Explore to Learn: Latent Exploration Through Disentangled Synergy Patterns for Reinforcement Learning in Overactuated Control

Yiming Wang[1], Kaiyan Zhao[2], Xu Li[1], Yan Li[3], Jiayu Chen[4], Steven Morad[1] Leong Hou U[1]*

[1]University of Macau
[2]Wuhan University
[3]Shenzhen Polytechnic University
[4]University of Hong Kong
{wang.yiming,yb57411}@connect.um.edu.mo, zhao.kaiyan@whu.edu.cn, jiayuc@hku.hk,
smorad@um.edu.mo, ryanlhu@um.edu.mo

Abstract

Control in high-dimensional action spaces remains a fundamental challenge in reinforcement learning (RL), primarily due to inefficient exploration of the action space. While recent methods attempt to guide exploration, they often fall short of achieving the agility and coordination exhibited in biological motor control. Inspired by how organisms exploit muscle synergies for efficient movement, we propose *Explore to Learn (ETL)*, a two-stage framework that first discovers fundamental synergy patterns and then leverages them for task-specific policy learning. In the first stage, ETL discovers underlying synergy patterns by deploying a targeted exploration policy. These patterns are modeled as latent directions in a low-dimensional space, along which the agent is guided to collect diverse and structured muscle activation trajectories. A variational autoencoder (VAE) is then trained to encode high-dimensional actions into a latent space whose dimensions correspond to the synergy patterns. In the second stage, the policy is trained entirely in this synergy-aware latent space, producing synergy coefficients that the decoder maps back to full-dimensional muscle actions. This structured representation significantly reduces the complexity of learning, while the decoder is further fine-tuned to enhance expressiveness and generalization across downstream tasks. Extensive experiments across musculoskeletal environments and the DMControl suite demonstrate that ETL consistently outperforms prior methods in both exploration efficiency and control performance, achieving superior scalability and generalization in overactuated control tasks.

1 Introduction

Exploration is a persistent challenge in reinforcement learning (RL), especially in complex environments. Classical strategies such as ϵ-greedy and Boltzmann exploration inject random actions (Mnih et al. 2015), while modern methods use intrinsic rewards to drive deep exploration. Count-based (Bellemare et al. 2016a) and curiosity-driven approaches (Burda et al. 2018b; Pathak et al. 2017a) incentivize novelty or prediction error. More recent methods (Raileanu, Rocktäschel et al. 2020; Wang et al. 2024) exploit latent states to define scalable exploration bonuses.

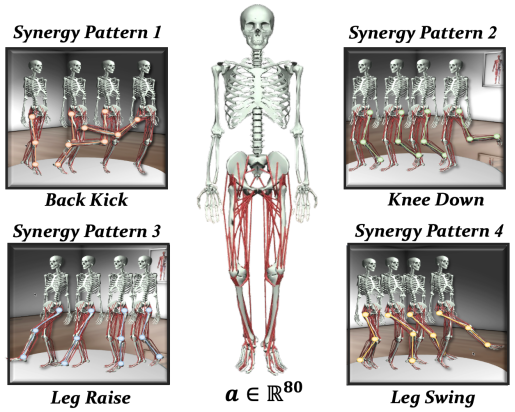

Figure 1: Biological motor control leverages low dimensional muscle synergies for efficient coordination. Inspired by this principle, ETL learns disentangled synergy patterns (e.g., motor primitives like back kick, knee down, leg raise and leg swing) to reduce the burden of overactuated control ($a \in \mathbb{R}^{80}$), enabling tractable and structured exploration.

Despite recent progress, *exploration in high-dimensional action spaces* remains challenging. Tasks such as musculoskeletal control require coordinating many actuators under nonlinear constraints, making policy learning unstable and inefficient. While methods like action noise injection (Chiappa et al. 2023) or self-organization priors (Schumacher et al. 2023) offer partial solutions, they often struggle to match the *dexterity* needed for complex tasks like locomotion or in-hand manipulation. Biological motor control offers inspiration: humans and animals are thought to rely on low-dimensional muscle synergies, enabling efficient and adaptive movement generation (Balda, Pepperberg, and Kamil 1998; Bizzi, Mussa-Ivaldi, and Giszter 1991). As shown in Fig. 1, operating in such compact latent manifolds promotes coordinated locomotion (Caggiano, Cheung, and Bizzi 2016), as illustrated by synergy patterns like back kick, knee down, leg raise, and leg swing.

*Corresponding author.
Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Inspired by how organisms *explore* diverse behaviors to *learn* synergy patterns, we propose Explore to Learn (ETL), a framework that discovers muscle synergies through latent exploration and leverages them as an action representation to improve policy learning in overactuated control tasks. The core intuition is simple yet powerful: *Discovering reusable motor primitives through exploration, then exploiting them for scalable policy learning.* During exploration, synergy patterns are modeled as latent vectors that guide the agent to explore in the latent space. This targeted exploration policy induces richer and more diverse muscle activation trajectories aligned with patterns. A Gaussian Mixture VAE (GMVAE) is trained on the collected actions to map high-dimensional controls into a compact latent space whose dimensionality matches the number of synergy patterns. The alignment associates each latent variable with a distinct motor primitive, yielding disentangled representations and enabling the policy to directly modulate synergy activations for efficient, interpretable control. By restricting exploration to synergy-informed directions, ETL alleviates the curse of dimensionality and accelerates the discovery of meaningful motor primitives. During task learning, the policy operates entirely in the latent space: instead of generating muscle activations, it outputs synergy coefficients that the VAE decodes into full activation vectors, which confines exploration to a low-dimensional synergy manifold, decoupling musculoskeletal complexity from the search process. The decoder is further fine-tuned during downstream training to keep the learned synergies adaptable to task-specific requirements.

The main contributions of this paper are as follows: (1) We propose ETL, a new two-stage learning paradigm for high-dimensional control that first learns a synergy-based action representation through principled exploration and then leverages it for efficient policy learning. (2) We introduce a novel exploration mechanism for synergy discovery, where exploration is actively guided within a behavioral metric space to yield a diverse and structured dataset for action representation learning. (3) We develop a gaussian mixture variational autoencoder-based mechanism for learning compact action representations within musculoskeletal systems and demonstrate the benefit of decoder fine-tuning for reconstructing original actions in downstream tasks. (4) We conduct extensive experiments showing that ETL enhances exploration efficiency and generalization across a variety of challenging musculoskeletal environments.

2 Background

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, $P\left(s' \mid s, a\right)$ 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 $(s_0, a_0, s_1, \dots,)$ by learning a value function (or value network) V_π from the interaction that approximates $V^\pi(s_0) \approx \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right]$.

Latent Exploration. To facilitate exploration, many RL methods augment extrinsic rewards r^e with exploration bonuses r^b, forming $r = r^e + r^b$. While curiosity-driven methods (Bellemare et al. 2016a) are popular choices for r^b, they often underperform in high-dimensional or sparse-reward settings. Latent exploration mitigates this by learning a compact embedding $\phi : \mathcal{S} \to \mathcal{Z}$, where intrinsic rewards are computed more effectively. The encoder ϕ filters irrelevant features, simplifying exploration.

Synergy Pattern. A synergy pattern is a latent *motor primitive* that encodes a coordinated activation profile across multiple actuators. Formally, let $a \subseteq \mathbb{R}^n$ denote the high-dimensional action space. The latent action representation $\mathbf{z^a} = [z^1_{sp}, \dots, z^d_{sp}] \in \mathbb{R}^d$ consists of d disentangled dimensions, where each z^i_{sp} reflects the activation strength of a distinct synergy pattern. A decoder $\texttt{Dec} : \mathbb{R}^d \to \mathbb{R}^n$ maps this latent representation to the original action space, yielding a full action as $a = \texttt{Dec}(\mathbf{z^a})$. This formulation enables structured control by composing high-dimensional actions from interpretable and reusable motor primitives.

3 Related Work

Exploration in RL. Exploration remains a fundamental challenge in RL. Classical approaches, including ϵ-greedy (Sutton, Barto et al. 1998) and count-based methods (Bellemare et al. 2016b; Ostrovski et al. 2017a; Yang et al. 2023a), rely on stochasticity or visitation frequency to encourage exploration. Modern methods introduce intrinsic rewards, such as curiosity-driven objectives (Yang et al. 2023b; Stanton and Clune 2016; Burda et al. 2018a; Yang et al. 2024), prediction errors from learned dynamics (Pathak et al. 2017b; Burda et al. 2019; Pathak, Gandhi, and Gupta 2019; Wang et al. 2025), or entropy maximization in latent spaces (Seo et al. 2021). State novelty is often estimated via density modeling: pseudo-count techniques (Bellemare et al. 2016a; Ostrovski et al. 2017b; Tang et al. 2017; Zhao et al. 2024) promote visits to low-density regions, while hybrid methods (Raileanu, Rocktäschel et al. 2020; Badia et al. 2020b,a; Chen et al. 2024) integrate these signals with model-based or policy-based objectives. Recent work also rewards latent state transitions with high novelty (Zhang et al. 2021; Henaff et al. 2022; Wang et al. 2023, 2024), enabling deeper and more structured exploration.

Overactuated Control. Overactuated systems, such as musculoskeletal control, pose significant exploration challenges due to actuator redundancy and the curse of dimensionality. Prior work mitigates this by injecting structured noise (Chiappa et al. 2023), leveraging self-organizing controllers (Schumacher et al. 2023), or learning coordination priors like muscle networks (Luo et al. 2023), torque constraints (Jiang et al. 2019), and energy-based models. Biologically inspired approaches simplify control via muscle synergies, often by manually grouping actuators (Joos, Péan, and Goksel 2020), reducing action dimensionality (Tieck

et al. 2018; Tahami, Jafari, and Fallah 2014), or extracting low-dimensional structures via PCA (Al Borno, Hicks, and Delp 2020; Zhao et al. 2022), typically relying on expert data. In contrast, our approach discovers synergy patterns autonomously through latent exploration, requiring no supervision and generalizing across diverse domains.

4 Methodology

Motivation. Biological organisms often exhibit a natural "***Explore to Learn***" cycle, where spontaneous, exploratory movements unveil latent motor patterns, or muscle synergies, that eventually guide more dexterous behaviors (Hacques et al. 2021). For instance, infants initially generate diverse, seemingly random muscle activations and gradually identify co-activation patterns that support stable and purposeful movement. Inspired by this, we propose Explore to Learn (ETL), a two-stage framework for discovering and leveraging synergy-based action representations. As shown in Fig. 2, in the first stage, ETL guides exploration along synergy-guided directions in latent space, encouraging the agent to uncover behaviorally meaningful muscle activations. A Gaussian Mixture VAE is then trained to encode these high-dimensional actions into compact, disentangled synergy representations. In the second stage, the policy learns in this latent space, outputting synergy coefficients that are decoded into full muscle actions. This structured representation enables more efficient exploration and learning in overactuated control settings. ETL thus operationalizes a biologically grounded paradigm: discovering reusable motor primitives through exploration, then exploiting them for scalable policy learning. The details are as follows.

4.1 Synergy Pattern-driven Latent Exploration

Desiderata. Biological studies (Dominici et al. 2011; Wainwright 2002) suggest that a small number of underlying *synergy patterns* can be linearly combined to generate a wide repertoire of motor behaviors. Capturing this diversity in muscle activation trajectories is **essential** for learning disentangled motor primitives. To this end, we identify two key desiderata for synergy pattern-driven exploration:

(D1) **Compact latent exploration:** Rather than exhaustively visiting all states in the MDP, which is infeasible in complex environments, exploration should be conducted in a compact latent space $\mathcal{Z}$ to ensure tractability and efficiency.

(D2) **Behavioral diversity via latent directions:** Synergy patterns should correspond to latent directions that induce *distinct behaviors*. These behaviors should collectively span diverse regions of the state space, promoting disentangled representation learning.

Objective. To address **(D1)**, we propose learning an exploration policy π_{exp} through latent exploration guided by synergy patterns. We first learn a state encoder $\phi : \mathcal{S} \to \mathcal{Z}$ and define a set of unit latent vectors $z_{sp}^1, z_{sp}^2, \ldots, z_{sp}^d \subset \mathcal{Z}_{sp}$, each representing a hypothesized synergy direction. These vectors serve as directional priors that guide the agent to explore behaviorally meaningful trajectories. As shown in Fig. 2 (left), navigating along different latent directions z_{sp} allows the agent to generate diverse yet coherent muscle activations, facilitating the discovery of distinct synergy patterns. To encourage such traversal, we define the following inner-product-based exploration bonus:

$$r^b(s, s', z_{sp}) = \langle \phi(s') - \phi(s), z_{sp} \rangle \tag{1}$$

This bonus function (1) provides an reward signal that guides the exploration policy π_{exp} to collect transitions aligned with the latent direction z_{sp}, effectively encouraging the emergence of structured and diverse motor behaviors.

To achieve **(D2)**, ensuring that distinct latent directions lead to distinct behaviors, we employ a *behavioral metric-based state encoder* ϕ. This design is motivated by the insight that distinct synergy patterns should correspond to functionally different motor behaviors. Thus, the encoder should preserve behavioral dissimilarity in the latent space.

Definition 1 (Behavioral Metric)**.** *Let $\mathcal{M}$ be the set of pseudometrics: $d : S \times S \to \mathbb{R}_{\geq 0}$. For any $d \in \mathcal{M}$, the operator of behavioral metric (Castro et al. 2021; Zhang et al. 2020) is defined as:*

$$\mathcal{F}^\pi(d)(s_i, s_j) = \underbrace{|\bar{r}^\pi(s_i) - \bar{r}^\pi(s_j)|}_{\textit{reward difference}} + \underbrace{\gamma\, \mathcal{D}(d; P^\pi_{s_i}, P^\pi_{s_j})}_{\textit{distribution divergence}} \tag{2}$$

where the expected reward $\bar{r}^\pi(s) = \mathbb{E}_{a\sim\pi(\cdot\,|\,s)}[r(s,a)]$, the policy–induced transition distribution is defined as $P^\pi_s = \mathbb{E}_{a\sim\pi(\cdot\,|\,s)}[P(\cdot\,|\,s,a)]$ and $\mathcal{D}$ is a divergence between the two next–state distributions that may itself depend on d.

The core idea is to learn an embedding space in which distances reflect behavioral dissimilarity: states leading to different long-term behaviors are mapped farther apart. As such, the agent will explore different behaviorally distinct actions to form the basis of modeled synergy patterns.

Limitation in Overactuated Control. Recent behavioral metric, such as MICo (Castro et al. 2021) and DBC (Zhang et al. 2020) learn state encoders by minimizing the discrepancy between behavioral distances and latent distances:

$$\mathcal{L}(\phi) = \frac{1}{2}\mathbb{E}[(d^\pi_\phi(s_i, s_j) - \underbrace{|\bar{r}^\pi(s_i) - \bar{r}^\pi(s_j)|}_{\textit{reward difference}} - \underbrace{\gamma\, \mathcal{D}(d; P^\pi_{s_i}, P^\pi_{s_j})}_{\textit{distribution divergence}})]^2 \tag{3}$$

where $\mathcal{D}$ denotes the divergence between next-state distributions (e.g., Wasserstein in DBC, expected sample distance in MICo). While effective in dense-reward environments, these methods suffer from a critical failure under sparse rewards, especially in overactuated control tasks:

Theorem 1 (Representation Collapse)**.** *Let ξ denote a distribution over state pairs (s_i, s_j). Assume deterministic transitions and the existence of a stationary distribution. Then, for any behavioral metric d of the form (2), we have:*

$$\mathbb{E}_{(s_i,s_j)\sim\xi}[d^\pi_\phi(s_i, s_j)] = \frac{1}{1-\gamma}\mathbb{E}_{(s_i,s_j)\sim\xi}[|\bar{r}^\pi_{s_i} - \bar{r}^\pi_{s_j}|] \tag{4}$$

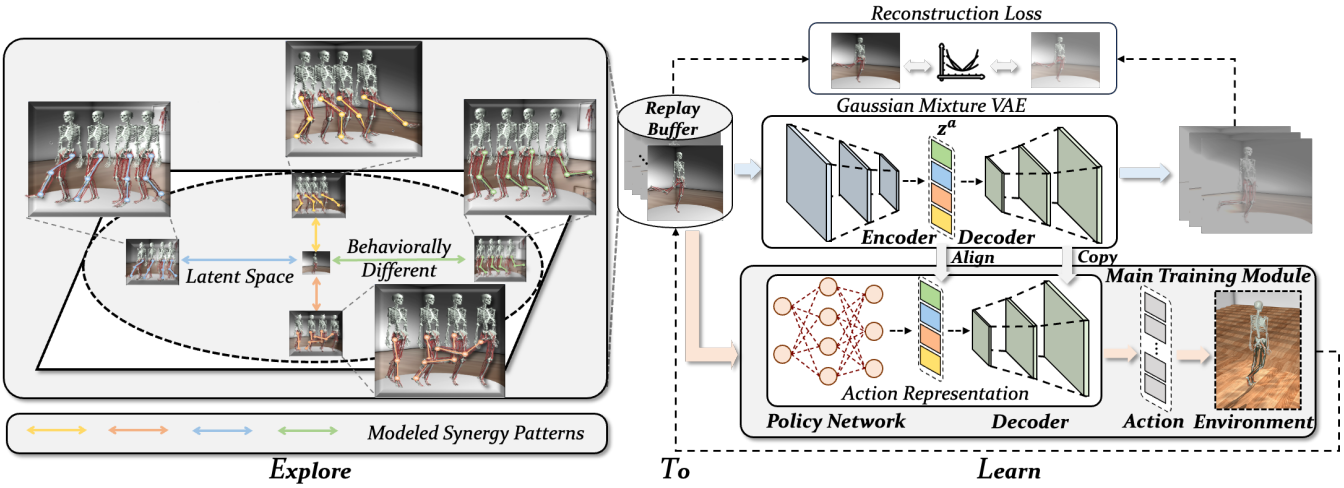

Figure 2: Overview of ETL (Explore to Learn) framework. ETL operates in two stages: **Explore** and **Learn**. In the exploration stage (left), a latent exploration policy is used to sample diverse synergy patterns, modeled as directions in a low-dimensional latent space. These patterns produce behaviorally distinct muscle activations, which are stored in the replay buffer. A Gaussian Mixture VAE (top right) is then trained to encode these trajectories into compact, disentangled synergy representations. In the learning stage (bottom right), a policy network operates in the learned latent space, outputting synergy coefficients that are decoded into full muscle activations. The decoder is further fine-tuned during policy training to enhance control expressiveness.

Under sparse rewards, the right-hand side of Eq. (4), i.e., $\mathbb{E}_{(s_i,s_j)\sim\xi}[|\bar{r}^\pi_{s_i} - \bar{r}^\pi_{s_j}|] \approx 0$*. Consequently, the learned embedding collapses to a constant value:* $\hat{d}_\phi(s_i) = d_{\phi_c}$.

Proof in Appendix A. This collapse renders the latent space behaviorally indistinguishable, severely limiting its utility in overactuated control settings, such as musculoskeletal tasks where nuanced distinctions are essential for meaningful exploration, highlighting a core limitation in behavior-driven representation learning under sparse feedback.

Behavioral Discriminability Regularization. To address the representation collapse issue, we introduce a *behavioral discriminability regularization* term that encourages diversity in the learned state embeddings even when reward signals are weak. Specifically, for each sampled state pair (s_i, s_j), we penalize latent contraction if the reward difference falls below a small threshold ϵ:

$$\mathcal{L}_{\text{reg}} = \mathbb{E}_{s_i,s_j}\left[\mathbb{I}\left[|\bar{r}^\pi_{s_i} - \bar{r}^\pi_{s_j}| < \epsilon\right] \cdot \|\phi(s_i) - \phi(s_j)\|^2\right] \quad (5)$$

where $\mathbb{I}[\cdot]$ denotes the indicator function. This loss prevents the encoder ϕ from mapping low-reward-difference state pairs to the same embedding unless they are behaviorally indistinguishable. We integrate this regularization into the behavioral metric objective as follows:

$$\begin{aligned}\mathcal{L}(\phi) = \frac{1}{2}\mathbb{E}_{s_i,s_j}[(d_\phi(s_i,s_j) - |\bar{r}^\pi_{s_i} - \bar{r}^\pi_{s_j}| \\ - \gamma\mathcal{D}_\phi(s_i,s_j))^2] + \lambda\mathcal{L}_{\text{reg}}\end{aligned} \quad (6)$$

where λ is a weighting coefficient. The first term enforces alignment with behavioral differences; the second promotes discriminability under sparse rewards.

Theorem 2 (Theoretical Guarantee)**.** *Adding* $\mathcal{L}_{reg}$ *to the behavioral metric loss preserves the original behavioral metric guarantees under deterministic dynamics.*

Proof provided in Appendix A. Specifically, we establish convergence and behavioral similarity guarantees for the resulting metric-based encoder, ensuring stable training. Empirically (see Section 5), this regularization significantly improves the latent space structure and facilitates the discovery of synergy patterns in overactuated control tasks.

4.2 Synergy-based Policy Learning

Structured Action Representation Learning. Overactuated control tasks (e.g., musculoskeletal systems) involve redundant and high-dimensional action spaces, where each action corresponds to coordinated activations of numerous actuators. Learning directly in such spaces poses significant challenges for both policy optimization and exploration. To address this, we aim to construct a structured latent action space that captures the underlying coordination patterns among actuators. Motivated by evidence that motor control arises from a low-dimensional set of muscle synergies, we use a Variational Autoencoder (VAE) (Kingma, Welling et al. 2013) to encode muscle activations into a compact latent space, where each dimension captures a distinct synergy pattern. However, standard VAEs struggle to model the multimodal nature of muscle activations. To overcome this, we adopt a Gaussian Mixture VAE, enabling more expressive and structured latent representations within our framework.

Unlike conventional VAEs with a unimodal Gaussian prior over the latent action variable $\mathbf{z^a}$, GMVAE adopts a mixture-of-Gaussians (Dilokthanakul et al. 2016) prior:

$$p(c) = \texttt{Cat}(c; P), \quad p(\mathbf{z^a}|c) = \mathcal{N}(\mathbf{z^a}; \mu_c, \Sigma_c) \quad (7)$$

where c is a categorical latent variable indicating the mixture of patterns, and P is a fixed uniform prior over d synergy patterns. The marginal prior over $\mathbf{z^a}$ becomes:

$$p(\mathbf{z^a}) = \sum_{c=1}^{d} P_c \mathcal{N}(\mathbf{z^a}; \mu_c, \Sigma_c) \tag{8}$$

This allows the latent space to explicitly represent d distinct clusters, each capturing a different mode of synergy pattern. During training, given a high-dimensional action $a \in \mathcal{A}$, the GMVAE encoder learns to infer the posterior over both the discrete component $q(c|a) = \texttt{Cat}(c; \psi(a))$ and the latent action variable $q(\mathbf{z^a}|a, c) = \mathcal{N}(\mathbf{z^a}; \mu_\psi(a, c), \Sigma_\psi(a, c))$. The model is trained by maximizing the following variational lower bound:

$$\begin{aligned} \mathcal{L}_{\text{GMVAE}}(a) = \ & \mathbb{E}_{q(c|a)} \Big[\mathbb{E}_{q(\mathbf{z^a}|a,c)} \big[\log p(a|\mathbf{z^a}) \big] \\ & - \text{KL}\big(q(\mathbf{z^a}|a, c) \| p(\mathbf{z^a}|c)\big) \Big] - \text{KL}\big(q(c|a) \| p(c)\big) \end{aligned} \tag{9}$$

Crucially, the use of GMVAE provides two advantages: (1) it promotes clustering of muscle activations into interpretable synergy modes, facilitating more efficient exploration; and (2) it improves representation disentanglement, leading to more stable and accurate decoding of muscle activations during downstream policy learning.

Decoder Fine-tuning and Synergy-aware Policy Learning. After unsupervised pretraining, we fix the GMVAE encoder and retain the decoder as a mapping from the latent synergy action space $\mathcal{Z}^a$ to the original action space $\mathcal{A}$. During downstream task learning, the policy $\pi(\mathbf{z_t^a}|s_t)$ operates entirely in the low-dimensional latent space, producing synergy coefficients $\mathbf{z_t^a}$ that are decoded into high-dimensional actions via $a_t = \texttt{Dec}(\mathbf{z_t^a})$. To accommodate task-specific coordination demands beyond those captured during exploration, we fine-tune the decoder alongside policy learning. While the policy adapts based on environmental rewards, the decoder is jointly updated to align latent actions with reward-relevant motor outputs. This is achieved via a supervised reconstruction loss:

$$\mathcal{L} = \mathbb{E}_{(\mathbf{z_t^a}, a_t) \sim \mathcal{D}} \left[\left| \texttt{Dec}_\psi(\mathbf{z_t^a}) - a_t \right|^2 \right] \tag{10}$$

where $(\mathbf{z_t^a}, a_t)$ pairs are collected during interaction. The encoder remains frozen to preserve latent consistency, enabling the policy to reason over stable, interpretable motor primitives. This synergy-aware architecture effectively decouples policy learning from high-dimensional actuation, reducing optimization complexity while enabling adaptive control. Empirically, this design improves convergence speed, training stability, and generalization across diverse tasks. We outline the full training procedure in Algorithm 1.

5 Experiments

The overall objective of our experiment is to evaluate the performance of ETL in comparison to other baselines, we conduct comprehensive experiments on various settings of locomotion and manipulation tasks with high-dimensional action space to assess the effectiveness of ETL (Project site: https://sites.google.com/view/explore-to-learn/).

Algorithm 1: Explore to Learn (ETL)

Initialize: exploration policy π_{exp}, task-specific policy π_θ, state encoder ϕ, number of synergy-patterns d, GMVAE encoder–decoder ψ, replay buffer $\mathcal{D}$

`Phase 1: Latent Exploration`

for episode $= 1$ to N_{explore} **do**
  Sample unit latent vector z_{sp} (synergy pattern)
  for $t = 1$ to MAX_STEP_PER_EPISODE **do**
    Sample action a_t guided by synergy-pattern z_{sp}
    Compute exploration reward F_t ▷Eq. (1)
    Record transition in the buffer $\mathcal{D}$
    Update exploration policy π_{exp} via policy gradient
    Train GMVAE $(\texttt{Enc}_\psi, \texttt{Dec}_\psi)$ on collected action data in $\mathcal{D}$ ▷Eq. (9)
  end for
end for

`Phase 2: Policy Learning`

while *not converged* **do**
  for $t = 1$ to MAX_STEP_PER_EPISODE **do**
    Observe state s_t
    Sample synergy latent $\mathbf{z_t^a} \sim \pi_\theta(\cdot \mid s_t)$
    Decode action $a_t = \texttt{Dec}_\psi(\mathbf{z_t^a})$
    Update π_θ using policy gradient
    Fine-tune decoder $\texttt{Dec}_\psi$ using collected $(\mathbf{z_t^a}, a_t)$ pairs to minimize the reconstruction loss ▷Eq. (10)
  end for
end while

Figure 3: Five high-dimensional tasks from the DMControl suite, with respective action dimensions: Humanoid Run ($a \in \mathbb{R}^{24}$), Humanoid Stand ($a \in \mathbb{R}^{24}$), Humanoid Walk ($a \in \mathbb{R}^{24}$), Dog Run ($a \in \mathbb{R}^{38}$), and Dog Walk ($a \in \mathbb{R}^{38}$).

Baselines. We compare ETL against four representative methods designed for high-dimensional control or latent exploration: (1) Lattice (Chiappa et al. 2023): A latent exploration approach that injects temporally correlated noise into the latent state of the policy to improve exploration in high-dimensional control tasks. (2) EME (Wang et al. 2024): The latest latent exploration technique that leverages state differences in a learned metric space to drive more effective and structured exploration. (3) DEP (Schumacher et al. 2023): A method that integrates differential extrinsic plasticity (DEP) into reinforcement learning, enabling rapid, state-space-covering exploration in musculoskeletal models within seconds of interaction. (4) DynSyn (He et al. 2024): A recent approach that derives synergy-based representations from dynamical structures and adapts them in a task-specific, state-dependent manner to enhance motor control.

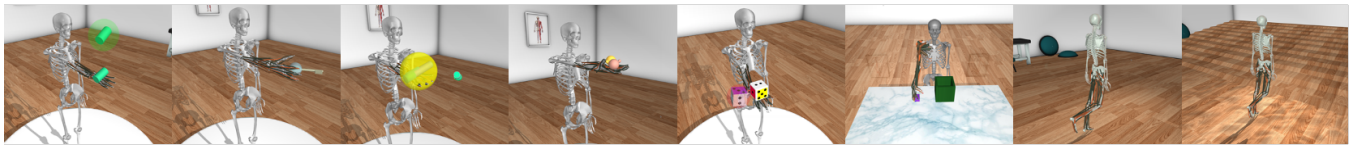

Figure 4: Eight musculoskeletal control tasks from MyoSuite, with corresponding action dimensions: Hand Pose ($a \in \mathbb{R}^{39}$), Key Turn ($a \in \mathbb{R}^{39}$), Pen Twirl ($a \in \mathbb{R}^{39}$), Baoding ($a \in \mathbb{R}^{39}$), Die Rotation ($a \in \mathbb{R}^{39}$), Relocate ($a \in \mathbb{R}^{63}$), Leg Walk ($a \in \mathbb{R}^{80}$), and Stair-Terrain-Walk ($a \in \mathbb{R}^{80}$).

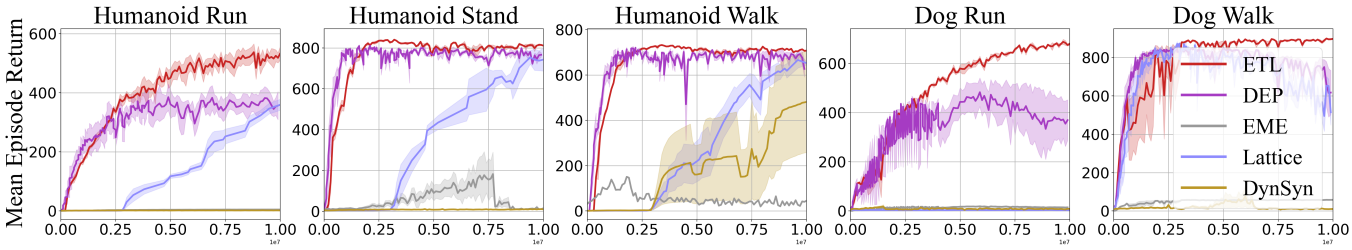

Figure 5: Comparison results for the tasks from DMControl suite (Humanoid and Dog), x axis denotes the timesteps.

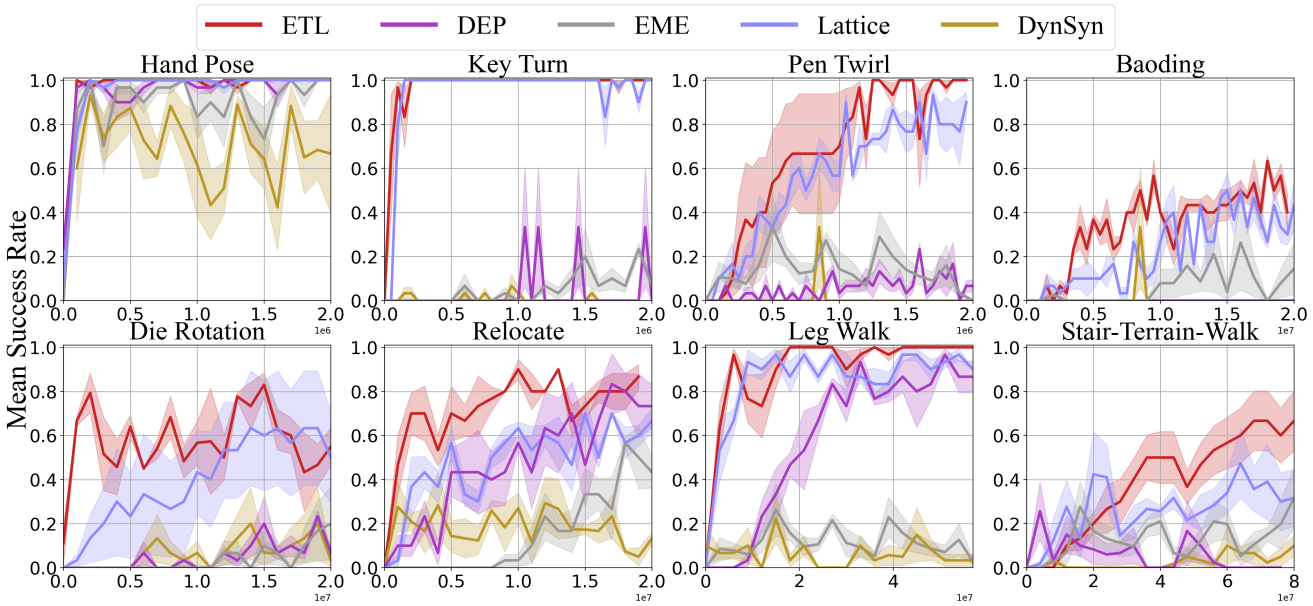

Figure 6: Comparison results for the tasks in Myosuite (averaged over 5 seeds), x axis denotes the timesteps.

5.1 Experimental Setting

DMControl. We evaluate the effectiveness of ETL on five high-dimensional locomotion tasks from the DMControl Suite (Tassa et al. 2018): Humanoid Run, Humanoid Stand, Humanoid Walk, Dog Run and Dog Walk. As shown in Fig. 3, the action space dimensionality is 38 for the Dog embodiment and 24 for the Humanoid embodiment. These tasks are widely used to benchmark policy learning in continuous control settings with complex dynamics.

Myosuite. To assess the applicability of ETL in physiologically realistic motor control scenarios, we further evaluate it on eight challenging tasks from the MyoSuite benchmark (Caggiano et al. 2022a), a physics-based musculoskeletal simulation environment. MyoSuite emphasizes fine-grained biomechanical coordination and high-dimensional actuation. As illustrated in Fig. 4, we focus on two musculoskeletal models: **MyoHand** ($a \in \mathbb{R}^{39}$) and **MyoLeg** ($a \in \mathbb{R}^{80}$). We include five representative tasks from the default MyoSuite suite: Hand Pose, Key Turn, Pen Twirl, Leg Walk, and Stair-Terrain-Walk, which test dexterous manipulation and coordinated locomotion. Additionally, we consider three advanced tasks drawn from the NeurIPS MyoChallenge benchmarks (Caggiano et al. 2022b, 2023) designed to promote human-level dexterity and agility: Baoding, Die Rotation, and Relocate. These tasks involve dynamic object manipulation and pose significant challenges

for exploration and motor control. Full task specifications and experimental settings are provided in Appendix C.

5.2 Experimental Results and Analysis

Overall Performance. As shown in Fig. 5, ETL consistently outperforms all baselines on high-dimensional DMControl tasks, demonstrating superior sample efficiency and learning stability. While DEP and Lattice perform comparably on simpler tasks (e.g., Humanoid Stand, Walk), their performance degrades in more complex Dog tasks, where ETL's advantage becomes more pronounced.

In MyoSuite tasks (Fig. 6), ETL significantly outperforms all baselines, particularly on challenging manipulation tasks like *Baoding* and *Die Rotation*, which require precise coordination and adaptive exploration. Bonus-based (e.g., EME) and noise-driven (e.g., Lattice) methods fail to scale in high-dimensional action spaces, while DEP and DynSyn suffer from local optima or rigid priors. In contrast, ETL autonomously discovers disentangled synergy patterns that generalize across tasks without manual priors. Latent exploration yields structured trajectories that enable the GMVAE to learn compact, transferable action representations. As a result, ETL achieves the best performance on 6 of 8 tasks and matches top scores on the remaining two, demonstrating superior efficiency in overactuated control.

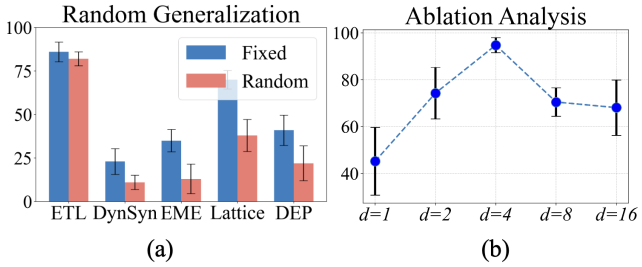

Figure 7: (a) Average performance across all Myosuite tasks under fixed and random goal settings. (b) Ablation study on the number of synergy patterns d in Leg Walk locomotion task, averaged over five seeds.

Random Goal Generalization. To evaluate the generalization capability of ETL, we consider a more challenging setting in which task goals are randomized during both training and evaluation in the MyoSuite environment. As illustrated in Fig. 7(a), ETL exhibits strong generalization performance under this setting. Notably, its performance remains nearly unchanged compared to the fixed-goal scenario, demonstrating robust adaptability to varying task objectives. In contrast, other baseline methods suffer a substantial degradation in performance, often exceeding a 50% drop, highlighting their limited capacity to handle goal variability. These results indicate that ETL's synergy-based representation enables effective transfer across diverse settings, further validating its scalability and robustness in realistic control scenarios.

Ablation Study. We conduct ablation studies to evaluate the effect of two key components in the ETL framework: (1) the number of synergy patterns d and (2) the contributions of

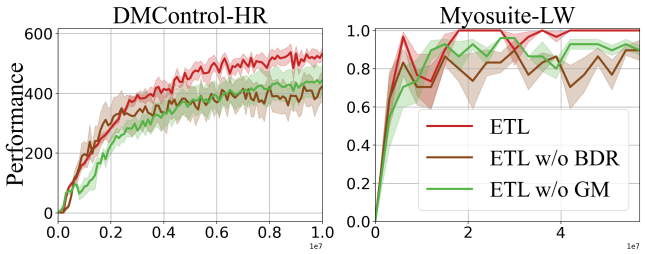

Figure 8: Comparison results between ETL and its variants in Humanoid Run (HR) and Leg Walk (LW) tasks.

individual architectural components. To analyze the impact of the number of synergy patterns, as suggested by neuroscience studies(Caggiano, Cheung, and Bizzi 2016), we vary d across $1, 2, 4, 8, 16$ and report results on the *Leg Walk* task in Fig.7(b). In our experiments, we observe that performance improves as d increases from 1 to 4, indicating that a richer set of synergy patterns enables the model to capture coordinated muscle activations more accurately. However, performance degrades as d increases further to 8 and 16. This suggests that overly large synergy spaces may introduce representational noise or redundancy, hindering policy learning and leading to overfitting.

We further investigate the contributions of two critical components in ETL by introducing the following ablated variants: (1) **ETL w/o BDR**: removes the Behavioral Discriminability Regularization from the encoder loss. (2) **ETL w/o GM**: replaces the Gaussian Mixture VAE with a standard unimodal VAE. As shown in Fig. 8, removing BDR results in a noticeable performance drop, confirming its role in preventing representation collapse in sparse reward settings. The performance declines even more when the Gaussian Mixture prior is removed, highlighting the necessity of capturing multi-modal coordination patterns through a structured latent space. Together, these results validate that both the behavioral discriminability regularization and the Gaussian Mixture module are essential for learning meaningful representations and achieving robust performance. Additional ablation results are provided in Appendix B.

6 Conclusion

In this work, we propose Explore to Learn (ETL), a two-stage framework for efficient learning in high-dimensional, overactuated control tasks. In the first stage, ETL discovers muscle synergy patterns by guiding exploration along structured latent directions, enabling the collection of diverse muscle activation trajectories. A variational autoencoder is then trained to encode actions into a compact synergy-based latent space. In the second stage, the policy operates entirely in this space, producing synergy coefficients that are decoded into full-dimensional actions. The decoder is further fine-tuned to enhance expressiveness and generalization. Experiments across musculoskeletal and locomotion benchmarks show that ETL consistently outperforms strong baselines in both sample efficiency and final performance.

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

Al Borno, M.; Hicks, J. L.; and Delp, S. L. 2020. The effects of motor modularity on performance, learning and generalizability in upper-extremity reaching: a computational analysis. *Journal of the Royal Society Interface*, 17(167): 20200011.

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*.

Balda, R. P.; Pepperberg, I. M.; and Kamil, A. C. 1998. *Animal cognition in nature: the convergence of psychology and biology in laboratory and field*. Academic Press.

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.

Bizzi, E.; Mussa-Ivaldi, F. A.; and Giszter, S. 1991. Computations underlying the execution of movement: a biological perspective. *Science*, 253(5017): 287–291.

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

Burda, Y.; Edwards, H.; Storkey, A.; and Klimov, O. 2018b. Exploration by random network distillation. *arXiv preprint arXiv:1810.12894*.

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

Caggiano, V.; Cheung, V. C.; and Bizzi, E. 2016. An optogenetic demonstration of motor modularity in the mammalian spinal cord. *Scientific reports*, 6(1): 35185.

Caggiano, V.; Wang, H.; Durandau, G.; Sartori, M.; and Kumar, V. 2022a. MyoSuite–A contact-rich simulation suite for musculoskeletal motor control. *arXiv preprint arXiv:2205.13600*.

Caggiano, V.; Wang, H.; Durandau, G.; Song, S.; Tassa, Y.; Sartori, M.; and Kumar, V. 2022b. MyoChallenge: Learning contact-rich manipulation using a musculoskeletal hand. https://sites.google.com/view/myochallenge.

Caggiano, V.; Wang, H.; Durandau, G.; Song, S.; Tassa, Y.; Sartori, M.; and Kumar, V. 2023. MyoChallenge 23: Towards Human-Level Dexterity and Agility. https://sites.google.com/view/myosuite/myochallenge/myochallenge-2023.

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

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*.

Chiappa, A. S.; Marin Vargas, A.; Huang, A.; and Mathis, A. 2023. Latent exploration for reinforcement learning. *Advances in Neural Information Processing Systems*, 36: 56508–56530.

Dilokthanakul, N.; Mediano, P. A.; Garnelo, M.; Lee, M. C.; Salimbeni, H.; Arulkumaran, K.; and Shanahan, M. 2016. Deep unsupervised clustering with gaussian mixture variational autoencoders. *arXiv preprint arXiv:1611.02648*.

Dominici, N.; Ivanenko, Y. P.; Cappellini, G.; d'Avella, A.; Mondì, V.; Cicchese, M.; Fabiano, A.; Silei, T.; Di Paolo, A.; Giannini, C.; et al. 2011. Locomotor primitives in newborn babies and their development. *Science*, 334(6058): 997–999.

Hacques, G.; Komar, J.; Dicks, M.; and Seifert, L. 2021. Exploring to learn and learning to explore. *Psychological Research*, 85(4): 1367–1379.

He, K.; Zuo, C.; Ma, C.; and Sui, Y. 2024. DynSyn: dynamical synergistic representation for efficient learning and control in overactuated embodied systems. In *Proceedings of the 41st International Conference on Machine Learning*, 18115–18132.

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

Jiang, Y.; Van Wouwe, T.; De Groote, F.; and Liu, C. K. 2019. Synthesis of biologically realistic human motion using joint torque actuation. *ACM Transactions On Graphics (TOG)*, 38(4): 1–12.

Joos, E.; Péan, F.; and Goksel, O. 2020. Reinforcement learning of musculoskeletal control from functional simulations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 135–145. Springer.

Kingma, D. P.; Welling, M.; et al. 2013. Auto-encoding variational bayes.

Luo, S.; Androwis, G.; Adamovich, S.; Nunez, E.; Su, H.; and Zhou, X. 2023. Robust walking control of a lower limb rehabilitation exoskeleton coupled with a musculoskeletal model via deep reinforcement learning. *Journal of neuroengineering and rehabilitation*, 20(1): 34.

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.

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.

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.; Rocktäschel, T.; et al. 2020. Ride: Rewarding impact-driven exploration for procedurally-generated environments. *arXiv preprint arXiv:2002.12292*.

Schumacher, P.; Haeufle, D. F. B.; Büchler, D.; Schmitt, S.; and Martius, G. 2023. DEP-RL: Embodied Exploration for Reinforcement Learning in Overactuated and Musculoskeletal Systems. In *ICLR*. OpenReview.net.

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.

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.

Sutton, R. S.; Barto, A. G.; et al. 1998. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge.

Tahami, E.; Jafari, A. H.; and Fallah, A. 2014. Learning to control the three-link musculoskeletal ARM using actor–critic reinforcement learning algorithm during reaching movement. *Biomedical Engineering: Applications, Basis and Communications*, 26(05): 1450064.

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.

Tassa, Y.; Doron, Y.; Muldal, A.; Erez, T.; Li, Y.; Casas, D. d. L.; Budden, D.; Abdolmaleki, A.; Merel, J.; Lefrancq, A.; et al. 2018. Deepmind control suite. *arXiv preprint arXiv:1801.00690*.

Tieck, J. C. V.; Pogančić, M. V.; Kaiser, J.; Roennau, A.; Gewaltig, M.-O.; and Dillmann, R. 2018. Learning continuous muscle control for a multi-joint arm by extending proximal policy optimization with a liquid state machine. In *International Conference on Artificial Neural Networks*, 211–221. Springer.

Wainwright, P. C. 2002. The evolution of feeding motor patterns in vertebrates. *Current Opinion in Neurobiology*, 12(6): 691–695.

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.; Li, Y.; and U, L. H. 2025. BILE: an effective behavior-based latent exploration scheme for deep reinforcement learning. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, 6497–6505.

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*.

Yang, M.; Dong, R.; Wang, Y.; Liu, F.; Du, Y.; Zhou, M.; and Hou U, L. 2023a. TieComm: learning a hierarchical communication topology based on Tie theory. In *International Conference on Database Systems for Advanced Applications*, 604–613. Springer.

Yang, M.; Wang, Y.; Yu, Y.; Zhou, M.; et al. 2023b. MixLight: Mixed-agent cooperative reinforcement learning for traffic light control. *IEEE Transactions on Industrial Informatics*, 20(2): 2653–2661.

Yang, M.; Zhao, K.; Wang, Y.; Dong, R.; Du, Y.; Liu, F.; Zhou, M.; and U, L. H. 2024. Team-wise effective communication in multi-agent reinforcement learning. *Autonomous Agents and Multi-Agent Systems*, 38(2): 36.

Zhang, A.; McAllister, R. T.; Calandra, R.; Gal, Y.; and Levine, S. 2020. Learning Invariant Representations for Reinforcement Learning without Reconstruction. *ArXiv*, abs/2006.10742.

Zhang, T.; Xu, H.; Wang, X.; Wu, Y.; Keutzer, K.; Gonzalez, J. E.; and Tian, Y. 2021. 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*.

Zhao, K.; Wen, H.; Zhang, Z.; Atzori, M.; Müller, H.; Xie, Z.; and Scano, A. 2022. Evaluation of methods for the extraction of spatial muscle synergies. *Frontiers in neuroscience*, 16: 732156.