_t, \psi(s_{t+k})) - \phi(s_{t+1}, \psi(s_{t+k+1})) \quad (8)$$

Here, $\psi(s_{t+k})$ denotes the predicted future state from D-SPN, and ϕ is the state encoder.

To quantify uncertainty during exploration, we generate M rollouts $\{\hat{s}_{t+k}^{(m)}\}_{m=1}^M$ from the D-SPN posterior and compute the covariance matrix of their latent projections:

$$\Sigma_t = \text{Cov}\,[\,\phi(\hat{s}_{t+k}^{(1:M)})\,] \quad (9)$$

We then sample a unit direction $z \sim \text{Unif}(\mathbb{S}^{d-1})$ from the unit sphere in latent space, and define the directional-variance exploration reward:

$$F(s_t, s_{t+1}, z) = \underbrace{\langle \Delta_t,\, z\rangle^2}_{\text{directed progress}} \times \underbrace{\langle \Sigma_t z,\, z\rangle}_{\text{uncertainty along } z} \quad (10)$$

Geometrically, this reward is high only when the agent moves in a novel direction and the predictive model remains uncertain along that direction. If either factor is small, the reward vanishes. This encourages exploration that is simultaneously diverse and informative, pushing the agent to span underexplored regions of the latent and thus real state space.

Information-Gain Perspective. The exploration reward is theoretically grounded: it ascends a provable lower bound

on the mutual information between the agent's actions and the future dynamics:

Theorem 2 (Directional-Variance Bonus Captures an Information-Gain Lower Bound)**.** *Let the predictive posterior*[2] *from D-SPN be* $p_\theta(z_{t+k} \mid s_t) = \mathcal{N}(\mu_t, \Sigma_t)$ *in latent space* $\mathcal{Z}$ *(*$z_{t+k} = \phi(s_{t+k})$*). Define the same-horizon latent displacement as Eq. (8), and let* $z \sim \mathrm{Unif}(\mathbb{S}^{d-1})$ *be a unit direction. Then the expected reward directional-variance* F_t *satisfies:*

$$\mathbb{E}_z[F_t(z)] = \frac{1}{d(d+2)} \left(\|\Delta_t\|^2 \cdot \mathrm{tr}(\Sigma_t) + 2 \cdot \Delta_t^\top \Sigma_t \Delta_t \right) \tag{11}$$

Moreover, this expectation provides a lower bound (up to a constant C*) on the mutual information between* Δ_t *and the predicted future latent state* z_{t+k}*:*

$$\mathbb{E}_z[F_t(z)] \;\propto\; \mathcal{I}(z_{t+k}; \Delta_t) + C \tag{12}$$

where the constant C *depends only on* $\|\Delta_t\|^2$ *and the dimensionality of latent space* d*.*

Proof in Appendix A. Thus, maximizing F_t (Eq. (10)) corresponds to ascending a stochastic lower bound on the information gained about future dynamics. The exploration reward automatically anneals as uncertainty decreases, scales to visual domains, and improves sample efficiency and policy coverage in sparse-reward, long-horizon tasks. We outline the full training procedure in Algorithm 1.

Algorithm 1: Latent State-Predictive Exploration (LSPE)

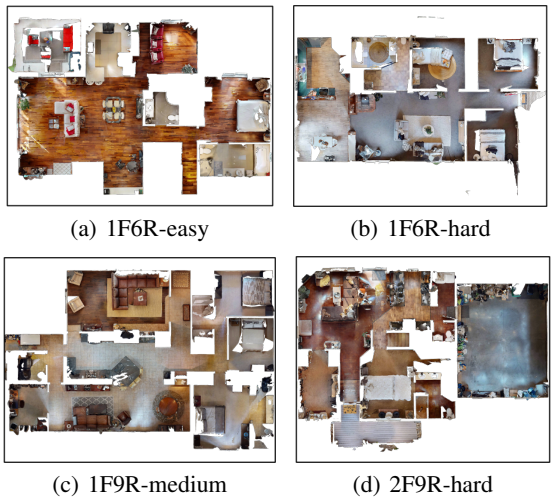
Initialize: policy π_θ, encoder ϕ, diffusion-based predictor ψ, replay buffer $\mathcal{D}$
while *not converged* **do**
  Sample unit latent vector$z \sim \mathrm{Unif}(\mathbb{S}^{d-1})$
  for $t = 1$ to MAX_STEP_PER_EPISODE **do**
    Sample action $a_t \sim \pi_\theta(\cdot \mid \phi(s_t))$ and reach s_{t+1}
    Record transition in the buffer $\mathcal{D}$
    Compute exploration reward F_t ▷Eq. (10)
    Reshape the reward: $r'_t = r_t + F_t$
    Update policy π_θ via policy gradient
    Sample minibatch $B \sim \mathcal{D}$
    Update predictor D-SPN via $\mathbb{E}_B[\mathcal{L}(\psi)]$ ▷Eq. (5)
    Joint update the state encoder (ϕ) and the D-SPN predictor (ψ) $\mathbb{E}_B[\mathcal{L}(\phi, \psi)]$ ▷Eq. (7)
  end for
end while

5 Experimental Results

To evaluate the performance of LSPE, we conduct extensive experiments on various long-horizon settings of different environments to assess the effectiveness of our algorithm.

Baselines. We compare LSPE with the following baselines: (1) ICM (Pathak et al. 2017b): A widely used curiosity-driven exploration approach. (2) RND (Burda et al. 2019a): A typical bonus-based exploration baseline. (3) EME (Wang et al. 2024): A dynamic bonus-based method tailored for high-dimensional environments and challenging exploration tasks. (4) METRA (Park, Rybkin, and Levine 2023): An unsupervised method that induces diverse skills and long-horizon behaviors. (5) SkillTree (Wen et al. 2025): A hierarchical framework that distills actions into discrete skills to enhance long-horizon exploration. (6) NoisyNet (Fortunato et al. 2018): A noise-based exploration approach that perturbs the value function to promote diverse actions.

[2]For clarity, we state the result in latent space; an analogous form holds in original state space by replacing $z = \phi(s)$ with s.



(a) 1F6R-easy (b) 1F6R-hard (c) 1F9R-medium (d) 2F9R-hard

Figure 4: Four indoor environments with different settings (different number of floors and rooms) with different complexity level (easy, medium, hard) for the navigation task.

5.1 Experimental Setting

Navigation Tasks on Habitat. Habitat (Manolis Savva et al. 2019) 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 1(a), in each episode the robot agent is placed in a different indoor space consisting of connected rooms, with the task of navigating from the initial position to a designated destination. We utilize the Habitat-Matterport 3D (HM3D) Research Dataset (Yadav et al. 2023), which contains high-quality renderings of indoor scenes. As depicted in Fig. 4, we select four indoor environments with varying settings and complexity levels. For instance, "1F6R-easy" denotes an environment with one floor and six rooms, with navigation complexity rated as easy.

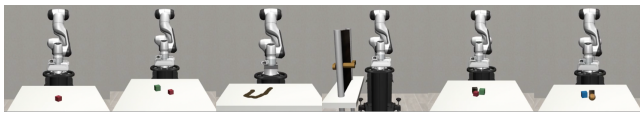


Figure 5: Manipulation tasks: Lift, Stack, Table Wiping, Door Opening, and two Stack Out-of-Distribution (OOD) variants. Stack OOD-A: add an interference term; Stack OOD-C: alter the block's shape and color.

Long-Horizon Control Tasks. We evaluate our method on four long-horizon manipulation tasks from Robosuite (Zhu et al. 2020): Lift, Stack, Table Wiping, and Door Opening. In all tasks, initial conditions are randomized per episode.

- Lift: Raise a randomly placed block above a target height.
- Stack: Place one block on top of the another from randomized initial positions.
- Door Opening: Rotate the handle and open a door from a randomized initial pose.
- Table Wiping: Remove randomly placed whiteboard markings by wiping the surface.

For further details, please refer to Appendix C.

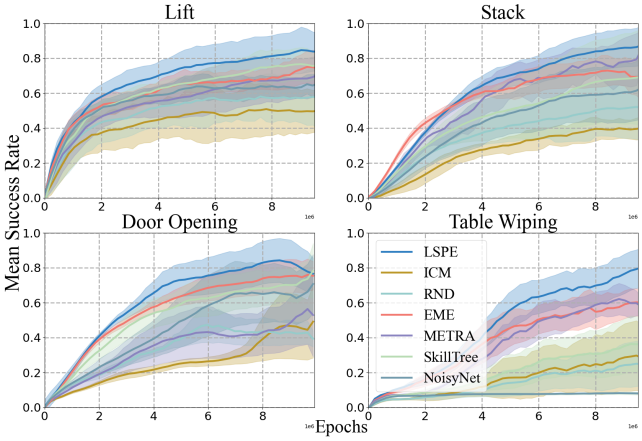


Figure 6: Comparison results for the manipulation tasks (Lift, Stack, Door Opening and Table Wiping). The solid lines and shaded areas represent the mean values and the standard errors, respectively, over five different seeds.

5.2 Experimental Results and Analysis

Overall Performance. As shown in Fig. 6, LSPE consistently outperforms all baseline methods across the manipulation tasks, highlighting its effectiveness in long-horizon, continuous control settings. This superior performance is attributed to LSPE's directional-variance exploration reward, which encourages structured exploration and accelerates policy learning. In contrast, the action-noise method NoisyNet fails to scale in high-dimensional spaces, where stochastic perturbations become ineffective. Similarly, unsupervised latent reward-based method METRA, exhibits limited guidance for exploration. The state-difference method EME struggles to provide sufficient exploration incentives due to the subtlety of visual changes in pixel-based inputs. Hierarchical approaches like SkillTree also underperform, likely due to insufficient low-level exploration. Curiosity-based methods such as ICM and RND suffer from unstable or weak novelty signals, further limiting their ability to drive meaningful exploration in complex environments.

In navigation tasks, LSPE exhibits strong scalability, consistently outperforming all baselines across four environments of increasing complexity. Each setting is evaluated

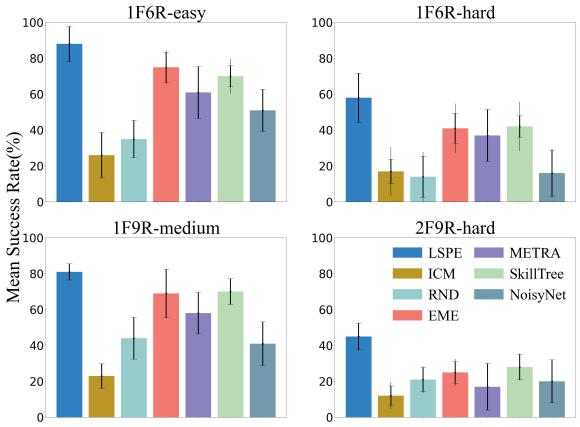


Figure 7: Comparison results for the navigation tasks across four environments with different complexity levels.

over five random seeds. In the easiest environment, LSPE achieves a success rate of approximately 90%, while other methods remain below 70%, except for EME. As task complexity increases, LSPE maintains robust performance, reaching around 80% in the medium setting and nearly 50% in the most challenging environment, which contains the largest number of rooms. In contrast, all baselines show significant performance degradation. These results demonstrate LSPE's robustness and generalization ability in long-horizon navigation across diverse and complex visual environments.

Ablation Results. We evaluate two LSPE variants to assess the contribution of key components: **LSPE w/o ERF**, which removes the exploration reward function, and **LSPE w MSD**, which replaces the temporal-aware encoder with a MICo-based alternative. As shown in Table 1, removing the exploration reward significantly reduces performance, though the variant still outperforms strong baselines such as ICM, RND, and METRA. Replacing the encoder also leads to performance degradation, highlighting the importance of temporal consistency in representation learning. Additional ablation studies on the diffusion model and prediction horizon are presented in Appendix B.

Figure 8: Navigation tasks tested in realistic indoor environments via Unitree Robotic Go2.

Generalization Ability. To assess LSPE's ability to generalize under distributional shifts, we introduce two Out-of-Distribution (OOD) variants of the default stacking task, illustrated in Fig. 5. In the **OOD-A** setting ("A" for Add),

Environment	Stack OOD-A	Stack OOD-C	RealEnv-Easy	RealEnv-Medium	RealEnv-Hard
ICM	19.5%±8.3%	11.2%±5.5%	30.5%±7.5%	20.4%±5.3%	9.1%±2.0%
RND	28.0%±7.2%	15.2%±4.5%	49.8%±6.0%	27.2%±5.5%	19.8%±5.8%
EME	39.8%±12.9%	28.6%±12.5%	72.5%±9.5%	70.2%±5.5%	55.4%±9.4%
METRA	25.7%±10.5%	21.3%±10.7%	39.8%±8.5%	30.2%±6.6%	18.2%±6.5%
SkillTree	44.3%±11.2%	39.7%±10.1%	50.5%±8.5%	38.4%±7.1%	21.2%±6.5%
NosiyNet	23.6%±7.5%	16.9%±7.6%	45.5%±8.8%	34.3%±8.7%	25.2%±7.5%
LSPE	**75.7**%±**12.6**%	**70.9**%±**13.3**%	**91.5**%±**8.5**%	**85.2**%±**7.6**%	**78.4**%±**7.8**%
LSPE w/o ERF	31.6%±9.1%	17.6%±11.0%	45.5%±8.5%	40.2%±7.8%	30.3%±6.6%
LSPE w MSD	55.7%±10.1%	38.9%±10.8%	69.5%±8.4%	58.9%±9.4%	49.7%±8.9%

Table 1: Mean success rates (averaged over 10 random seeds) in the out-of-distribution (OOD) generalization settings of the manipulation task, as well as performance of real-robot deployments across different physical indoor environments.

an immovable black block is added to the scene, acting as a distractor. In the **OOD-C** setting ("C" for Change), the color of two blocks is modified and one block is replaced with a sphere to introduce spurious correlations. Table 1 reports quantitative results. LSPE achieves the highest success rates in both OOD settings, demonstrating that its state encoder effectively filters out task-irrelevant factors (e.g., distractors), while its exploration reward encourages broad and meaningful interaction, enhancing generalization. Among the LSPE variants, LSPE w/ MSD achieves the second-best performance in OOD-A. Even LSPE w/o ERF remains among the top-five performers across both in-distribution and OOD tasks. These results highlight the robustness and strong generalization capacity of LSPE under diverse and challenging OOD conditions.

Scalability to Real World. To validate the real-world applicability of LSPE, we deploy the trained policy on a Unitree Go2 quadruped robot and evaluate its performance in physical indoor environments. We collects real-world visual trajectory data to train LSPE entirely in the target domain, eliminating the need for domain adaptation or simulation fine-tuning. As illustrated in Fig. 8, the robot is equipped with a front-facing RGB-D camera and operates in three progressively challenging indoor environments: **Easy** (a structured room with sparse furniture), **Medium** (an office-like space with moderate clutter), and **Hard** (a multi-room layout with occlusions, distractors, and narrow passages). Detailed setup descriptions are provided in Appendix C. As shown in Table 1, the LSPE policy consistently achieves the best performance across all physical environments. In contrast, baseline methods such as RND and NoisyNet, trained on the same real-world datasets, struggle to generalize and frequently get trapped in local regions. These results demonstrate that LSPE's structured latent exploration, guided by predictive modeling, scales robustly to realistic settings.

6 Related Work

Exploration is a fundamental challenge in reinforcement learning (RL), especially in long-horizon tasks. Classical strategies include ϵ-greedy (Sutton 2018), count-based methods (Bellemare et al. 2016b; Ostrovski et al. 2017a), and curiosity-driven approaches (Stanton and Clune 2016, 2018; Burda et al. 2018; Zhao et al. 2024). Recent works incorporate intrinsic rewards via learned dynamics (Pathak et al. 2017b; Burda et al. 2019a; Pathak, Gandhi, and Gupta 2019), latent dynamics models (Bai et al. 2021; Tao, François-Lavet, and Pineau 2020; Raileanu and Rocktäschel 2020; Mahankali et al. 2024), or entropy maximization over state embeddings (Seo et al. 2021; Yadav et al. 2023; Chen et al. 2024). Pseudo-count–based methods (Bellemare et al. 2016a; Ostrovski et al. 2017b; Tang et al. 2017) explore low-density regions via density estimation, and hybrid approaches (Raileanu and Rocktäschel 2020; Badia et al. 2020b,a) combine these with model-based objectives. Others directly reward state differences (Wang et al. 2023; Zhang et al. 2021b; Henaff et al. 2022; Wang et al. 2024). Skill-based abstraction has been explored for efficient long-horizon exploration (Yuan et al. 2023; Wen et al. 2025), as have imitation-augmented methods (Gupta et al. 2019), unsupervised approaches (Eysenbach et al. 2018; Park, Rybkin, and Levine 2023; Park, Kreiman, and Levine 2024; Ying et al. 2025) and hierarchical RL with temporal decoupling (Lee et al. 2022). Finally, model-based approaches leverage imagined rollouts for planning exploratory behavior (Shyam, Jaśkowski, and Gomez 2019; Ratzlaff et al. 2020; Hafner et al. 2019a), though their effectiveness depends heavily on model fidelity. Details in Appendix D.

7 Conclusion

In this work, we proposed Latent State-Predictive Exploration (LSPE), a novel and scalable framework for efficient exploration in long-horizon reinforcement learning. LSPE employs a compact, bisimulation-based state encoder to filter out irrelevant and noisy features from high-dimensional observations. To further enhance temporal structure and representation stability in long-horizon settings, we integrate a diffusion-based self-predictive module that infuses temporal information into the learned embeddings. Additionally, we introduce a theoretically grounded exploration reward that encourages the agent to visit novel and uncertain regions in the latent space, facilitating structured and efficient exploration. Extensive experiments on navigation and manipulation tasks demonstrate that LSPE consistently outperforms strong baselines across varying levels of difficulty. Finally, LSPE achieves robust real-world performance when deployed on a quadruped robot in physical indoor environments, confirming its scalability and practical effectiveness.

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) and 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. This work was performed in part at SICC which is supported by SKL-IOTSC, University of Macau.

References

Badia, A. P.; Piot, B.; Kapturowski, S.; Sprechmann, P.; Vitvitskyi, A.; Guo, Z. D.; and Blundell, C. 2020a. Agent57: Outperforming the atari human benchmark. In *International conference on machine learning*, 507–517. PMLR.

Badia, A. P.; Sprechmann, P.; Vitvitskyi, A.; Guo, D.; Piot, B.; Kapturowski, S.; Tieleman, O.; Arjovsky, M.; Pritzel, A.; Bolt, A.; et al. 2020b. Never give up: Learning directed exploration strategies. *arXiv preprint arXiv:2002.06038*.

Bai, C.; Liu, P.; Liu, K.; Wang, L.; Zhao, Y.; Han, L.; and Wang, Z. 2021. Variational dynamic for self-supervised exploration in deep reinforcement learning. *IEEE Transactions on neural networks and learning systems*, 34(8): 4776–4790.

Bellemare, M.; Srinivasan, S.; Ostrovski, G.; Schaul, T.; Saxton, D.; and Munos, R. 2016a. Unifying count-based exploration and intrinsic motivation. *Advances in neural information processing systems*, 29.

Bellemare, M.; Srinivasan, S.; Ostrovski, G.; Schaul, T.; Saxton, D.; and Munos, R. 2016b. Unifying count-based exploration and intrinsic motivation. *Advances in neural information processing systems*, 29.

Burda, Y.; Edwards, H.; Pathak, D.; Storkey, A.; Darrell, T.; and Efros, A. A. 2018. Large-scale study of curiosity-driven learning. *arXiv preprint arXiv:1808.04355*.

Burda, Y.; Edwards, H.; Storkey, A. J.; and Klimov, O. 2019a. Exploration by random network distillation. In *ICLR (Poster)*. OpenReview.net.

Burda, Y.; Edwards, H.; Storkey, A. J.; and Klimov, O. 2019b. Exploration by random network distillation. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Castro, P. S.; Kastner, T.; Panangaden, P.; and Rowland, M. 2021. MICo: Improved representations via sampling-based state similarity for Markov decision processes. *Advances in Neural Information Processing Systems*, 34: 30113–30126.

Chen, Y.; Zhao, K.; Wang, Y.; Yang, M.; Zhang, J.; and Niu, X. 2024. Enhancing LLM Agents for Code Generation with Possibility and Pass-rate Prioritized Experience Replay. *arXiv preprint arXiv:2410.12236*.

Eysenbach, B.; Gupta, A.; Ibarz, J.; and Levine, S. 2018. Diversity is all you need: Learning skills without a reward function. *arXiv preprint arXiv:1802.06070*.

Fortunato, M.; Azar, M. G.; Piot, B.; Menick, J.; Hessel, M.; Osband, I.; Graves, A.; Mnih, V.; Munos, R.; Hassabis, D.; Pietquin, O.; Blundell, C.; and Legg, S. 2018. Noisy Networks For Exploration. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Gupta, A.; Kumar, V.; Lynch, C.; Levine, S.; and Hausman, K. 2019. Relay policy learning: Solving long-horizon tasks via imitation and reinforcement learning. *arXiv preprint arXiv:1910.11956*.

Ha, D.; and Schmidhuber, J. 2018. World models. *arXiv preprint arXiv:1803.10122*, 2(3).

Hafner, D.; Lillicrap, T.; Ba, J.; and Norouzi, M. 2019a. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*.

Hafner, D.; Lillicrap, T.; Fischer, I.; Villegas, R.; Ha, D.; Lee, H.; and Davidson, J. 2019b. Learning latent dynamics for planning from pixels. In *International conference on machine learning*, 2555–2565. PMLR.

Henaff, M.; Raileanu, R.; Jiang, M.; and Rocktäschel, T. 2022. Exploration via elliptical episodic bonuses. *Advances in Neural Information Processing Systems*, 35: 37631–37646.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Lee, S.; Kim, J.; Jang, I.; and Kim, H. J. 2022. Dhrl: A graph-based approach for long-horizon and sparse hierarchical reinforcement learning. *Advances in Neural Information Processing Systems*, 35: 13668–13678.

Mahankali, S.; Hong, Z.-W.; Sekhari, A.; Rakhlin, A.; and Agrawal, P. 2024. Random Latent Exploration for Deep Reinforcement Learning. *arXiv preprint arXiv:2407.13755*.

Manolis Savva; Abhishek Kadian; Oleksandr Maksymets; Zhao, Y.; Wijmans, E.; Jain, B.; Straub, J.; Liu, J.; Koltun, V.; Malik, J.; Parikh, D.; and Batra, D. 2019. Habitat: A Platform for Embodied AI Research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A. A.; Veness, J.; Bellemare, M. G.; Graves, A.; Riedmiller, M.; Fidjeland, A. K.; Ostrovski, G.; et al. 2015. Human-level control through deep reinforcement learning. *nature*, 518(7540): 529–533.

Nichol, A. Q.; and Dhariwal, P. 2021. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, 8162–8171. PMLR.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Ostrovski, G.; Bellemare, M. G.; Oord, A.; and Munos, R. 2017a. Count-based exploration with neural density models. In *International conference on machine learning*, 2721–2730. PMLR.

Ostrovski, G.; Bellemare, M. G.; Oord, A.; and Munos, R. 2017b. Count-based exploration with neural density models. In *International conference on machine learning*, 2721–2730. PMLR.

Park, S.; Kreiman, T.; and Levine, S. 2024. Foundation policies with hilbert representations. *arXiv preprint arXiv:2402.15567*.

Park, S.; Rybkin, O.; and Levine, S. 2023. Metra: Scalable unsupervised rl with metric-aware abstraction. *arXiv preprint arXiv:2310.08887*.

Pathak, D.; Agrawal, P.; Efros, A. A.; and Darrell, T. 2017a. Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*, 2778–2787. PMLR.

Pathak, D.; Agrawal, P.; Efros, A. A.; and Darrell, T. 2017b. Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*, 2778–2787. PMLR.

Pathak, D.; Gandhi, D.; and Gupta, A. 2019. Self-supervised exploration via disagreement. In *International conference on machine learning*, 5062–5071. PMLR.

Raileanu, R.; and Rocktäschel, T. 2020. RIDE: Rewarding Impact-Driven Exploration for Procedurally-Generated Environments. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Ratzlaff, N.; Bai, Q.; Fuxin, L.; and Xu, W. 2020. Implicit generative modeling for efficient exploration. In *International Conference on Machine Learning*, 7985–7995. PMLR.

Seo, Y.; Chen, L.; Shin, J.; Lee, H.; Abbeel, P.; and Lee, K. 2021. State entropy maximization with random encoders for efficient exploration. In *International Conference on Machine Learning*, 9443–9454. PMLR.

Shyam, P.; Jaśkowski, W.; and Gomez, F. 2019. Model-based active exploration. In *International conference on machine learning*, 5779–5788. PMLR.

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

Stanton, C.; and Clune, J. 2018. Deep curiosity search: Intra-life exploration improves performance on challenging deep reinforcement learning problems. *arXiv preprint arXiv:1806.00553*.

Sutton, R. S. 2018. Reinforcement learning: An introduction. *A Bradford Book*.

Tang, H.; Houthooft, R.; Foote, D.; Stooke, A.; Xi Chen, O.; Duan, Y.; Schulman, J.; DeTurck, F.; and Abbeel, P. 2017. # exploration: A study of count-based exploration for deep reinforcement learning. *Advances in neural information processing systems*, 30.

Tao, R. Y.; François-Lavet, V.; and Pineau, J. 2020. Novelty search in representational space for sample efficient exploration. *Advances in Neural Information Processing Systems*, 33: 8114–8126.

Wang, Y.; Yang, M.; Dong, R.; Sun, B.; Liu, F.; et al. 2023. Efficient potential-based exploration in reinforcement learning using inverse dynamic bisimulation metric. *Advances in Neural Information Processing Systems*, 36.

Wang, Y.; Zhao, K.; Liu, F.; et al. 2024. Rethinking Exploration in Reinforcement Learning with Effective Metric-Based Exploration Bonus. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Wen, Y.; Li, S.; Zuo, R.; Yuan, L.; Mao, H.; and Liu, P. 2025. SkillTree: Explainable Skill-Based Deep Reinforcement Learning for Long-Horizon Control Tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 21491–21500.

Yadav, K.; Ramrakhya, R.; Ramakrishnan, S. K.; Gervet, T.; Turner, J.; Gokaslan, A.; Maestre, N.; Chang, A. X.; Batra, D.; Savva, M.; et al. 2023. Habitat-matterport 3d semantics dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4927–4936.

Ying, C.; Chen, H.; Zhou, X.; Hao, Z.; Su, H.; and Zhu, J. 2025. Exploratory Diffusion Model for Unsupervised Reinforcement Learning. *arXiv preprint arXiv:2502.07279*.

Yuan, H.; Zhang, C.; Wang, H.; Xie, F.; Cai, P.; Dong, H.; and Lu, Z. 2023. Skill reinforcement learning and planning for open-world long-horizon tasks. *arXiv preprint arXiv:2303.16563*.

Zhang, A.; McAllister, R. T.; Calandra, R.; Gal, Y.; and Levine, S. 2021a. Learning Invariant Representations for Reinforcement Learning without Reconstruction. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Zhang, H.; Wang, H.; Huang, X.; Chen, W.; and Kan, Z. 2024. Exploiting Hybrid Policy in Reinforcement Learning for Interpretable Temporal Logic Manipulation. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 13795–13800. IEEE.

Zhang, T.; Xu, H.; Wang, X.; Wu, Y.; Keutzer, K.; Gonzalez, J. E.; and Tian, Y. 2021b. Noveld: A simple yet effective exploration criterion. *Advances in Neural Information Processing Systems*, 34: 25217–25230.

Zhao, K.; Wang, Y.; Chen, Y.; Li, Y.; Niu, X.; et al. 2024. Efficient Diversity-based Experience Replay for Deep Reinforcement Learning. *arXiv preprint arXiv:2410.20487*.

Zhu, Y.; Wong, J.; Mandlekar, A.; Martín-Martín, R.; Joshi, A.; Lin, K.; Maddukuri, A.; Nasiriany, S.; and Zhu, Y. 2020. robosuite: A Modular Simulation Framework and Benchmark for Robot Learning. In *arXiv preprint arXiv:2009.12293*.