RETHINKING UNSUPERVISED SKILL DISCOVERY: A TRAJECTORY-LEVEL PERSPECTIVE

Anonymous authors
Paper under double-blind review

ABSTRACT

Unsupervised skill discovery enables reinforcement learning agents to acquire reusable behaviors without relying on laborious, task-specific reward engineering. However, existing MI-based approaches typically discriminate skills at the level of individual states. Consequently, this state-centric formulation often forces skills to partition the environment and specialize in disjoint regions, preventing agents from learning temporally extended behaviors that generalize across varying initial conditions and making downstream skill chaining highly challenging. In this work, we propose **TIME (Trajectory-level Information Maximization and Exploration)**, a novel framework that discovers reusable skills by modeling the temporal evolution of agent-generated trajectories rather than isolated state occupancy. TIME introduces a conditional MI objective that allows all skills to access the full state space while encouraging them to induce distinguishable long-horizon behaviors. Furthermore, to overcome the limited state coverage of pure MI methods, TIME incorporates a simple yet effective Lipschitz-constrained training mechanism that prevents the exploitation of trivial differences. Evaluated across a diverse set of challenging continuous-control environments, TIME learns coherent and highly reusable behaviors while outperforming prior unsupervised skill discovery methods, yielding substantial performance gains on downstream tasks through hierarchical skill chaining. Code is available at `https://anonymous.4open.science/r/TIME-32C3`.

1 INTRODUCTION

Reinforcement learning (RL) has achieved remarkable success across a range of challenging domains, including magnetic control Tracey et al. (2024), autonomous racing Song et al. (2023), and robotics Tang et al. (2025). Despite these advances, RL systems remain heavily dependent on meticulously engineered reward functions-an often laborious, brittle, and inherently task-specific process.To address this limitation, unsupervised skill discovery Eysenbach et al. (2019); Zhang et al. (2025); Hu et al. (2024) has emerged as a promising direction by learning reusable behaviors without relying on external rewards. In this setting, an agent is encouraged to interact with the environment to produce diverse behaviors and discover a set of skills that can later be reused to accelerate learning across a variety of downstream tasks when combined with Hierarchical Reinforcement Learning (HRL) Pateria et al. (2021) or skill-selection mechanisms.

A widely adopted approach to unsupervised skill discovery is to maximize the Mutual Information (MI) Walters-Williams & Li (2009) between the skill variable and the environment's states Laskin et al. (2022); Liu & Abbeel (2021); Eysenbach et al. (2019); Hu et al. (2024). This objective encourages each skill to produce behaviors that lead to distinguishable states, thereby increasing behavioral diversity across skills. Despite their empirical success, a key limitation of these methods is that skill discrimination is primarily defined at the level of individual states rather than at the level of evolving behaviors. As a result, skills often specialize in disjoint regions of the state space, effectively partitioning the environment into separate areas. While this improves distinguishability between skills, it does not necessarily capture dynamically meaningful behaviors that generalize across different starting conditions.

This limitation raises a fundamental question: *what makes a skill useful for control-the states it reaches, or the way it evolves?* Consider navigation in a maze environment. Reaching a target

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

requires executing a sequence of coordinated actions that progressively move the agent toward the goal. However, when skills are tied to fixed state regions, there is no guarantee that a single skill can reliably guide the agent from varying initial states to the desired destination. This mismatch limits the controllability and reuse of learned skills (as illustrated in Figure 1). In contrast, meaningful behaviors are naturally defined by their temporal evolution-for example, moving toward an exit regardless of the starting position-suggesting that skills should be characterized by how states evolve, rather than by isolated state occupancy.

Motivated by this observation, we introduce **TIME (Trajectory-level Information Maximization and Exploration)**, a novel framework for discovering reusable skills by modeling the structure of agent-generated trajectories. The key principle of TIME is to promote diversity over future trajectories induced by different skills, while allowing all skills to access the entire state space, rather than restricting them to disjoint regions. As a result, skills are no longer defined by where they operate, but by how they influence the temporal evolution of states. To realize this idea, TIME redefines the unsupervised skill discovery objective as maximizing the conditional MI between the skill variable and future trajectories given the current state. Theoretically, this objective can be interpreted as jointly optimizing two complementary goals: (i) encouraging different skills to induce distinguishable future behaviors, and (ii) discouraging skills from being tied to isolated regions of the state space. This objective is implemented via a novel intrinsic reward mechanism, supported by a theoretical analysis based on a tractable variational lower bound of the conditional MI.

In addition, prior work Park et al. (2022); Zhang et al. (2025); Park et al. (2024) has shown that MI-based methods suffer from insufficient state coverage, as distinguishability does not imply broad exploration. Existing approaches introduce alternative objectives, such as metric-aware MI Zhang et al. (2025); Park et al. (2024) or direction-based exploration Park et al. (2022), but deviate from the original MI objective, do not address its state-centric limitation, and often incur more complex training procedures. In contrast, the trajectory-centric objective proposed in our work is, under the conditional MI formulation, theoretically incompatible with these approaches. To unlock the full potential of TIME, a simple yet effective Lipschitz-constrained training mechanism is introduced. By regulating the skill discriminator's sensitivity, this mechanism prevents the model from exploiting trivial differences, thereby improving state coverage without altering the objective or requiring more complex formulations.

In our experiments, we evaluate TIME on four continuous-control benchmarks: `Ant`, `HalfCheetah`, `Humanoid`, and `Quadruped`. Across all environments, TIME consistently discovers diverse and semantically meaningful behaviors while maintaining state-agnostic skill representations, thereby overcoming a key limitation of prior unsupervised RL approaches. Moreover, the learned skills improve downstream task performance by an average of **37.3%** relative to strong baselines.

2 PRELIMINARIES

Trajectory-based Markov Decision Process (t-MDP) Following prior work Zhang et al. (2025) and to align with our trajectory-centric perspective, we formulate unsupervised skill discovery in a trajectory-based Markov Decision Process (t-MDP) defined as $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, P, \mu_0, \mathcal{Z} \rangle$, where $\mathcal{S} \subseteq \mathbb{R}^n$ and $\mathcal{A} \subseteq \mathbb{R}^m$ denote the state and action spaces, $P(s' \mid s, a)$ is the transition dynamics, $\mu_0(s)$ is the initial state distribution, and $\mathcal{Z} \subseteq \mathbb{R}^d$ is a latent skill space with prior $p(z)$. A skill-conditioned policy $\pi(a \mid s, z)$ induces a trajectory distribution $p_\pi(\tau \mid z) = \mu_0(s_0) \prod_{t=0}^{T-1} \pi(a_t \mid s_t, z) P(s_{t+1} \mid s_t, a_t)$, where $\tau = \{(s_t, a_t, s_{t+1})\}_{t=0}^{T-1}$. In this setting, we omit external rewards and instead rely on intrinsic objectives. In contrast to conventional state-centric formulations that define intrinsic rewards over individual states, we consider trajectory-dependent intrinsic rewards to capture temporally extended behaviors, i.e., $r(s_{t:t+k}, z)$ with $r : \mathcal{S}^k \times \mathcal{Z} \to \mathbb{R}$, where $s_{t:t+k}$ denotes the k-step future trajectory segment induced by executing skill z from state s_t. The objective of unsupervised skill discovery is thus to learn a policy that maximizes the expected discounted return over skill-conditioned rollouts, $J(\pi) = \mathbb{E}_{z \sim p(z)} \mathbb{E}_{\tau \sim p_\pi(\tau|z)} \left[\sum_{t=0}^{T-1} \gamma^t r(s_{t:t+k}, z) \right]$, where $\gamma \in [0, 1)$ is the discount factor.

MI-Based Skill Discovery MI-based skill discovery methods, such as DIAYN Eysenbach et al. (2019) and related approaches Laskin et al. (2022); Liu & Abbeel (2021); Zhang et al. (2025),

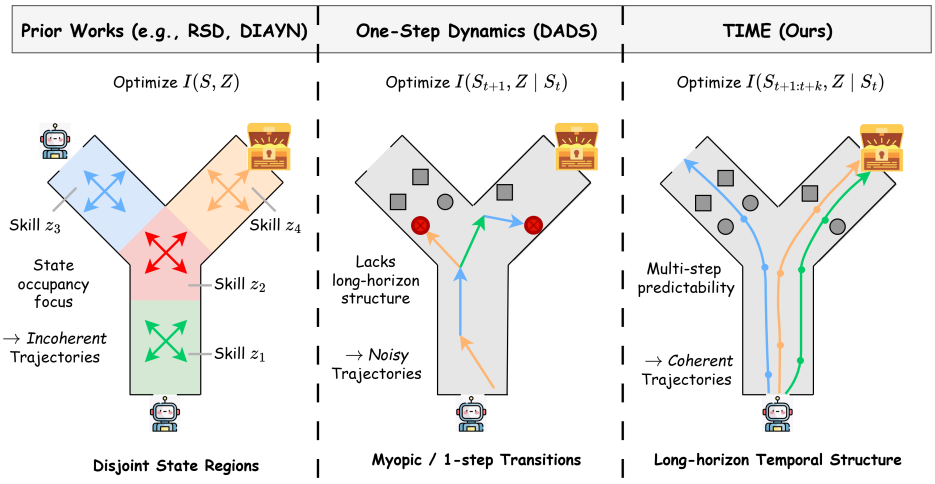


Figure 1: **Motivation for long-horizon skill learning in a maze with obstacles. Left (Prior MI):** Methods partition the state space, meaning no single skill covers the full goal traversal. **Middle (DADS Sharma et al. (2020)):** One-step dynamics lead to noisy, inconsistent trajectories that lack long-horizon structure. **Right (TIME):** By learning skills over multi-step future trajectories, our method achieves coherent, long-horizon behaviors effective for downstream tasks.

aim to learn diverse behaviors by maximizing the mutual information between states and skills, $I(S; Z) = H(Z) - H(Z \mid S)$, where $H(\cdot)$ denotes entropy. In practice, this is optimized via a tractable *variational lower bound* using a learned discriminator $q(z \mid s)$ to approximate the true posterior $p(z \mid s)$, yielding $I(S; Z) \geq \mathbb{E}_{s,z}\left[\log q(z \mid s) - \log p(z)\right]$. While recent works introduce alternative objectives Park et al. (2022) or metric-aware extensions Park et al. (2024); Zhang et al. (2025) to improve state coverage, these formulations remain fundamentally state-centric, focusing on spatial partitioning rather than temporal structures. In contrast, our work adopts a trajectory-centric perspective by modeling dependencies between skills and future state sequences, which enables capturing temporally extended behavioral structure.

3 Methodology

Key ideas. Building on the observation that meaningful skills are better characterized by the temporal evolution they induce rather than by the states they occupy, TIME formulates unsupervised skill discovery at the trajectory level. Specifically, TIME learns latent behaviors that generate distinguishable future trajectories from shared initial conditions, enabling skills to be defined by how they transform the environment over time rather than by localized state occupancy. To achieve this, we optimize a **trajectory-centric mutual information objective** (Equation 1) that emphasizes multi-step temporal dynamics over instantaneous spatial distinctions.

Our framework proceeds in two stages: (i) **Unsupervised Discovery (Section 3.1)**: optimizing a tractable variational lower bound of the MI objective via both trajectory- and state-centric discriminators; and (ii) **Downstream Adaptation**: leveraging the learned library for hierarchical reinforcement learning. Notably, to enhance state coverage, we introduce a **Lipschitz-constrained training mechanism** (Section 3.2) for the trajectory discriminator. This prevents the exploitation of trivial distinctions and promotes broad exploration without altering the core objective. The pipeline and pseudocode are detailed in Figure 2 and Algorithm 1.

3.1 Trajectory-Centric Skill Discovery

In contrast to state-centric formulations, we propose a trajectory-centric objective that maximizes the mutual information between a future trajectory segment and the current skill, conditioned on the

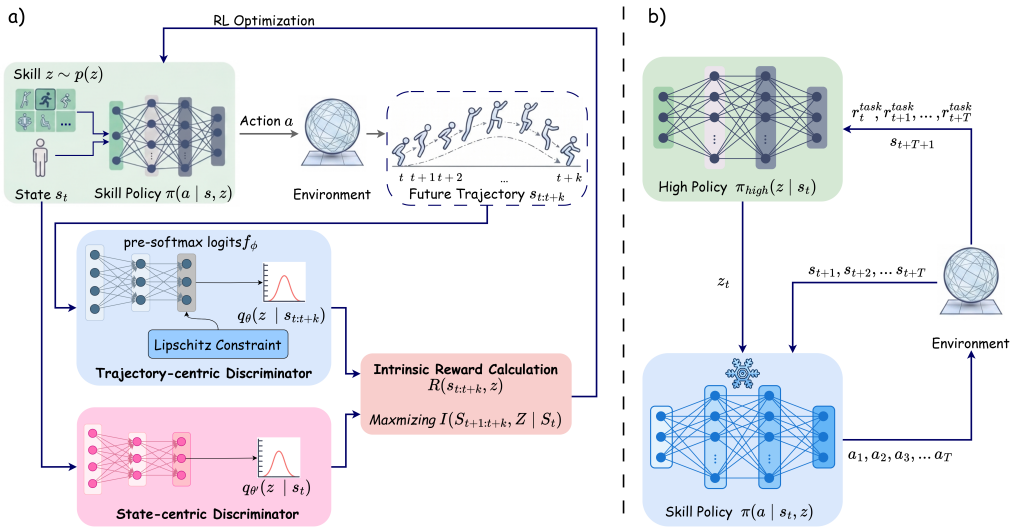


Figure 2: **Overview of the TIME framework.** In unsupervised skill discovery, trajectory- and state-centric discriminators are jointly used to construct an intrinsic reward for learning the skill policy, with the former regularized by a Lipschitz constraint. In the downstream phase, HRL leverages the pretrained skills via a high-level policy for task optimization.

present state:

$$I(S_{t+1:t+k}; Z \mid S_t) = H(Z \mid S_t) \uparrow - H(Z \mid S_{t:t+k}) \downarrow$$
$$= \mathbb{E}_{s_{t:t+k}, z}\left[\log \frac{p(s_{t+1:t+k}, z \mid s_t)}{p(s_{t+1:t+k} \mid s_t) p(z \mid s_t)}\right]. \tag{1}$$

Specifically, maximizing this objective induces two complementary effects: **1)** Minimizing $H(Z \mid S_{t:t+k})$ encourages the future trajectory segment to be highly informative about the skill, ensuring that different skills induce distinguishable long-horizon behaviors. In other words, the executed trajectory should reveal which skill is being followed. **2)** Maximizing $H(Z \mid S_t)$ promotes skill diversity conditioned on the current state, discouraging skills from collapsing to state-dependent behaviors. This reduces the dependence between individual states and skills, allowing multiple skills to originate from the same state.

Intuitively, skills are represented as temporally extended patterns of behavior, characterized by the trajectories they generate rather than isolated states. However, exact computation of mutual information in this continuous control setting remains intractable. In particular, Equation 1 involves components that are not directly accessible in RL settings, such as the transition distribution $p(s_{t+1:t+k} \mid s_t)$. To address this, we instead optimize a tractable surrogate, which is a *variational lower bound* of the original objective. This auxiliary objective can be computed using learned models, and maximizing it provides a principled approximation to maximizing the true MI. Specifically, we derive the following variational lower bound:

$$I_{\text{TIME}}(S_{t+1:t+k}; Z \mid S_t) = \mathbb{E}_{s_{t:t+k}, z}\left[\log \frac{q_\theta(z \mid s_{t:t+k})}{p(z)\, q_{\theta'}(z \mid s_t)}\right] + \mathbb{E}_{s_t}\mathbb{E}_z\left[\log q_{\theta'}(z \mid s_t)\right]. \tag{2}$$

Here, $q_\theta(z \mid s_{t:t+k})$ denotes a trajectory-based skill discriminator that approximates the posterior $p(z \mid s_{t:t+k})$, while $q_{\theta'}(z \mid s_t)$ is a state-based discriminator that approximates $p(z \mid s_t)$. This variational formulation yields a tractable surrogate objective for the original MI. In the following, we formally show that Equation 2 defines a lower bound on $I(S_{t+1:t+k}; Z \mid S_t)$, and characterize the theoretical conditions under which this bound becomes tight.

Theorem 1 (Trajectory MI Lower Bound Characterization)**.** *Let $q_\theta(z \mid s_{t:t+k})$ and $q_{\theta'}(z \mid s_t)$ be variational distributions, and define*

$$q_{\theta'}(z, s_t) := q_{\theta'}(z \mid s_t)\, p(s_t), \quad q_\theta(z, s_{t:t+k}) := q_\theta(z \mid s_{t:t+k})\, p(s_{t:t+k}).$$

Then,

$$I(S_{t+1:t+k}; Z \mid S_t) \;\geq\; I_{\mathit{TIME}}(S_{t+1:t+k}; Z \mid S_t) \tag{3}$$

if and only if

$$\begin{aligned} &+ \mathrm{KL}\big(p(z)p(s_t) \,\|\, q_{\theta'}(z, s_t)\big) \\ &- \mathrm{KL}\big(p(z, s_t) \,\|\, q_{\theta'}(z, s_t)\big) \\ &+ \mathrm{KL}\big(p(z, s_{t:t+k}) \,\|\, q_{\theta}(z, s_{t:t+k})\big) \;\geq\; 0. \end{aligned} \tag{4}$$

Discussion. Theorem 1 characterizes exactly when the tractable objective is a lower bound. The condition is favored when both discriminators accurately approximate their respective posteriors. Appendix A gives the proof and two useful consequences: Corollary 1 bounds the approximation gap by the sum of the variational errors, while Corollary 2 shows that the bound is exact when the trajectory posterior is recovered and $I(S_t; Z) = 0$, the ideal state-agnostic setting.

Implementation. To optimize the proposed variational objective I_{TIME}, we interpret it as an intrinsic reward and use it to guide policy learning. Specifically, we define the reward associated with a trajectory segment $s_{t:t+k}$ and skill z as

$$r(s_{t:t+k}, z) = \log \frac{q_\theta(z \mid s_{t:t+k})}{p(z)\, q_{\theta'}(z \mid s_t)} + \mathbb{E}_z\big[\log q_{\theta'}(z \mid s_t)\big]. \tag{5}$$

In the case where the prior over skills is uniform, i.e., $p(z) = \frac{1}{N}$ over a discrete set of N latent skills $\{z_i\}_{i=1}^N$, the reward can be rewritten as

$$r(s_{t:t+k}, z) = \log q_\theta(z \mid s_{t:t+k}) - \log q_{\theta'}(z \mid s_t) + \frac{1}{N}\sum_{i=1}^{N} \log q_{\theta'}(z_i \mid s_t) + \log N. \tag{6}$$

Overall, this reward encourages representations that are informative about future trajectories while being minimally dependent on the current state, aligning with the objective of maximizing conditional mutual information.

3.2 Lipschitz-Constrained Discriminator

Standard MI objectives often suffer from limited state coverage, as agents tend to exploit trivial local variations to maximize skill discriminability. While prior works address this via metric-aware modifications Park et al. (2024); Zhang et al. (2025) or objective replacement Park et al. (2022), these approaches either deviate from the MI formulation or require careful metric design. Moreover, as discussed in Appendix A, our proposed trajectory-centric MI objective is not readily compatible with such metric-based formulations.

In contrast, to unlock the full potential of TIME, we propose to mitigate this limitation at the level of the trajectory-centric discriminator q_ϕ, without introducing additional reward terms. Our key idea is to enforce smoothness in the trajectory-centric discriminator, such that similar trajectories yield similar posterior distributions over skills. Intuitively, if two trajectory segments are nearly identical, the trajectory-centric should not confidently assign them to different skills. Conversely, to achieve a high reward, the agent must generate trajectories that are sufficiently distinct, i.e., well separated in the state space. This shifts the burden of exploration from explicit reward design to the trajectory-centric discriminator's inductive bias. To instantiate this principle, we enforce a *Lipschitz constraint* directly on the pre-softmax logits of the discriminator, avoiding distortions introduced by the softmax operation. Let $f_\phi(s_{t:t+k}) \in \mathbb{R}^{|\mathcal{Z}|}$ denote these logits, where $q_\phi(z \mid s_{t:t+k}) = \mathrm{softmax}(f_\phi(s_{t:t+k}))$. We constrain the network layers such that:

$$\|f_\phi(s_{t:t+k}) - f_\phi(s_{t':t'+k})\| \leq L \cdot \|s_{t:t+k} - s_{t':t'+k}\|, \tag{7}$$

where L is a Lipschitz constant, this constraint ensures that small changes in trajectory space induce only small changes in the predicted skill distribution, preventing the discriminator from exploiting spurious high-frequency variations. As a result, the agent is encouraged to generate trajectories that are globally distinguishable rather than merely locally separable.

The following result makes the connection between this constraint and state coverage explicit. For compactness, write a trajectory segment as $\tau = (s_t, \ldots, s_{t+k})$ and equip the concatenated state vector with the Euclidean norm.

Proposition 1 (Reward requires trajectory-space coverage). *Let $q_\phi(z \mid \tau) = \mathrm{softmax}(f_\phi(\tau))$, and suppose that f_ϕ satisfies Eq. equation 7. Consider two segments τ and τ' generated under distinct skills $z \neq z'$. If both skills are decoded with confidence*

$$q_\phi(z \mid \tau) \geq e^{-\delta} \quad \textit{and} \quad q_\phi(z' \mid \tau') \geq e^{-\delta}, \qquad 0 < \delta < \log 2, \tag{8}$$

then

$$\|\tau - \tau'\|_2 \geq \frac{\sqrt{2}}{L} \underbrace{\left[-\log\left(e^{\delta} - 1\right)\right]}_{m(\delta)}. \tag{9}$$

Consequently, there is at least one index $i \in \{0, \ldots, k\}$ for which

$$\left\|s_{t+i} - s'_{t'+i}\right\|_2 \geq \frac{\sqrt{2}\, m(\delta)}{L\sqrt{k+1}}. \tag{10}$$

The proof is deferred to Appendix A.2.

Proposition 1 shows that discriminator reward is *purchased with distance*: confidently identifying two different skills requires the policy to visit state sequences separated at a scale proportional to $m(\delta)/L$. With N skills attaining the same confidence, their segments therefore form an N-point packing in trajectory space at the scale in Eq. equation 9. A tighter Lipschitz constraint (smaller L) raises the minimum separation needed to obtain the same reward, whereas an unconstrained discriminator can create the required logit margin from arbitrarily small local differences. This explains why the constraint increases exploration pressure and, through Eq. equation 10, expands the set of visited states. It also predicts that coverage should grow with the number of skills, consistent with the cardinality ablation in Appendix D.2.

4 Experimental Evaluation

We evaluate four questions: whether TIME learns state-agnostic behaviors (Q1), coherent and predictable trajectories (Q2), broad state coverage (Q3), and reusable skills for downstream hierarchical control (Q4). The main text focuses on Q1, Q3, and Q4; the detailed trajectory-embedding diagnostic for Q2 is reported in Appendix B.2.

4.1 Experimental Setup

Environments. We evaluate TIME on continuous-control benchmarks including `Ant-v5` from Gym Brockman et al. (2016) for spatial navigation, along with `HalfCheetah`, `Quadruped`, and `Humanoid` from the DMC suite Tassa et al. (2018) for complex, high-dimensional locomotion.

Baselines. We compare TIME against representative unsupervised skill discovery methods, including MI-based approaches (**DIAYN Eysenbach et al. (2019), DADS Sharma et al. (2020)**), metric-aware methods (**METRA Park et al. (2024), RSD Zhang et al. (2025)**), and the disentanglement-based method **DUSDi Hu et al. (2024)**. These comparisons evaluate TIME's ability to learn trajectory-defined skills, improve exploration coverage, and support downstream hierarchical control.

Evaluation Metrics & Implementation. To address **Q1–Q4**, we combine qualitative trajectory visualizations with quantitative evaluations of skill diversity, state coverage, representation structure, and downstream control performance. Full implementation details and metric definitions are provided in Appendix B.

4.2 State-Agnostic and Temporally Extended Skills

To address **Q1**, we first visualize learned trajectories in Figure 3. Unlike state-centric methods that associate skills with disjoint spatial regions, TIME allows multiple skills to reach overlapping states while remaining distinguishable through their temporally extended future behaviors.

We further quantify this property using the conditional entropy $H(Z|S)$, where higher entropy indicates weaker dependence between skills and individual states. As shown in Table 1, TIME

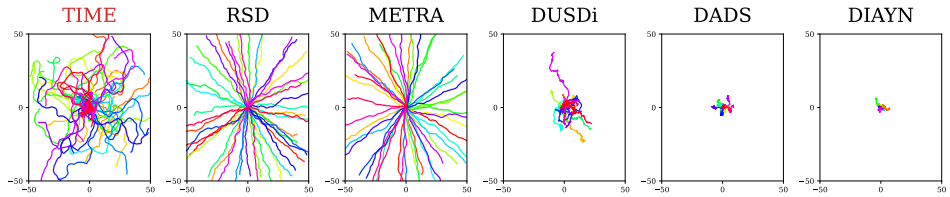


Figure 3: **Qualitative Analysis of Discovered Skills on `Ant-v5`.** 2D spatial trajectories ($x - y$) generated by policies learned in the `Ant-v5` environment. Distinct colors denote individual latent skills z. While TIME achieves extensive state coverage comparable to metric-aware baselines (METRA, RSD), it uniquely discovers temporally extended, overlapping macro-behaviors—a structural advantage driven by our trajectory-level MI objective.

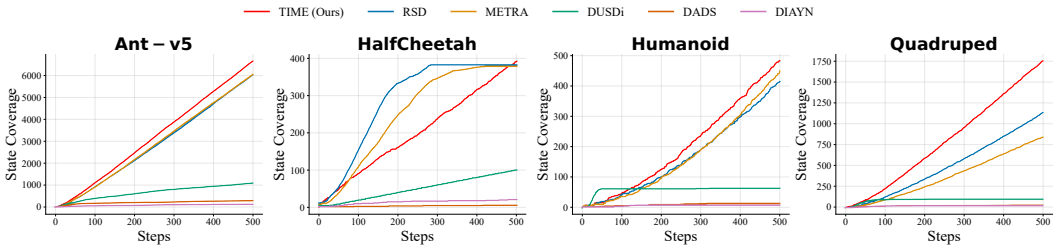


Figure 4: **State coverage over training steps.** TIME matches or exceeds metric-aware baselines (RSD, METRA), demonstrating that our Lipschitz-constrained discriminator effectively prevents local exploitation and promotes broad exploration.

achieves the highest entropy on three of the four environments, outperforming the strongest baseline by **157.1%** on `Ant-v5`, **18.2%** on `HalfCheetah`, and **34.0%** on `Quadruped`. On the more complex `Humanoid` environment, TIME remains competitive, outperforming METRA by **3.1%** while remaining within **8.1%** of RSD. These results confirm that TIME effectively learns state-agnostic skills while preserving broad exploration coverage. Additional qualitative visualizations and full trajectory comparisons across all environments are provided in Section 4.3.

The complementary Q2 analysis in Appendix B.2 confirms that TIME's trajectory embeddings are non-degenerate and well separated across skills.

Table 1: State-skill conditional entropy $H(Z|S)$, reported as mean $\pm$ standard deviation over five seeds. Higher is better.

Environment	METRA	RSD	TIME (Ours)
Ant-v5	0.068 ± 0.007	0.070 ± 0.006	$\mathbf{0.180 \pm 0.012}$
HalfCheetah	2.193 ± 0.075	1.864 ± 0.107	$\mathbf{2.593 \pm 0.087}$
Humanoid	0.638 ± 0.045	$\mathbf{0.716 \pm 0.049}$	0.658 ± 0.035
Quadruped	0.216 ± 0.018	0.247 ± 0.022	$\mathbf{0.331 \pm 0.015}$

4.3 Evaluating Broad State Coverage

To address **Q3**, we evaluate exploration breadth by measuring the cumulative number of unique spatial regions visited during pre-training (Figure 4). The results confirm that existing mutual information-based approaches, including DIAYN, DADS, and DUSDi, tend to exploit localized behavioral variations, leading to limited environment coverage. In contrast, TIME significantly improves exploration efficiency, consistently outperforming the best-performing MI-based baseline (DUSDi) by substantial margins ranging from **3.9**$\times$ (HalfCheetah) up to **16.6**$\times$ (`Quadruped`). More remarkably, despite using a standard, non-metric MI objective, TIME surpasses specialized metric-aware baselines (METRA, RSD) explicitly designed for exploration. When compared against the strongest metric-aware competitor in each respective environment, TIME consistently achieves higher final coverage, notably expanding the state space by **+54.2%** on `Quadruped` (vs. RSD), and maintaining a strict advantage in `Ant-v5` (**+9.9%** vs. METRA), `Humanoid` (**+7.4%** vs. METRA), and `HalfCheetah` (**+2.6%** vs. RSD). These empirical results directly validate the theoretical intuition established in Section 3.2. To further isolate

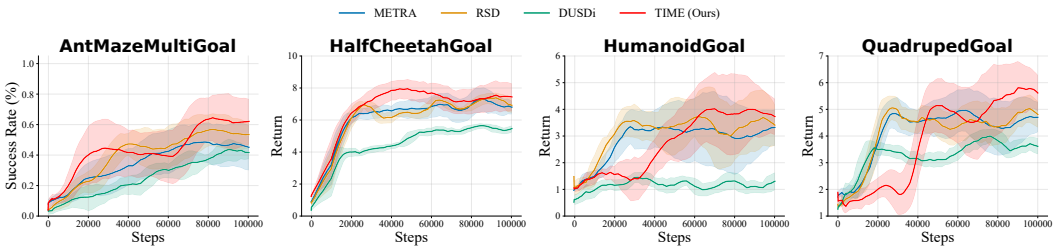


Figure 5: **Downstream Task Performance.** Learning curves of TIME and baseline methods on goal-reaching environments. Solid lines denote smoothed returns or success rates over training, and shaded regions indicate the corresponding rolling standard deviation.

the impact of the Lipschitz constraint within our framework, we compare TIME against an unregularized ablation variant in Appendix D.

4.4 Evaluating Downstream Task Performance

To address **Q4** and evaluate the effectiveness of TIME on downstream tasks, we assess the learned skills on four goal-reaching environments, comparing against three recent baselines: RSD, METRA, and DUSDi. Detailed descriptions of these environments are provided in Appendix B. In this setting, a meta-controller is trained to sequentially compose pre-trained unsupervised skills to guide the agent toward specified target locations.

Table 2: **Final Convergence Performance.** Results are reported as mean $\pm$ standard deviation over five seeds. TIME achieves the highest mean performance across all evaluated environments.

Algorithm	AntMazeMultiGoal	HalfCheetahGoal	HumanoidGoal	QuadrupedGoal
TIME (Ours)	**0.618 $\pm$ 0.148**	**7.400 $\pm$ 0.452**	**3.722 $\pm$ 0.662**	**5.605 $\pm$ 0.665**
RSD	0.533 $\pm$ 0.083	6.880 $\pm$ 0.430	3.409 $\pm$ 0.766	4.804 $\pm$ 0.484
METRA	0.448 $\pm$ 0.146	6.764 $\pm$ 0.413	3.323 $\pm$ 0.651	4.706 $\pm$ 0.557
DUSDi	0.414 $\pm$ 0.037	5.448 $\pm$ 0.210	1.332 $\pm$ 0.326	3.622 $\pm$ 0.301

Asymptotic Performance and Scalability. As shown in Figure 5 and Table 2, TIME consistently achieves the best mean final performance across all environments over five seeds. Averaged across tasks, TIME improves upon prior methods by **+37.3%** overall (**+12.3%** over RSD, **+19.6%** over METRA, and **+79.8%** over DUSDi). The gain over the strongest baseline (RSD) is **+15.9%** on `AntMazeMultiGoal`, **+7.6%** on `HalfCheetahGoal`, **+9.2%** on `HumanoidGoal`, and **+16.7%** on `QuadrupedGoal`. We attribute these improvements to the structure of TIME's high-level action space: by defining skills through *trajectory-level* effects rather than state occupancy. This decouples high-level decision making from local state constraints, allowing the meta-controller π_{high} to focus on *which* behavior to invoke, rather than whether it is feasible in the current state.

TIME is less sample-efficient early in the most challenging tasks and exhibits higher training variance; Appendix B.5 analyzes this trade-off in detail.

5 Related Work

Mutual Information-Based Skill Discovery. Unsupervised skill discovery typically learns task-agnostic behaviors by maximizing mutual information between visited states and latent skills Eysenbach et al. (2019). Subsequent works extend this framework such as state clustering Bai et al. (2024), state-space factorization Wang et al. (2024); Hu et al. (2024), evolutionary speciation Wickman & Zhang (2025), human-aligned objectives Hussonnois et al. (2025), and enhanced representation learning, exploration prioritization, or sampling strategies Levy et al. (2025); Ma et al. (2026); Zhang et al. (2025). Despite these advances, existing approaches remain fundamentally *state-centric* while TIME shifts the mutual information objective from isolated state occupancy to *trajectory-level*

temporal evolution, enabling the discovery of coherent and state-agnostic behaviors that generalize across diverse initial conditions.

Trajectory-Level and Progressive Exploration. Recent approaches address localized occupancy via trajectory-centric formulations. DADS Sharma et al. (2020) maximizes MI over one-step transitions, lacking explicit long-horizon structure. PDRL He et al. (2024) encourages temporally extended skills by exploring hard-to-reach states, but relies on heuristic intrinsic rewards rather than a principled information-theoretic foundation. TIME introduces a tractable variational lower bound for trajectory-level conditional MI, defining skills through coherent, multi-step behavioral dynamics.

Metric-Aware Learning and Lipschitz Constraints To combat the limited state coverage of pure MI maximization, prior works introduce auxiliary exploration bonuses Liu et al. (2025); Xiao et al. (2025), metric-aware MI objectives for spatial displacement (e.g., METRA Park et al. (2024)), or Lipschitz constraints Gulrajani et al. (2017); Barbara et al. (2024); Lin et al. (2021) on state discriminators (e.g., LSD Park et al. (2022)). TIME retains the standard MI objective and applies Lipschitz regularization directly to the *trajectory-centric discriminator*. This prevents the exploitation of spurious spatial variations, naturally promoting broad state coverage.

6 Conclusion

In this work, we present TIME, an unsupervised skill discovery framework that shifts the focus from isolated state occupancy to trajectory-level temporal evolution. By optimizing a tractable lower bound on conditional mutual information, TIME decouples skill identity from spatial localization, enabling coherent, state-agnostic behaviors. Additionally, enforcing a Lipschitz constraint on the trajectory discriminator naturally promotes broad exploration without relying on handcrafted metrics. Empirically, TIME consistently outperforms leading baselines across diverse continuous control benchmarks in both exploration and downstream hierarchical control.

While effective, the TIME framework presents certain limitations that motivate future work. First, optimizing the proposed objective requires training trajectory- and state-centric discriminators jointly. This dual-discriminator architecture introduces additional computational overhead and can make unsupervised pre-training more challenging to stabilize than conventional single-discriminator approaches. Second, the expressive multi-step skills learned by TIME inherently expand the high-level decision space for downstream HRL. Although this expressivity mitigates premature convergence and yields higher asymptotic performance, it can reduce early-stage sample efficiency and increase optimization variance during adaptation. Promising future directions include exploring unified representation learning to simplify the discriminator architecture, developing stronger regularization techniques for the learned skill space, and extending TIME to multi-agent HRL settings.

References

Chenjia Bai, Rushuai Yang, Qiaosheng Zhang, Kang Xu, Yi Chen, Ting Xiao, and Xuelong Li. Constrained ensemble exploration for unsupervised skill discovery. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 2418–2442. PMLR, 21–27 Jul 2024.

Nicholas H Barbara, Ruigang Wang, and Ian R Manchester. On robust reinforcement learning with lipschitz-bounded policy networks. In *Symposium on Systems Theory in Data and Optimization*, pp. 137–152. Springer, 2024.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function. In *International Conference on Learning Representations (ICLR)*, 2019.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017.

Ziming He, Chao Song, Jingchen Li, and Haobin Shi. Pdrl: Towards deeper states and further behaviors in unsupervised skill discovery by progressive diversity. *IEEE Transactions on Cognitive and Developmental Systems*, 17(3):495–509, 2024.

Jiaheng Hu, Zizhao Wang, Peter Stone, and Roberto Martín-Martín. Disentangled unsupervised skill discovery for efficient hierarchical reinforcement learning. *Advances in neural information processing systems*, 37:76529–76552, 2024.

Maxence Hussonnois, Thommen George Karimpanal, and Santu Rana. Human-aligned skill discovery: Balancing behaviour exploration and alignment. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems*, pp. 1025–1033, 2025.

Michael Laskin, Hao Liu, Xue Bin Peng, Denis Yarats, Aravind Rajeswaran, and Pieter Abbeel. Unsupervised reinforcement learning with contrastive intrinsic control. In *Advances in Neural Information Processing Systems*, volume 35, pp. 34478–34491, 2022.

Andrew Levy, Alessandro G Allievi, and George Konidaris. Representation learning and skill discovery with empowerment. In *Reinforcement Learning Conference*, 2025.

Zinan Lin, Vyas Sekar, and Giulia Fanti. Why spectral normalization stabilizes gans: Analysis and improvements. *Advances in neural information processing systems*, 34:9625–9638, 2021.

Hao Liu and Pieter Abbeel. Aps: Active pretraining with successor features. In *International Conference on Machine Learning*, pp. 6736–6747. PMLR, 2021.

Xin Liu, Yaran Chen, Guixing Chen, Haoran Li, and Dongbin Zhao. Balancing state exploration and skill diversity in unsupervised skill discovery. *IEEE Transactions on Cybernetics*, 2025.

Nelson Ma, Junyu Xuan, Guangquan Zhang, and Jie Lu. Discovering mixture skills for unsupervised reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 24262–24270, 2026.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

Sherjil Ozair, Corey Lynch, Yoshua Bengio, Aaron Van den Oord, Sergey Levine, and Pierre Sermanet. Wasserstein dependency measure for representation learning. *Advances in Neural Information Processing Systems*, 32, 2019.

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

Seohong Park, Oleh Rybkin, and Sergey Levine. Metra: Scalable unsupervised rl with metric-aware abstraction. In *International Conference on Learning Representations (ICLR)*, 2024.

Shubham Pateria, Budhitama Subagdja, Ah-hwee Tan, and Chai Quek. Hierarchical reinforcement learning: A comprehensive survey. *ACM Computing Surveys (CSUR)*, 54(5):1–35, 2021.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015.

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

Archit Sharma, Shixiang Gu, Sergey Levine, Vikash Kumar, and Karol Hausman. Dynamics-aware unsupervised discovery of skills. In *International Conference on Learning Representations (ICLR)*, 2020.

Yunlong Song, Angel Romero, Matthias Müller, Vladlen Koltun, and Davide Scaramuzza. Reaching the limit in autonomous racing: Optimal control versus reinforcement learning. *Science Robotics*, 8(82):eadg1462, 2023.

Ebrahim Hamid Sumiea, Said Jadid Abdulkadir, Hitham Seddig Alhussian, Safwan Mahmood Al-Selwi, Alawi Alqushaibi, Mohammed Gamal Ragab, and Suliman Mohamed Fati. Deep deterministic policy gradient algorithm: A systematic review. *Heliyon*, 10(9), 2024.

Chen Tang, Ben Abbatematteo, Jiaheng Hu, Rohan Chandra, Roberto Martín-Martín, and Peter Stone. Deep reinforcement learning for robotics: A survey of real-world successes. *Annual Review of Control, Robotics, and Autonomous Systems*, 8(1):153–188, 2025.

Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, et al. Deepmind control suite. *arXiv preprint arXiv:1801.00690*, 2018.

Brendan D Tracey, Andrea Michi, Yuri Chervonyi, Ian Davies, Cosmin Paduraru, Nevena Lazic, Federico Felici, Timo Ewalds, Craig Donner, Cristian Galperti, et al. Towards practical reinforcement learning for tokamak magnetic control. *Fusion Engineering and Design*, 200:114161, 2024.

Janett Walters-Williams and Yan Li. Estimation of mutual information: A survey. In *International Conference on Rough Sets and Knowledge Technology*, pp. 389–396. Springer, 2009.

Zizhao Wang, Jiaheng Hu, Caleb Chuck, Stephen Chen, Roberto Martín-Martín, Amy Zhang, Scott Niekum, and Peter Stone. Skild: Unsupervised skill discovery guided by factor interactions. *Advances in Neural Information Processing Systems*, 37:77748–77776, 2024.

Ryan Wickman and Xiaofei Zhang. Speciated unsupervised skill discovery. In *2025 IEEE International Conference on Data Mining (ICDM)*, pp. 1633–1641. IEEE, 2025.

Ting Xiao, Jiakun Zheng, Rushuai Yang, Kang Xu, Qiaosheng Zhang, Peng Liu, Zhe Wang, and Chenjia Bai. Unsupervised skill discovery through skill regions differentiation. *IEEE Transactions on Neural Networks and Learning Systems*, 2025.

He Zhang, Ming Zhou, Shaopeng Zhai, Ying Sun, and Hui Xiong. Efficient skill discovery via regret-aware optimization. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2025.

A Theoretical Analysis and Proofs

A.1 Proofs of Theorem 1 and its Corollaries

Proof of Theorem 1. We start from the conditional mutual information objective

$$I(S_{t+1:t+k}; Z \mid S_t). \tag{11}$$

Applying the chain rule of mutual information gives

$$I(S_{t:t+k}; Z) = I(S_t; Z) + I(S_{t+1:t+k}; Z \mid S_t), \tag{12}$$

which implies

$$I(S_{t+1:t+k}; Z \mid S_t) = I(S_{t:t+k}; Z) - I(S_t; Z). \tag{13}$$

Next, consider our variational objective:

$$\begin{aligned} I_{\text{TIME}} &= \mathbb{E}_{s_{t:t+k},z}\left[\log \frac{q_\theta(z \mid s_{t:t+k})}{p(z)\, q_{\theta'}(z \mid s_t)}\right] + \mathbb{E}_{s_t}\mathbb{E}_z\left[\log q_{\theta'}(z \mid s_t)\right] \\ &= \underbrace{\mathbb{E}_{s_{t:t+k},z}\left[\log \frac{q_\theta(z \mid s_{t:t+k})}{p(z)}\right]}_{(A)} - \underbrace{\left(\mathbb{E}_{s_{t:t+k},z}\left[\log q_{\theta'}(z \mid s_t)\right] - \mathbb{E}_{s_t}\mathbb{E}_z\left[\log q_{\theta'}(z \mid s_t)\right]\right)}_{(B)}. \end{aligned} \tag{14}$$

We first analyze term (A):

$$\begin{aligned} I(S_{t:t+k}; Z) - (A) &= \mathbb{E}_{s_{t:t+k},z}\left[\log \frac{p(z \mid s_{t:t+k})}{p(z)}\right] - \mathbb{E}_{s_{t:t+k},z}\left[\log \frac{q_\theta(z \mid s_{t:t+k})}{p(z)}\right] \\ &= \mathbb{E}_{s_{t:t+k},z}\left[\log \frac{p(z \mid s_{t:t+k})}{q_\theta(z \mid s_{t:t+k})}\right] \\ &= \mathrm{KL}\Big(p(z, s_{t:t+k}) \,\|\, q_\theta(z, s_{t:t+k})\Big), \end{aligned} \tag{15}$$

where $q_\theta(z, s_{t:t+k}) = q_\theta(z \mid s_{t:t+k})\, p(s_{t:t+k})$.

Next, we analyze term (B):

$$\begin{aligned} I(S_t; Z) - (B) &= \mathbb{E}_{s_t,z}\left[\log \frac{p(z \mid s_t)}{p(z)}\right] - \left(\mathbb{E}_{s_t,z}\left[\log q_{\theta'}(z \mid s_t)\right] - \mathbb{E}_{s_t}\mathbb{E}_z\left[\log q_{\theta'}(z \mid s_t)\right]\right) \\ &= \mathbb{E}_{s_t,z}\left[\log p(z \mid s_t)\right] - \mathbb{E}_{s_t,z}\left[\log p(z)\right] - \mathbb{E}_{s_t,z}\left[\log q_{\theta'}(z \mid s_t)\right] + \mathbb{E}_{s_t}\mathbb{E}_z\left[\log q_{\theta'}(z \mid s_t)\right] \\ &= \mathbb{E}_{s_t,z}\left[\log \frac{p(z \mid s_t)}{q_{\theta'}(z \mid s_t)}\right] - \mathbb{E}_{s_t}\mathbb{E}_z\left[\log \frac{p(z)}{q_{\theta'}(z \mid s_t)}\right] \\ &= \mathbb{E}_{s_t,z}\left[\log \frac{p(z \mid s_t)p(s_t)}{q_{\theta'}(z \mid s_t)p(s_t)}\right] - \mathbb{E}_{s_t}\mathbb{E}_z\left[\log \frac{p(z)p(s_t)}{q_{\theta'}(z \mid s_t)p(s_t)}\right] \\ &= \mathrm{KL}\Big(p(z, s_t) \,\|\, q_{\theta'}(z, s_t)\Big) - \mathrm{KL}\Big(p(z)p(s_t) \,\|\, q_{\theta'}(z, s_t)\Big), \end{aligned} \tag{16}$$

where $q_{\theta'}(z, s_t) = q_{\theta'}(z \mid s_t)p(s_t)$.

Finally, combining Eq. equation 13, Eq. equation 15, and Eq. equation 16, we obtain

$$\begin{aligned} &I(S_{t+1:t+k}; Z \mid S_t) - I_{\text{TIME}}(S_{t+1:t+k}; Z \mid S_t) \\ &= \Big(I(S_{t:t+k}; Z) - (A)\Big) - \Big(I(S_t; Z) - (B)\Big) \\ &= \mathrm{KL}\Big(p(z, s_{t:t+k}) \,\|\, q_\theta(z, s_{t:t+k})\Big) + \mathrm{KL}\Big(p(z)p(s_t) \,\|\, q_{\theta'}(z, s_t)\Big) - \mathrm{KL}\Big(p(z, s_t) \,\|\, q_{\theta'}(z, s_t)\Big). \end{aligned} \tag{17}$$

Therefore,

$$I(S_{t+1:t+k}; Z \mid S_t) \geq I_{\text{TIME}}(S_{t+1:t+k}; Z \mid S_t) \tag{18}$$

if and only if

$$\begin{aligned} &\mathrm{KL}\big(p(z)p(s_t) \,\|\, q_{\theta'}(z, s_t)\big) \\ -\,&\mathrm{KL}\big(p(z, s_t) \,\|\, q_{\theta'}(z, s_t)\big) \\ +\,&\mathrm{KL}\big(p(z, s_{t:t+k}) \,\|\, q_{\theta}(z, s_{t:t+k})\big) \geq 0, \end{aligned} \tag{19}$$

which is exactly Eq. equation 4. This completes the proof. □

Corollary 1 (Approximation Error Bound)**.** *Assume* $\mathrm{KL}(p(z \mid s_t) \| q_{\theta'}(z \mid s_t)) < \epsilon_1$ *and* $\mathrm{KL}(p(z \mid s_{t:t+k}) \| q_{\theta}(z \mid s_{t:t+k})) < \epsilon_2$*. If* $\mathrm{KL}(p(z)p(s_t) \| q_{\theta'}(z, s_t)) < \mathrm{KL}(p(z, s_t) \| q_{\theta'}(z, s_t))$*, then*

$$|I(S_{t+1:t+k}; Z \mid S_t) - I_{\mathrm{TIME}}(S_{t+1:t+k}; Z \mid S_t)| < \epsilon_1 + \epsilon_2.$$

Proof of Corollary 1. First, we expand the joint KL divergence terms using the definitions $p(z, s_t) = p(z \mid s_t)p(s_t)$ and $q_{\theta'}(z, s_t) = q_{\theta'}(z \mid s_t)p(s_t)$:

$$\begin{aligned} \mathrm{KL}\big(p(z, s_t) \,\|\, q_{\theta'}(z, s_t)\big) &= \mathbb{E}_{p(z,s_t)}\left[\log \frac{p(z, s_t)}{q_{\theta'}(z, s_t)}\right] \\ &= \mathbb{E}_{p(z,s_t)}\left[\log \frac{p(z \mid s_t)p(s_t)}{q_{\theta'}(z \mid s_t)p(s_t)}\right] \\ &= \mathrm{KL}\big(p(z \mid s_t) \,\|\, q_{\theta'}(z \mid s_t)\big) < \epsilon_1. \end{aligned} \tag{20}$$

By the given condition $\mathrm{KL}\big(p(z)p(s_t) \,\|\, q_{\theta'}(z, s_t)\big) < \mathrm{KL}\big(p(z, s_t) \,\|\, q_{\theta'}(z, s_t)\big)$, we immediately have:

$$\mathrm{KL}\big(p(z)p(s_t) \,\|\, q_{\theta'}(z, s_t)\big) < \epsilon_1. \tag{21}$$

Note that the KL-divergence is always non-negative. Therefore, the absolute difference between these two terms is bounded by their maximum:

$$\begin{aligned} &\Big|\mathrm{KL}\big(p(z)p(s_t) \,\|\, q_{\theta'}(z, s_t)\big) - \mathrm{KL}\big(p(z, s_t) \,\|\, q_{\theta'}(z, s_t)\big)\Big| \\ &< \max\Big\{\mathrm{KL}\big(p(z)p(s_t) \,\|\, q_{\theta'}(z, s_t)\big), \mathrm{KL}\big(p(z, s_t) \,\|\, q_{\theta'}(z, s_t)\big)\Big\} \leq \epsilon_1. \end{aligned} \tag{22}$$

Similarly, for the multi-step trajectory term, we have:

$$\begin{aligned} \mathrm{KL}\big(p(z, s_{t:t+k}) \,\|\, q_{\theta}(z, s_{t:t+k})\big) &= \mathbb{E}_{p(z,s_{t:t+k})}\left[\log \frac{p(z \mid s_{t:t+k})p(s_{t:t+k})}{q_{\theta}(z \mid s_{t:t+k})p(s_{t:t+k})}\right] \\ &= \mathrm{KL}\big(p(z \mid s_{t:t+k}) \,\|\, q_{\theta}(z \mid s_{t:t+k})\big) < \epsilon_2. \end{aligned} \tag{23}$$

From the proof of Theorem 1, we know the exact gap, denoted as $\tilde{\Delta}$, is given by Eq. equation 17:

$$\Delta := I(S_{t+1:t+k}; Z \mid S_t) - I_{\mathrm{TIME}}(S_{t+1:t+k}; Z \mid S_t).$$

Applying the triangle inequality, we bound the absolute gap:

$$\begin{aligned} |\Delta| &= \Big|\mathrm{KL}\big(p(z, s_{t:t+k}) \,\|\, q_{\theta}(z, s_{t:t+k})\big) + \mathrm{KL}\big(p(z)p(s_t) \,\|\, q_{\theta'}(z, s_t)\big) - \mathrm{KL}\big(p(z, s_t) \,\|\, q_{\theta'}(z, s_t)\big)\Big| \\ &\leq \mathrm{KL}\big(p(z, s_{t:t+k}) \,\|\, q_{\theta}(z, s_{t:t+k})\big) + \Big|\mathrm{KL}\big(p(z)p(s_t) \,\|\, q_{\theta'}(z, s_t)\big) - \mathrm{KL}\big(p(z, s_t) \,\|\, q_{\theta'}(z, s_t)\big)\Big| \\ &< \epsilon_2 + \epsilon_1, \end{aligned} \tag{24}$$

which supports the claim. □

Corollary 2 (Tightness under Perfect Approximation)**.** *If* $q_{\theta}(z \mid s_{t:t+k}) = p(z \mid s_{t:t+k})$ *and* $I(S_t; Z) = 0$*, then*

$$I_{\mathrm{TIME}}(S_{t+1:t+k}; Z \mid S_t) = I(S_{t+1:t+k}; Z \mid S_t).$$

Proof of Corollary 2. If $q_\theta(z \mid s_{t:t+k}) \to p(z \mid s_{t:t+k})$, then the approximation is perfect and the corresponding KL divergence vanishes:

$$\mathrm{KL}\big(p(z, s_{t:t+k}) \,\|\, q_\theta(z, s_{t:t+k})\big) = \mathrm{KL}\big(p(z \mid s_{t:t+k}) \,\|\, q_\theta(z \mid s_{t:t+k})\big) = 0. \tag{25}$$

If $I(S_t; Z) \to 0$, it implies that the variables Z and S_t are independent, i.e., $p(z, s_t) = p(z)p(s_t)$. Consequently, the joint and product-of-marginal KL divergences with respect to $q_{\theta'}$ become identical:

$$\mathrm{KL}\big(p(z)p(s_t) \,\|\, q_{\theta'}(z, s_t)\big) = \mathrm{KL}\big(p(z, s_t) \,\|\, q_{\theta'}(z, s_t)\big). \tag{26}$$

Substituting these into the gap derived in Eq. equation 17 of Theorem 1, we obtain:

$$\begin{aligned} &I(S_{t+1:t+k}; Z \mid S_t) - I_{\text{TIME}}(S_{t+1:t+k}; Z \mid S_t) \\ &= 0 + \mathrm{KL}\big(p(z, s_t) \,\|\, q_{\theta'}(z, s_t)\big) - \mathrm{KL}\big(p(z, s_t) \,\|\, q_{\theta'}(z, s_t)\big) \\ &= 0. \end{aligned} \tag{27}$$

Therefore, $I_{\text{TIME}}(S_{t+1:t+k}; Z \mid S_t) = I(S_{t+1:t+k}; Z \mid S_t)$, and the equality holds perfectly. □

A.2 Proof of the Lipschitz Coverage Proposition

Proof of Proposition 1. The softmax probability satisfies

$$q_\phi(z \mid \tau) = \left(1 + \sum_{j \neq z} \exp(f_{\phi,j}(\tau) - f_{\phi,z}(\tau))\right)^{-1}.$$

Thus, $q_\phi(z \mid \tau) \geq e^{-\delta}$ implies $f_{\phi,z}(\tau) - f_{\phi,j}(\tau) \geq m(\delta)$ for every $j \neq z$. Applying the same argument to (z', τ') gives

$$f_{\phi,z}(\tau) - f_{\phi,z'}(\tau) \geq m(\delta), \qquad f_{\phi,z}(\tau') - f_{\phi,z'}(\tau') \leq -m(\delta).$$

Let $u = e_z - e_{z'}$, so $\|u\|_2 = \sqrt{2}$. Cauchy–Schwarz and Eq. equation 7 yield

$$2m(\delta) \leq \langle u, f_\phi(\tau) - f_\phi(\tau') \rangle \tag{28}$$

$$\leq \sqrt{2}\, \|f_\phi(\tau) - f_\phi(\tau')\|_2 \leq \sqrt{2} L \, \|\tau - \tau'\|_2 , \tag{29}$$

which proves Eq. equation 9. Finally, $\|\tau - \tau'\|_2^2 = \sum_{i=0}^{k} \|s_{t+i} - s'_{t'+i}\|_2^2$; if Eq. equation 10 failed at every index, this identity would contradict Eq. equation 9. □

A.3 Relation to Prior Work.

While most existing MI-based methods optimize objectives derived from strict variational lower bounds of $I(S; Z)$, extending such formulations to the conditional MI $I(S_{t+1:t+k}; Z \mid S_t)$ is substantially more challenging. The difficulty stems from the fact that the required distributions, such as $p(s_{t+1:t+k} \mid s_t)$ and $p(s_{t+1:t+k} \mid s_t, z)$, are generally unknown in RL environments and hard to model explicitly. Prior approaches such as DADS Sharma et al. (2020) address this challenge by optimizing a one-step approximation, $I(S_{t+1}; Z \mid S_t)$, through a learned forward dynamics model. However, this approximation lacks a principled variational lower bound to the true objective and does not naturally extend to long-horizon trajectory modeling. In contrast, our formulation directly captures long-term predictability without explicitly learning environment dynamics, yielding a tractable and theoretically grounded variational lower bound on trajectory-level conditional mutual information.

A.4 Inapplicability of Wasserstein Dependency Measure: A Theoretical Perspective

Recent skill discovery methods such as METRA Park et al. (2024) and RSD Zhang et al. (2025) maximize a Wasserstein dependency measure Ozair et al. (2019) as a proxy for mutual information. For global state-skill dependency $I(S; Z)$, the Wasserstein-based mutual information is defined as:

$$I_W(S; Z) = W_1\Big(p(s, z), p(s)p(z)\Big), \tag{30}$$

where W_1 is the 1-Wasserstein distance. By Kantorovich-Rubinstein duality, this is equivalent to:

$$I_W(S;Z) = \sup_{\|f\|_L \le 1} \left(\mathbb{E}_{p(s,z)}[f(s,z)] - \mathbb{E}_{p(s)p(z)}[f(s,z)] \right). \tag{31}$$

In standard RL, Eq. equation 31 is tractable to optimize. The agent can efficiently estimate the first term using joint transitions (s,z) collected by the behavior policy, and the second term using the product of marginals $p(s)p(z)$ by *simply randomly shuffling states and skills* within a minibatch from the replay buffer.

However, our proposed objective focuses on the *conditional trajectory-centric* mutual information, $I(S_{t+1:t+k}; Z \mid S_t)$. To adapt this to the Wasserstein dependency framework, we define the conditional Wasserstein dependency measure as the expected distance between the joint conditional distribution and the product of conditional marginals over the state distribution:

$$I_W(S_{t+1:t+k}; Z \mid S_t) = \mathbb{E}_{s_t \sim p(s_t)} \left[W_1\Big(p(s_{t+1:t+k}, z \mid s_t),\ p(s_{t+1:t+k} \mid s_t) p(z \mid s_t) \Big) \right]. \tag{32}$$

By Kantorovich-Rubinstein duality, for any *fixed* state s_t, the 1-Wasserstein distance inside the expectation is:

$$\begin{aligned} &W_1\Big(p(s_{t+1:t+k}, z \mid s_t),\ p(s_{t+1:t+k} \mid s_t) p(z \mid s_t) \Big) \\ &= \sup_{\|f_{s_t}\|_L \le 1} \left(\mathbb{E}_{p(s_{t+1:t+k}, z \mid s_t)} \big[f_{s_t}(s_{t+1:t+k}, z) \big] - \mathbb{E}_{p(s_{t+1:t+k} \mid s_t) p(z \mid s_t)} \big[f_{s_t}(s_{t+1:t+k}, z) \big] \right). \end{aligned} \tag{33}$$

Instead of learning an independent, state-specific Lipschitz function f_{s_t} for each s_t, one might consider a shared conditional critic $f(s_{t+1:t+k}, z, s_t)$ that takes s_t as an additional input. Because the class of shared conditional critics $f(s_{t+1:t+k}, z, s_t)$ forms a restricted subset of the more general collection of arbitrary measurable state-specific critics $\{f_{s_t}\}_{s_t}$, the resulting objective induces the following variational lower bound surrogate:

$$\begin{aligned} &I_W(S_{t+1:t+k}; Z \mid S_t) \ge \tilde{I}_W(S_{t+1:t+k}; Z \mid S_t) \\ &:= \sup_{\|f\|_L \le 1} \mathbb{E}_{s_t \sim p(s_t)} \Bigg[\underbrace{\mathbb{E}_{p(s_{t+1:t+k}, z \mid s_t)} \big[f(s_{t+1:t+k}, z, s_t) \big]}_{(C)} - \underbrace{\mathbb{E}_{p(s_{t+1:t+k} \mid s_t) p(z \mid s_t)} \big[f(s_{t+1:t+k}, z, s_t) \big]}_{(D)} \Bigg]. \end{aligned} \tag{34}$$

While term (C) can be naturally estimated from standard RL trajectories (since the agent follows $\pi(a \mid s, z)$ to generate $s_{t+1:t+k}$ starting from s_t), evaluating the penalty term (D) reveals a fundamental limitation. Unlike the unconditional case where one can simply shuffle the replay buffer, **sampling from the conditional marginals $p(s_{t+1:t+k} \mid s_t)p(z \mid s_t)$ cannot be achieved by random shuffling**, because the conditioning variable s_t must strictly match for both the trajectory and the skill.

This matching requirement presents two severe practical challenges for standard off-policy RL:

1. **Intractability of the Skill Posterior:** Sampling $z \sim p(z \mid s_t)$ requires either access to the true posterior or a consistent learned estimator $q(z \mid s_t)$ of it. In episodic skill discovery, $z \sim p(z)$ is sampled uniformly at the beginning of an episode. Thus, the true posterior $p(z \mid s_t)$ depends on the entire trajectory history leading to s_t. Standard replay buffers do not directly provide samples from the conditional posterior $p(z \mid s_t)$. Consequently, any learned estimator $q(z \mid s_t)$ is likely to introduce approximation bias. These approximation errors would directly propagate into the Wasserstein objective, leading to unreliable dependency estimation.
2. **Impracticality of Marginalized Rollouts:** Sampling $s_{t+1:t+k} \sim p(s_{t+1:t+k} \mid s_t)$ requires rolling out a trajectory from s_t following the *marginal policy* $\pi_{\text{marg}}(a \mid s_t) = \int \pi(a \mid s_t, z) p(z \mid s_t) dz$. Estimating this integral at every environment step is computationally prohibitive. More crucially, obtaining valid k-step forward trajectories strictly conditioned on s_t without the governing z is generally intractable in standard RL, unless we assume strong environment capabilities such as a perfectly resettable generative simulator or a highly accurate model-based world model.

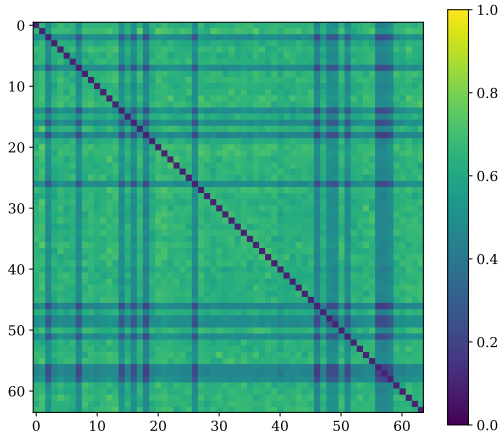


Figure 6: **Pairwise distance matrix of TIME trajectory embeddings on `Ant-v5`.** The cosine distances show intra-skill coherence along the diagonal and inter-skill diversity off the diagonal.

Because the negative term (D) is highly challenging to effectively sampled from the replay buffer without introducing severe bias or requiring impractical simulator assumptions, the metric-based exploration strategy employed by METRA does not directly extend to our conditional formulation. This fundamental sampling limitation directly motivates our variational approach (Theorem 1), which avoids explicit sampling from marginalized conditional distributions.

B Detailed Metric Implementation

This section provides comprehensive details on the implementation and mathematical formulation of the evaluation metrics discussed in Section 4.

B.1 State-Skill Conditional Entropy (Q1)

In Section 4.2, we quantitatively assess the degree of state-skill decoupling induced by TIME by measuring the conditional entropy $H(Z \mid S)$. A higher conditional entropy indicates that knowing the agent's spatial location S provides little information about the active skill Z, thereby suggesting that skills are less tightly coupled to isolated spatial regions.

Implementation. Since the environments feature continuous state spaces, we approximate the discrete probability distributions by mapping the 2D spatial coordinates (the xy-plane) of the agent into a discretized grid. Specifically, we partition the environment into a grid with a bin size of 1×1. Let s_{bin} denote a discretized spatial region. We collect trajectories from the fully trained policies and compute the empirical joint visitation counts to estimate the marginal $\hat{p}(s_{\text{bin}})$ and the conditional $\hat{p}(z \mid s_{\text{bin}})$. The conditional entropy is then calculated as:

$$H(Z \mid S) \approx -\sum_{s_{\text{bin}}} \hat{p}(s_{\text{bin}}) \sum_{z} \hat{p}(z \mid s_{\text{bin}}) \log \hat{p}(z \mid s_{\text{bin}})$$

Baseline Exclusion. Foundational MI-based methods (e.g., DIAYN, DADS) and highly localized state-centric methods (e.g., DUSDi) are omitted from this metric for interpretability reasons. As observed in the exploration coverage analysis, these methods typically occupy only a limited subset of the environment, resulting in sparse empirical support over the discretized state space. Under such limited coverage, the estimated conditional entropy becomes highly sensitive to discretization and sampling noise, making direct quantitative comparisons less informative.

B.2 Pairwise Trajectory Distance Matrix (Q2)

We investigate whether the latent skills discovered by TIME map to distinct and temporally distinguishable multi-step trajectories. Unlike state-centric methods that define skills by non-overlapping spatial goals, TIME differentiates skills via their temporal evolution.

Implementation. To quantitatively assess this, we construct a pairwise cosine distance matrix for the trajectory embeddings learned in the `Ant-v5` environment (as visualized in Figure 6). During evaluation, we uniformly sample $N = 64$ latent skills. To ensure that the trajectory variance originates strictly from the skill conditioning rather than varying initial conditions, the agent is initialized at a strictly fixed starting state s_0. For each skill z_i, the skill policy $\pi_\omega(a \mid s, z_i)$ executes a rollout to generate a k-step trajectory $\tau_i = (s_t)_{t=0}^{k}$.

These trajectories are then processed by our learned trajectory discriminator q_θ to extract their dense representations $e_i = q_\theta(\tau_i)$. The pairwise dissimilarity between any two skills z_i and z_j is computed using the cosine distance of their respective embeddings:

$$D_{\cos}(z_i, z_j) = 1 - \frac{e_i \cdot e_j}{\|e_i\|\|e_j\|}$$

This metric robustly captures the inter-skill diversity and intra-skill coherence directly in the trajectory space.

The off-diagonal entries have a mean distance of **0.578**, with a minimum of **0.177** and a maximum of **0.757**. Thus, even the closest pair remains separated, while structurally different traversals occupy distant regions of the learned trajectory space. These results support the claim that TIME learns a diverse, non-degenerate repertoire of temporally extended behaviors.

B.3 Exploration Coverage (Q3)

In Section 4.3, we evaluate the exploration breadth of the pre-trained agents by measuring the cumulative number of unique spatial regions visited throughout the training process.

Implementation. Consistent with the methodology used for the conditional entropy calculation, we project the agent's continuous (x, y) coordinates into a discretized 2D grid with a bin size of 1×1. A grid cell is marked as "visited" the first time the agent's center of mass enters its boundaries. We track the cumulative count of these distinct visited cells over the total environment steps.

This grid-based state counting protocol strictly adheres to the evaluation standards established by prior metric-aware methods, specifically METRA Park et al. (2024) and RSD Zhang et al. (2025). By employing the exact same bin resolution and counting mechanism, we ensure a fair, direct, and rigorous comparison of exploration efficiency across all baselines.

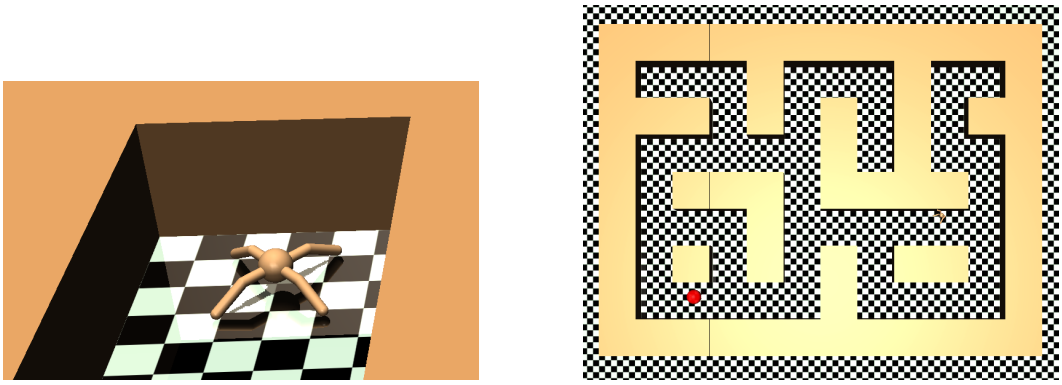
B.4 Downstream Goal-Reaching Environments (Q4)

To address **Q4** and evaluate the reusability of the discovered skills, we benchmark the pre-trained models on four challenging downstream tasks based on the DMC suite and OpenAI Gym: `AntMazeMultiGoal`, `HalfCheetahGoal`, `HumanoidGoal`, and `QuadrupedGoal`.

In all downstream evaluations, a hierarchical reinforcement learning setup is employed. The weights of the pre-trained low-level skill policy $\pi_\omega(a \mid s, z)$ are frozen. We then train a high-level meta-controller $\pi_{\text{high}}(z \mid s)$ using standard RL algorithms. The meta-controller operates at a temporally extended resolution: at every T environment steps, it observes the current state and selects a latent skill z, which the low-level policy subsequently executes for the next T transitions. Complete implementation details and hyperparameter configurations for this meta-controller are provided in Appendix C.

Point-to-Point Navigation For the `HalfCheetahGoal`, `QuadrupedGoal`, and `HumanoidGoal` environments, the objective is to direct the agent to a single, randomly generated target location.

- **Goal Generation:** The target coordinates are sampled uniformly from predefined intervals that scale with the specific environment's dynamics. For `HalfCheetahGoal`, the 1D



(a) The Ant Agent (b) Top-down view of the maze

Figure 7: **The `AntMazeMultiGoal` Environment. (a)** A close-up view of the quadrupedal Ant agent. **(b)** A top-down layout of the multi-goal maze. The agent must sequentially navigate through dynamically generated waypoints, providing a rigorous testbed for evaluating coherent, long-horizon skill chaining.

target is sampled from $[-100, 100]$. For the 2D locomotion tasks, targets are sampled from $[-7.5, 7.5]^2$ for `QuadrupedGoal`, and $[-5, 5]^2$ for `HumanoidGoal`.

- **Reward and Termination:** A target is considered successfully reached if the agent's center of mass comes within a Euclidean distance of 1.0 from the goal coordinates. Upon reaching the goal, the agent receives a sparse terminal reward of $+10$, and the episode terminates immediately. The performance metric reported for these tasks is the total episodic return.

B.5 Downstream Learning Dynamics

Although TIME achieves the strongest final performance, its more expressive trajectory-level skill space can make early high-level optimization harder. On the most challenging environments, `HumanoidGoal` and `QuadrupedGoal`, TIME is initially less sample-efficient than RSD and METRA; for example, its `HumanoidGoal` area under the learning curve is **2.75**, compared with **3.12** for RSD. TIME also has higher peak variability during training (average maximum standard deviation **1.128**, versus **0.955** for METRA). We attribute this trade-off to the richer set of futures induced by each trajectory-level skill: the meta-controller faces a more complex search space early in training, but this flexibility ultimately avoids the local optima that limit state-centric skills.

Sequential Waypoint Navigation To evaluate continuous, long-horizon skill chaining, we employ the `AntMazeMultiGoal` environment, which requires the agent to navigate through a dynamic sequence of target waypoints. This setup explicitly tests the temporal coherence of the learned behaviors and their robustness in complex, multi-stage planning scenarios (illustrated in Figure 7).

- **Goal Generation:** A target goal is initially spawned randomly in the environment. Once the agent successfully reaches the active goal, a new target is immediately generated. To ensure sustained traversal and prevent degenerate, localized hovering behaviors, the new goal is strictly constrained to spawn at a minimum Euclidean distance of 7.5 from the agent's current position. An episode contains a maximum of 25 sequential goals.
- **Reward Structure (Training):** To facilitate the optimization of the meta-controller, the environment provides a dense shaped reward at each timestep, formulated as: $r_t = \exp(-\|p_t - g_t\|_2)$ where p_t is the agent's current (x, y) position and g_t is the location of the active goal.
- **Evaluation Metric:** Although the dense exponential reward is utilized to train π_{high}, the true performance of the agent is evaluated based on its task completion capability. Consequently, as reported in our main results (Section 4.4), the evaluation metric for `AntMazeMultiGoal` is the *Success Rate (%)*, calculated as the total number of goals successfully reached divided by the maximum possible number of goals (25).

C METHOD IMPLEMENTATION DETAILS

In this section, we provide the complete pseudocode for the TIME framework and detail the network architectures, hyperparameter configurations, and optimization choices used in our experiments.

Optimization Algorithm. For the underlying unsupervised skill discovery process, we formulate the optimization as a continuous-control reinforcement learning problem. We employ Deep Deterministic Policy Gradient (DDPG) Sumiea et al. (2024) as the backbone off-policy RL algorithm to train the skill-conditioned behavior policy π_ω.

Lipschitz Constraint Implementation. To encourage the 1-Lipschitz continuity required by our theoretical formulation (Eq. 7), we apply Spectral Normalization Miyato et al. (2018) to all layers of the trajectory discriminator q_θ except the final output layer. This technique constrains the spectral norm of the weight matrices during training, leading to a more stable empirical estimation of the proposed variational lower bound.

The complete training procedure for the TIME framework is summarized in Algorithm 1.

Algorithm 1 TIME Pre-training Framework

Initialize: Skill policy π_ω, trajectory discriminator q_θ, state discriminator $q_{\theta'}$, replay buffer $\mathcal{D}$, and trajectory segment length k
Apply Spectral Normalization to all layers up to the logits of q_θ
for each skill training episode **do**
    Sample a latent skill $z \sim p(z)$ uniformly from the prior distribution
    Collect state transitions by executing actions $a_t \sim \pi_\omega(a \mid s_t, z)$
    Store the collected environment transitions in the replay buffer $\mathcal{D}$
    for each training update step **do**
        Sample a mini-batch of trajectory segments $(s_{t:t+k}, a_{t:t+k-1}, z)$ from $\mathcal{D}$
        Update trajectory discriminator $q_\theta(z \mid s_{t:t+k})$ via cross-entropy loss
        Update state discriminator $q_{\theta'}(z \mid s_t)$ via cross-entropy loss
        Compute the dense intrinsic reward $r(s_{t:t+k}, z)$ for the current batch
        Update skill policy π_ω and its corresponding critic via DDPG
    end for
end for

C.1 PRE-TRAINING HYPERPARAMETER CONFIGURATION

The default hyperparameter values shared across all evaluated continuous-control environments during the pre-training phase are comprehensively listed in Table 3.

Reward Coefficients. The coefficients λ_1, λ_2, and λ_3 balance the magnitudes of the corresponding intrinsic-reward terms. We use these fixed coefficients to stabilize policy and critic training by keeping the empirical reward distribution approximately centered and symmetric around zero, thereby avoiding a persistent positive or negative reward bias and overly large value updates.

C.2 HIERARCHICAL META-CONTROLLER CONFIGURATION

For the downstream goal-reaching tasks (Section 4.4), we freeze the pre-trained low-level skill policy and train a high-level meta-controller $\pi_{\text{high}}(z \mid s)$ to compose these skills. To optimize the meta-controller, we utilize Proximal Policy Optimization (PPO) Schulman et al. (2017), a robust on-policy reinforcement learning algorithm well-suited for stable hierarchical control.

The meta-controller policy is parameterized as an MLP. During training, it collects rollouts across multiple parallel environments and updates its parameters using the PPO clipped objective alongside Generalized Advantage Estimation (GAE) Schulman et al. (2015). The complete hyperparameter configuration for the PPO meta-controller is detailed in Table 4.

Table 3: Hyperparameters for TIME Pre-training.

Category	Hyperparameter	Value
RL Agent (DDPG)	Total Training Frames	2×10^6
	MLP Hidden Dimension	1024
	Batch Size	1024
	Learning Rate (Actor & Critic)	1×10^{-4}
	Target Network Polyak Update (τ)	0.01
	Action Exploration Std. Dev.	0.2
	Action Exploration Std. Dev. Clip	0.3
Reward & Skill Space	Number of Latent Skills (N)	64
	Trajectory Segment Length (k)	16
	Intrinsic Reward Scaling Factor	1.0
	State Conditional Log Coef. (λ_1)	0.2
	State Conditional Mean Coef. (λ_2)	0.2
	Skill Prior Coef. (λ_3)	0.5
Discriminators	Trajectory Discriminator LR (q_θ)	5×10^{-5}
	State Discriminator LR ($q_{\theta'}$)	1×10^{-5}
	Discriminator MLP Hidden Dim	1024

Table 4: Hyperparameters for High-Level Meta-Controller (PPO).

Category	Hyperparameter	Value
Environment & Rollout	Parallel Environments	8
	Total Training Frames	1×10^6
	Rollout Steps per Env	256
	Skill Execution Length (T)	16
Network Architecture	MLP Hidden Dimension	256
	Optimizer	Adam
	Learning Rate	3×10^{-4}
PPO Optimization	Discount Factor (γ)	0.99
	GAE Parameter (λ)	0.95
	PPO Clip Ratio (ϵ)	0.2
	Number of Epochs	4
	Minibatch Size	512
	Value Loss Coefficient	0.5
	Entropy Coefficient	0.01
	Max Gradient Norm	0.5
	Advantage Normalization	True

C.3 Computational Resources

All experiments were conducted on a compute server equipped with a single NVIDIA RTX 4090 GPU (24GB VRAM), an AMD EPYC 7H12 64-Core Processor (allocated 32 vCPUs), and 125GB of RAM. To ensure optimal simulation throughput and overall performance, the experiments were configured to utilize an average of 16 vCPUs concurrently.

The evaluation and training runs reported in this paper were executed continuously over a period of approximately 14 to 16 days, corresponding to a total compute budget of roughly 384 GPU-hours. Note that this estimate solely accounts for the final reported results; additional computational resources were expended during the preliminary phases of debugging, hyperparameter tuning, and exploratory analyses not detailed in the final manuscript.

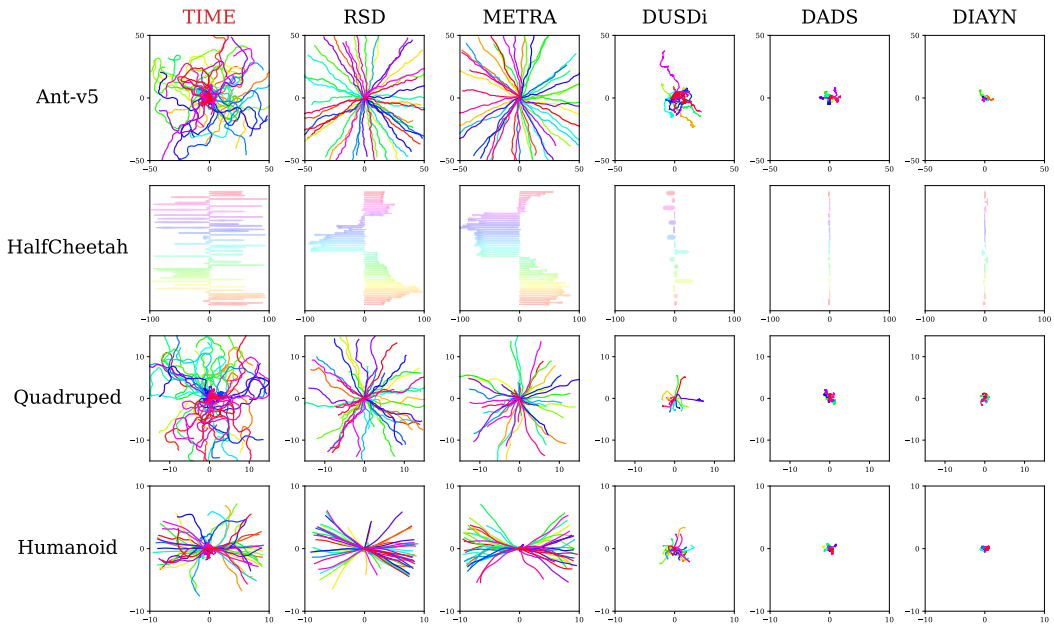


Figure 8: **Qualitative Analysis of Discovered Skills.** 2D spatial trajectories ($x - y$) generated by policies learned across four continuous-control environments. Distinct colors denote individual latent skills z. While TIME achieves extensive state coverage comparable to metric-aware baselines (METRA, RSD), it uniquely discovers temporally extended, overlapping macro-behaviors—a fundamental structural advantage driven by our trajectory-level MI objective.

D Ablation Studies and Extended Qualitative Results

In this section, we provide comprehensive ablation studies and extended visualizations to further analyze the internal mechanics and behavioral properties of the TIME framework. Specifically, our analysis encompasses the following key aspects:

- **Extended Behavioral Visualizations:** We present a full suite of qualitative skill rollouts generated by TIME across all four evaluated continuous-control environments (`Ant-v5`, `HalfCheetah`, `Quadruped`, and `Humanoid`). These are systematically compared against the trajectory profiles of the five baseline methods to highlight the qualitative advantages of trajectory-level skill discovery.
- **Impact of Skill Cardinality:** We perform an ablation on the number of latent skills by evaluating TIME variants trained with $N \in \{16, 32, 64\}$. This comparative analysis assesses how the dimensionality of the discrete skill space influences the granularity, coverage, and temporal diversity of the learned multi-step behaviors.
- **Impact of Trajectory Horizon:** We vary the trajectory segment length $k \in \{1, 8, 16, 32\}$ and evaluate the resulting skills on downstream goal-reaching tasks. This study examines the temporal context required to learn useful, long-horizon skills.
- **Efficacy of the Lipschitz Constraint:** To isolate the contribution of our theoretically motivated regularization, we compare the exploration performance (cumulative state coverage) of TIME against an unconstrained variant that omits the Lipschitz constraint on the trajectory discriminator. This evaluates the specific impact of Spectral Normalization in stabilizing the MI objective and preventing representation collapse during exploration.

D.1 Extended Behavioral Visualizations: TIME vs. Baselines

To complement the `Ant-v5` analysis in Section 4.2, Figure 8 extends our qualitative evaluation across all four continuous-control domains. Consistent with our earlier observations, foundational MI methods (DIAYN, DADS) and fine-grained state-centric approaches (DUSDi) suffer from severe

exploration bottlenecks, remaining highly localized around the initial state. While metric-aware baselines (METRA, RSD) effectively achieve broad spatial coverage, their learned skills (indicated by distinct colors) remain strictly disjoint and rigidly tied to specific spatial sectors.

In stark contrast, TIME consistently discovers temporally extended, highly overlapping macro-behaviors across all environments. Whether on `HalfCheetah`, `Quadruped`, or `Humanoid`, multiple distinct skills generated by TIME can traverse the same spatial regions. By optimizing mutual information over full trajectory segments rather than isolated states, TIME allows the agent to differentiate skills based on locomotion dynamics rather than spatial endpoints, successfully decoupling behavioral identity from spatial localization.

D.2 Impact of Skill Cardinality on Behavioral Granularity

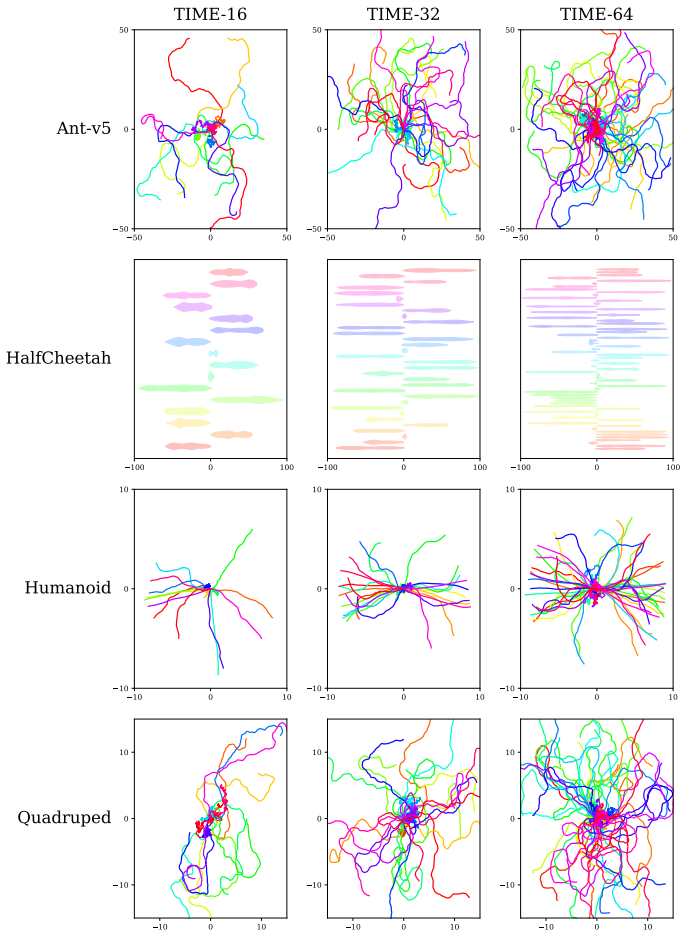


Figure 9: **Impact of Skill Cardinality on Trajectory Coverage.** 2D spatial rollouts generated by TIME on `Ant-v5` when trained with a latent skill space of $N \in \{16, 32, 64\}$. Increasing the size of the skill dictionary proportionally expands spatial coverage, allowing the agent to capture a denser and more expressive repertoire of complex, multi-step behaviors.

A critical hyperparameter in unsupervised skill discovery is the cardinality of the discrete skill space, denoted as N. To evaluate how the dimensionality of this latent space influences the granularity and coverage of the learned behaviors, we train variants of TIME with $N \in \{16, 32, 64\}$ skills. As illustrated in Figure 9, the empirical results demonstrate that achieving comprehensive coverage of the state space necessitates a relatively large number of skills (e.g., $N = 64$). Because the TIME framework explicitly optimizes for temporally extended, multi-step trajectories rather than simple state-reaching goals, each individual skill encodes a highly complex and specific structural movement pattern. Consequently, when the skill capacity is strictly bottlenecked ($N = 16$), the agent

is forced to learn a sparse set of coarse macro-behaviors, leaving large regions of the environment unexplored. Expanding the skill dictionary addresses this limitation by providing sufficient capacity to span the diverse variations of trajectory dynamics.

D.3 Impact of Trajectory Horizon on Downstream Performance

The trajectory segment length k controls the temporal horizon over which TIME distinguishes skill-conditioned behaviors. To assess its effect, we train TIME with $k \in \{1, 8, 16, 32\}$ and evaluate the learned skills on the `HalfCheetahGoal` and `QuadrupedGoal` downstream tasks. Table 5 reports the mean performance and standard deviation over five seeds.

Table 5: **Ablation of the trajectory horizon k.** Downstream performance is reported as mean $\pm$ standard deviation over five seeds. The reported configuration ($k = 16$) performs best on both tasks.

Horizon	HalfCheetahGoal	QuadrupedGoal
$k = 1$	0.414 ± 0.212	0.123 ± 0.072
$k = 8$	6.215 ± 1.521	3.890 ± 1.231
$k = 16$ (reported)	$\mathbf{7.400 \pm 0.452}$	$\mathbf{5.605 \pm 0.665}$
$k = 32$	7.026 ± 0.450	4.803 ± 0.657

The one-step variant ($k = 1$) performs poorly, confirming that transition-level discrimination does not provide sufficient temporal structure for reusable long-horizon skills. Increasing the horizon to $k = 8$ substantially improves performance, while the reported setting $k = 16$ achieves the strongest results on both tasks. Relative to $k = 8$, it improves performance by 19.1% on `HalfCheetahGoal` and 44.1% on `QuadrupedGoal`. Extending the horizon further to $k = 32$ yields a modest decline, suggesting that overly long segments can make skill discrimination and policy optimization more difficult. Overall, $k = 16$ provides the best balance between capturing temporally extended behavior and maintaining a tractable learning signal.

D.4 Efficacy of the Lipschitz-Constrained Discriminator

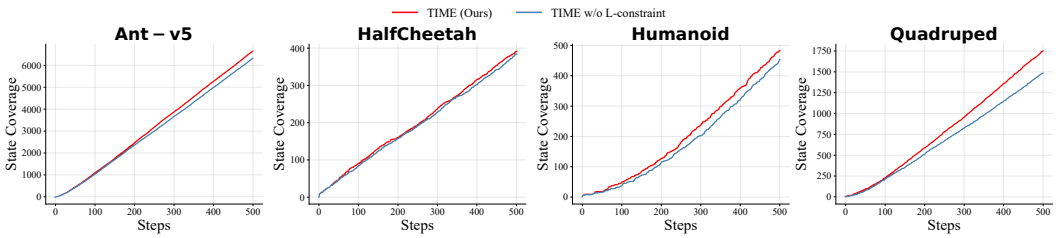


Figure 10: **Impact of the Lipschitz constraint on exploration.** Cumulative state coverage during pre-training for TIME and an ablated variant without spectral normalization (*TIME w/o L-constraint*). The constrained variant attains higher final coverage in all four environments.

In Section 3.2, we argued that controlling the smoothness of the trajectory discriminator prevents it from separating skills through arbitrarily small differences between trajectory segments. To isolate this effect, we compare TIME, which applies spectral normalization to f_ϕ, with an otherwise identical variant in which spectral normalization is removed. All results in this subsection are reported as mean $\pm$ standard deviation over five independent pre-training runs.

Figure 10 and Table 6 show that the constrained discriminator improves exploration consistently. Relative to the unconstrained variant, final coverage increases by 16.90% on `Quadruped`, 7.64% on `Humanoid`, 5.10% on `Ant-v5`, and 3.39% on `HalfCheetah`. The largest improvement occurs in `Quadruped`, where the higher-dimensional control problem provides more opportunities for the discriminator to exploit small, local trajectory differences.

We additionally verify that spectral normalization controls the discriminator quantity it is intended to regulate. For each trained discriminator, we report two complementary estimates. The product of the

Table 6: **Final state coverage with and without the Lipschitz constraint.** Bold values indicate the higher-coverage variant.

Environment	With constraint	Without constraint
`Quadruped`	$\mathbf{1750.7 \pm 54.2}$	1497.6 ± 50.9
`Humanoid`	$\mathbf{481.7 \pm 18.3}$	447.5 ± 18.4
`Ant-v5`	$\mathbf{6491.3 \pm 206.1}$	6176.1 ± 192.8
`HalfCheetah`	$\mathbf{393.4 \pm 10.5}$	380.5 ± 13.0

layer spectral norms, $\prod_l \|W_l\|_2$, upper-bounds the network Lipschitz constant when the intervening activations are 1-Lipschitz. We also compute the largest observed ratio

$$\max_{\tau,\tau' \in \mathcal{B}} \frac{\|f_\phi(\tau) - f_\phi(\tau')\|_2}{\|\tau - \tau'\|_2}, \tag{35}$$

over segment pairs sampled from the replay buffer $\mathcal{B}$, which provides an empirical lower bound on the global constant.

Table 7: **Measured Lipschitz constant of the trained trajectory discriminator.** Lower denotes the maximum observed ratio in Eq. equation 35; upper denotes $\prod_l \|W_l\|_2$.

Environment	Lower (w/)	Upper (w/)	Lower (w/o)	Upper (w/o)
`Ant-v5`	0.811 ± 0.026	1.079 ± 0.031	1.845 ± 0.065	3.803 ± 0.120
`HalfCheetah`	0.804 ± 0.026	1.065 ± 0.030	1.829 ± 0.069	3.756 ± 0.123
`Humanoid`	0.840 ± 0.027	1.114 ± 0.031	1.908 ± 0.069	3.928 ± 0.124
`Quadruped`	0.821 ± 0.025	1.089 ± 0.030	1.863 ± 0.066	3.833 ± 0.121

As shown in Table 7, spectral normalization keeps the upper estimate close to one (1.065–1.114), whereas removing it raises the estimate to 3.756–3.928. The empirical lower estimate exhibits the same pattern, increasing from 0.804–0.840 with the constraint to 1.829–1.908 without it. Together with the coverage results, these measurements support the proposed mechanism: a smoother trajectory discriminator makes locally trivial distinctions less rewarding and encourages the policy to produce more widely separated trajectories.