Focused Skill Discovery: Learning to Control Specific State Variables while Minimizing Side Effects

Anonymous authors
Paper under double-blind review

Keywords: Skill Discovery, Hierarchical Reinforcement Learning

Summary

Skills are essential for unlocking higher levels of problem solving. A common approach to discovering these skills is to learn ones that reliably reach different states, thus empowering the agent to control its environment. However, existing skill discovery algorithms often overlook the natural state variables present in many reinforcement learning problems, meaning that the discovered skills lack control of specific state variables. This can significantly hamper exploration efficiency, make skills more challenging to learn with, and lead to reward hacking on downstream tasks. We introduce a general method that enables these skill discovery algorithms to learn *focused skills*—skills that target and control specific state variables. Our approach improves state space coverage by a factor of three, unlocks new learning capabilities, and successfully mitigates reward hacking on downstream tasks.

Contribution(s)

1. This paper presents a general method allowing skill discovery algorithms to learn skills which control individual state variables.
   Context: Prior work Hu et al. (2024) explores a similar method, but is limited to skills that are discovered by maximizing mutual information.
2. We apply this method to three different skill discovery algorithms, showing that it improves exploration efficiency by a factor of three and unlocks new learning capabilities.
   Context: None
3. These skills can automatically avoid negative side effects when the goal is underspecified.
   Context: Prior work has discussed underspecified objectives but has not explored how skill discovery as a pretraining step can help mitigate these effects.
4. Compared to an existing method, we show that our approach can be significancy more effective at discovering skill which avoid side effects.
   Context: Hu et al. (2024) penalize skills by minimizing mutual information on non-target variables. We show that this is not sufficient in order to minimize side effects.
5. We show that our method is an effective pretraining step for mitigating reward hacking.
   Context: Prior work has explored skill discovery in safety-critical environments(Zhang et al., 2024; Kim et al., 2022).

Focused Skill Discovery: Learning to Control Specific State Variables while Minimizing Side Effects

Anonymous authors
Paper under double-blind review

Abstract

Skills are essential for unlocking higher levels of problem solving. A common approach to discovering these skills is to learn ones that reliably reach different states, thus empowering the agent to control its environment. However, existing skill discovery algorithms often overlook the natural state variables present in many reinforcement learning problems, meaning that the discovered skills lack control of specific state variables. This can significantly hamper exploration efficiency, make skills more challenging to learn with, and lead to reward hacking on downstream tasks. We introduce a general method that enables these skill discovery algorithms to learn *focused skills*—skills that target and control specific state variables. Our approach improves state space coverage by a factor of three, unlocks new learning capabilities, and successfully mitigates reward hacking on downstream tasks.

1 Introduction

Skills are learned behaviours that allow an agent to decompose a challenging problem into a set of easier sub-problems. In reinforcement learning (RL), a key challenge is *skill discovery*: finding a useful collection of skills, either from the agent's experiences or from an explicit task description. The main difficulty stems from needing to determine the best way to decompose a given problem before the agent has been told which problem to solve.

A popular approach to discovering skills is to find ones that can reliably reach different areas of the state space, allowing the agent to both control and explore its environment. Skills learned in this way facilitate control, since each skill will consistently bring the agent to the same area of the state space. They improve exploration, since set of states visited by each skill is unique. These skills can be discovered without any task description by maximizing the mutual information between the agent's selected skill and the state of the environment, or by aligning the representations of skills and states within a latent space.

However, while this approach is effective at generating a diverse set of behaviours, those behaviours tend not to provide the agent with much actual control over the individual state variables present in many reinforcement learning (RL) environments. For example, in a robot navigation task where the state is decomposed into the positions of the agent and other objects, skills might learn to navigate to different positions without learning to collect—or avoid— specific objects. The issue with these objectives is that skills are only encouraged to reach different states, regardless of the individual state variables that are changed in the process.

Learning a collection of *focused skills* that change one state variable at a time offers a number of benefits. First, it can significantly improve exploration when combining skills. For example, if one skill picks up a wrench while another skill picks up a hammer, then the agent can pick up both the tool and the hammer by executing one skill after the other. Second, these skills avoid making unnecessary changes to the environment, which make learning more efficient on downstream tasks where some state variables don't need to change. Lastly, by avoiding changes to non-target state variables, focused skills significantly limit the agent's tendency to pursue alternative objectives when the reward function is misspecified.

Our Contributions. We introduce a general method that enables skill discovery algorithms to control specific state variables. Rather than treating the agent's state as a unified whole, we leverage the factored state representation to learn skills focused on changing just one state variable at a time. We apply our method to a variety of skill discovery algorithms and show that focused skills are able to reach three times as many states for same number of skill executions compared to their un-focused counterparts. In downstream task, we show that these skills can solve problems that their un-focused skills struggle with and can automatically avoid side effects with no modification to the agent's goal. We compare our method to a recent skill-focusing method of Hu et al. (2024), highlighting that our method is more effective at learning focused skills, achieves stronger performance on downstream tasks and is less prone to reward hacking.

2 Background

We model interactions between an agent and its environment as a factored Markov decision process.

Definition 1. A **factored Markov decision process** $(\mathcal{S}, \mathcal{A}, P, R, \mu, \gamma)$ consists of a set of factored states $\mathcal{S} = \mathcal{S}^1 \times \cdots \times \mathcal{S}^N$, a set of actions $\mathcal{A}$, a transition probability function $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$, a reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, an initial state distribution μ and a discount factor $\gamma \in [0, 1)$.

Many existing RL domains are factored Markov Decision Processes (MDPs). For example, the state of a cart-pole balancing task might be specified in terms of the positions of the cart and pole, along with their respective linear and angular velocities. The state of a board game task might be specified as a list of locations and piece types. In this paper, we assume that each state variable s^i is in $\mathbb{R}^{d_i}$.

We use capital R to denote the reward of the Markov decision process (MDP), which is not observed when learning skills. We assume learning proceeds in two phases: first, the agent interacts with the environment reward-free, and has the opportunity to construct a set of skills (defined below). Next, the agent must learn a policy over its actions and skills that maximizes expected return, the discounted sum of future rewards from R.

We model skills using the options framework (Sutton et al., 1999), which we modify slightly to allow skills to depend on their interaction history. A *history* $h_t = (s_0, a_0, \ldots, a_{t-1}, s_t)$ is a sequence of state-action pairs that begins and ends with states. $\mathcal{H}$ denotes the set of all histories of finite length. For each state variable i, we also let $\mathcal{H}^i$ be the set of finite sequences of the form $(s_0^i, a_0, \ldots, a_{t-1}, s_t^i)$.

Definition 2. A **skill** $(\mathcal{I}, \pi, \beta)$ consists of an initiation set $\mathcal{I} \subseteq \mathcal{S}$, a policy $\pi : \mathcal{H} \times \mathcal{A} \to [0, 1]$ and a termination condition $\beta : \mathcal{H} \to [0, 1]$.

A skill might learn to pickup a tool, reach certain positions in a board game or swing a tennis racket. Defining skills in terms of histories —rather than environment states— allows us to conveniently model skills which end after a fixed number of timesteps. We assume that all skills terminate with probability one, and let $\Pr_{\pi,P}(\cdot|s)$ the distribution over states induced by following policy π from state s according to the transition probability function P until the skill terminates.

Definition 3. A **focused skill** $(\mathcal{I}, \pi, \beta, \mathcal{V})$ is a skill $(\mathcal{I}, \pi, \beta)$ together with a non-empty set of **target variables** $\mathcal{V} \subset \{1, \ldots, N\}$.

The target variables $\mathcal{V}$ are a proper subset of $\{1, \ldots, N\}$; otherwise the skill would not focus on any specific variables. Changes to non-target variables are called side effects.

Definition 4. The **side effects** of a focused skill $(\mathcal{I}, \pi, \beta, \mathcal{V})$ in state $s_0 \in \mathcal{I}$ the expected number of state variables $j \notin \mathcal{V}$ that differ between s_0 and the final state s_T of the skill:

$$\sum_{s_T \in \mathcal{S}} \Pr_{\pi,P}(s_T|s_0) \sum_{j \notin \mathcal{V}} \mathbf{1}\{s_T^j \neq s_0^j\}. \tag{1}$$

A focused skill might learn to pick up a tool *without* knocking over other objects, and would be said to have side effects if it knocks over objects while picking up the tool. We aim to minimize side effects with a side effects penalty.

Definition 5. Let $\mathcal{P}[N]$ be the power set of $\{1, \dots, N\}$. A function $\ell : \mathcal{H} \times \mathcal{P}[N] \to \mathbb{R}_{\geq 0}$ is called a **side effects penalty** if, for any history $h_t \in \mathcal{H}$ and $\mathcal{L} \in \mathcal{P}[N]$, ℓ satisfies the following two properties:

- If $h_0^j \neq h_t^j$ for some $j \in \mathcal{L}$, then $\ell(h, \mathcal{L}) > 0$,
- If $h_0^j = h_t^j$ for all $j \in \mathcal{L}$ then $\ell(h, \mathcal{L}) = 0$

When the side effects penalty depends only on the initial state s_0 and the final state s_t of a history h_t, we will overload notation and write $\ell(s_0, s_t, \mathcal{L})$ instead of $\ell(h_t, \mathcal{L})$.

Skill discovery is the process of finding a set of skills $\{(\mathcal{I}_z, \pi_z, \beta_z) : z \in \mathcal{Z}\}$, where $\mathcal{Z}$ is an index set used to label each skill. We will often refer to the index z as the skill *itself*, in order to avoid re-writing $(\mathcal{I}_z, \pi_z, \beta_z)$.

When skills are focused, it is convenient to decompose the skill index set as $\mathcal{Z} = \mathcal{Z}^1 \times \cdots \times \mathcal{Z}^N$. The elements z^i, referred to as *skill components*, indicate the effect that a skill z has on variable i. We can think of focused skills as skills $z = (z^1, \dots z^n)$ for which $z^i \neq 0$ if and only if i is a target variable of z. We would expect two focused skills z_1 and z_2 to have the same effect on a target variable i if $z_1^i = z_2^i$.

Without access to the MDP reward R, many skill discovery methods learn skills which maximize a skill reward.

Definition 6. A **skill reward** is a function $r : \mathcal{Z} \times \mathcal{H} \to \mathbb{R}$.

An example skill reward might incentivize one skill for making coffee and another skill for making toast. When the skill reward is a function only the initial state s_0 and final state s_t of a history h_t, we will overload notation and write $r(s_0, z, s_t)$ instead of $r(z, h_t)$. Skill rewards will always be denoted a lower case r and are agnostic to the MDP reward R.

3 Related Work

Unsupervised skill discovery aims to learn a collection of useful skills without an explicit task description. It can be thought of as a pretraining step to overcome challenges of exploration and data efficiency on downstream tasks. These methods typically assume that for all skills, the initiation set is equal to $\mathcal{S}$ and that skills terminate after a pre-defined number of timesteps.

Mutual-information-based skill discovery. A common approach to skill discovery is to maximize the mutual information be tween skills and states, allowing the agent to control its environment. Gregor et al. (2017)'s Variational Intrinsic Control (VIC) algorithm maximizes the conditional mutual information $I(Z; S_T|s_0)$, where s_0 is a starting state, Z is sampled from a distribution ν over $\mathcal{Z}$ and S_T is sampled by following skill Z from start state s_0 until termination. Since this mutual information is challenging to compute, VIC maximizes the lower bound developed by Barber & Agakov (2003):

$$\begin{aligned} I(Z; S_T|s_0) &= -H(Z|S_T, s_0) + H(Z|s_0) && (2) \\ &= \mathbb{E}[\log p(Z|S_T, s_0)] - \mathbb{E}[\log \nu(Z|s_0)] && (3) \\ &\geq \mathbb{E}[\log d(Z|S_T, s_0)] - \mathbb{E}[\log \nu(Z|s_0)]. && (4) \end{aligned}$$

Here $H(\cdot)$ and $H(\cdot|\cdot)$ are the entropy and conditional entropy. The *skill discriminator* $d(Z|S_T, s_0)$ is a learned estimate of $p(Z|S_T, s_0)$, the posterior distribution over skills. This lower bound gets tighter as d approaches the posterior distribution over skills. Therefore, the goal of VIC is to learn *both* the skill policies *and* a good skill discriminator. As shown by Gregor et al., this can be achieved by training a set of skill policies to maximize the reward

$$r_{\text{VIC}}(s_0, z, s_T) = \log(d(z|s_T, s_0)) - \log \nu(z|s_0), \tag{5}$$

while updating d to estimate the posterior distribution over skills. In a similar spirit, Eysenbach et al.'s A Diversity Is All You Need (DIAYN) algorithm maximizes the mutual information between skills and all states states that the skill visits during its execution, discovering skills with the reward

$$r_{\text{DIAYN}}(s_0, z, s_t) = \log(d(z|s_t)) - \log \nu(z). \tag{6}$$

A wide range of innovations to this basic approach have been studied in the literature (Achiam et al., 2018; Sharma et al., 2020; Campos et al., 2020; Hansen et al., 2020; Zhang et al., 2021; Kim et al., 2021; Liu & Abbeel, 2021). While maximizing mutual information produces diverse skills, Park et al. (2022) observe that these approaches struggle to discover skills for covering long distances in the state space.

Lipschitz-based skill discovery. Recent work has explored Lipschitz constraints to discover skills which maximize the distance travelled during skill execution. Lipschitz-constrained Skill Discovery (LSD) (Park et al., 2022) learns a Lipschitz-constrained function $\phi : \mathcal{S} \to \mathcal{Z}$ (i.e. $\|\phi(s') - \phi(s)\| \leq \|s' - s\|$ for all $s, s' \in \mathcal{S}$) and a set of skills that maximize the reward

$$r_{\text{LSD}}(s_0, z, s_T) = \langle \phi(s_T) - \phi(s_0), z \rangle, \tag{7}$$

where $\langle \cdot, \cdot \rangle$ is the inner product. This reward encourages $\phi(s_T) - \phi(s_0)$ to be large in the direction of the skill z. Due to the Lipschitz constraint, maximizing $\phi(s_T) - \phi(s_0)$ requires the Euclidean distance $\|s_T - s_0\|$ to be large. Lipschitz-based skill discovery algorithms can learn skills that maximize non-Euclidean notions of distance (Park et al., 2023) and have recently achieved state-of-the-art results in pixel-based environments (Park et al., 2024).

Leveraging State Variables in Skill Discovery. Several methods have used state variables to improve skill discovery. Skills can be encouraged to achieve subgoals on specific variables (Lee et al., 2020; Choi et al., 2023) or to cause specific interactions between state variables (Hu et al., 2022; Wang et al., 2024). While these algorithms improve control of state variables, they do not penalize side effects to other state variables. The benefits of skills which minimize side effects has been studied in the planning literature Allen et al. (2021).

The most relevant point of comparison to our work is Disentangled Unsupervised Skill Discovery (DUSDi) (Hu et al., 2024), a method designed to make mutual-information-based skill discovery algorithms learn focused skills. DUSDi maximizes the mutual information between skills and values on target variables while minimizing the mutual information between skills and values of non-target variables. However, we show that this objective is not always effective at learning focused skills because the mutual information penalty does not explicitly minimize side effects. We provide a more detailed comparison with DUSDi in Section 4.1 and compare the two methods in our experiments. A second distinction between DUSDi and our approach is that our approach is compatible with non mutual-information-based skill discovery algorithms, such as Lipschitz-based algorithms.

Reward Hacking. In downstream tasks, specifying reward functions can be challenging. Reward hacking (Amodei et al., 2016; Skalse et al., 2022; Laidlaw et al., 2025) studies issues that arise when the MDP reward function is only an approximation of the true objective. Kim et al. (2022) and Zhang et al. (2024) propose skill discovery algorithms to mitigate reward hacking in robotics tasks by forcing skills to stay within a pre-defined "safe" region of the state space with high probability. However, this constraint may be infeasible in stochastic environments. Moreover, without minimizing side effects in safe states, these skills may lead to undesired outcomes that are not safety-critical. For instance, it may be safe for the skills of a house-cleaning robot to move furniture around but would be undesirable for the skills to move furniture unnecessarily. An interesting are of future work might be to combine our approach with those of Kim et al. and Zhang et al., learning skills that satisfy hard safety constraints while minimizing side effects in safe states.

Algorithm 1 Focused Variational Intrinsic Control

```
for episode = $1, M$ do
    Sample $s_0$ from the initial state distribution $\mu$
    Sample skill $z$ from $\nu(\cdot|s_0)$
    Follow policy $\pi_z$ until termination state $s_T$
    for $i$ in $\mathcal{V}_z$ do
        Update the skill discriminator $d_i$ from $(s_0^i, z^i, s_T^i)$
    end for
    Calculate the reward $r_{\text{focused-VIC}}(s_0, z, s_T)$ using Equation 9.
    Update $\pi_z$ to maximize $r_{\text{focused-VIC}}$
    Update option prior $\nu(\cdot|s_0)$ based on $r_{\text{focused-VIC}}$
end for
```

4 Focused Skill Discovery

In this section, we present *focused skill discovery*: a general method of transforming existing skill discovery algorithms into ones that discover focused skills. Our method can be applied to any skill discovery algorithm that learns skills in a factored Markov decision process using a skill reward. We first outline our general concept and then illustrate how it can be applied to mutual-information-based and Lipschitz-based skill discovery algorithms. Lastly, we describe the key differences between our method and DUSDi (2024) which allow our method to minimize side effects more effectively.

We create a focused skill discovery algorithm from a baseline skill discovery algorithm by constructing a focused skill reward from the baseline algorithm's skill reward. Focused skill rewards incentivize focused skills to control their target variables while minimizing side effects to other state variables. The original skill discovery algorithm is *instantly transformed* into a focused skill discovery algorithm by replacing the baseline skill reward with the focused skill reward.

Focused skill rewards consist of two terms: one which encourages focused skills to manipulate their target variables and another which penalizes side effects. The first term is a sum of reward components $r_i : \mathcal{Z} \times \mathcal{H}^i \to \mathbb{R}$, each generated by restricting a copy of the original skill reward to $\mathcal{Z} \times \mathcal{H}^i$. The second term is a side effects penalty ℓ, which discourages the skill from affecting non-target variables. Together, these terms define a focused reward

$$r_{\text{focused}}(z, h_t) = \left[\sum_{i \in \mathcal{V}_z} r_i(z, h_t^i)\right] - \ell(h_t, \mathcal{V}_z^{\mathsf{c}}). \tag{8}$$

Example 1 (Focused skill rewards for VIC and DIAYN) To create focused skill rewards for VIC and DIAYN, we learn a separate skill discriminator d_i for each target variable i, which predicts skills based on the values of state variable i. This encourages skills to reliably reach different values of their respective target variables. Starting with the VIC reward in Equation 5, we obtain the focused VIC reward

$$r_{\text{focused-VIC}}(s_0, z, s_T) = \left[\sum_{i \in \mathcal{V}_z} \log(d_i(z|s_T^i, s_0^i)) - \log(\nu(z|s_0^i))\right] - \ell(s_0, s_T, \mathcal{V}_z^{\mathsf{c}}). \tag{9}$$

Similarly, the DIAYN reward in Equation 6 leads to the focused DIAYN reward

$$r_{\text{focused-DIAYN}}(s_0, z, s_t) = \left[\sum_{i \in \mathcal{V}_z} \log(d_i(z|s_t^i)) - \log(\nu(z))\right] - \ell(s_0, s_t, \mathcal{V}_z^{\mathsf{c}}). \tag{10}$$

Example 2 (Focused skill reward for LSD) A focused skill reward for Lipschitz-constrained skill discovery is constructed by learning a Lipschitz-constrained function $\phi_i : \mathcal{S}^i \to \mathcal{Z}$ for each target variable i. From the LSD reward in Equation 7, we derive the focused LSD reward

$$r_{\text{focused-LSD}}(s_0, z, s_T) = \left[\sum_{i \in \tau_z} \langle \phi_i(s_0^i) - \phi(s_T^i), z \rangle \right] - \ell(s_0, s_T, \mathcal{V}_z^{\mathsf{c}}). \tag{11}$$

This reward motivates skills to maximize the distance between initial and final states on their target variables.

When a skill reward is substituted for a focused skill reward in a skill discovery algorithm, we obtain a *focused skill discovery algorithm*—one which is capable of learning focused skills. Algorithm 1 applies this transformation to Variational Intrinsic Control, and algorithms for the focused versions of DIAYN and LSD are provided in Appendix A.

4.1 Comparison with DUSDi

While both focused skill discovery and DUSDi (2024) aim to learn focused skills, they differ in two key ways. First, the DUSDi objective is defined through mutual information, while our method remains agnostic for how the skill reward is generated. This allows our method to be applied to a wider range of skill discovery algorithms, such as LSD (2022). Second, our methods differ in how they mitigate side effects.

DUSDi mitigates side effects by minimizing the mutual information $I(Z^i, S^{\neg i})$ between skill components Z^i and the values of the remaining state variables, $S^{\neg i} = (S_1, \ldots, S_{i-1}, S_{i+1}, \ldots S_N)$. However, this mutual information does not necessarily penalize side effects; it only encourages skills to have the *same effects* on non-target state variables.

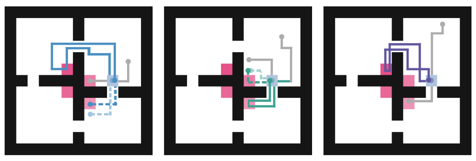
To understand the discrepency between side effects and "same effects", consider the case of a robotic arm with two state variables: one for the position of the arm's gripper and another indicating whether a cup of coffee on a nearby table has been knocked over. If i is the position variable of the arm, then $I(Z^i, S^{\neg i})$ can be minimized if *all* skill components for the arm gripper knock over the cup of coffee, because the components are indistinguishable from one another on the coffee cup variable if they all knock the cup over.

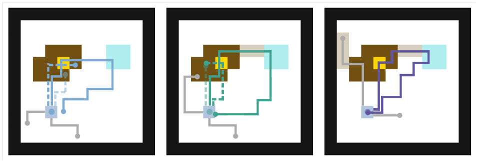
While this may seem like a contrived example, the issues with penalizing side effects through minimizing mutual information become apparent when state variables are *entangled*—i.e. a change in one state variable is correlated with changes in other state variables. In this case, during skill discovery, the skill discriminator may begin to encourage focused skills to make unnecessary changes their non-target state variables in order to minimize mutual information. This behaviour is observed empirically in our experiments.

On the other hand, focused skill discovery explicitly penalizes side effects, making it impossible to maximize a focused skill reward if skills cause unnecessary changes to their non-target variables. In our experiments below, we will study how this leads to substantial differences between skills learned with DUSDi and skills learned focused skill discovery.

5 Experiments

We seek to develop a principled understanding of how skills learned with focused skill discovery compare to baseline skill discovery algorithm and focused skills learned with DUSDi. We consider three gridworld environments that differ in their types of interactions and their opportunities for side effects. Qualitatively, we find that our method consistently learns skills that minimize side effects. Focused skills dramatically outperform unfocused skills across all downstream tasks we consider, and can reach three times as many states in the same amount of skill execution steps as unfocused


(a) FourRooms


(b) MudWorld

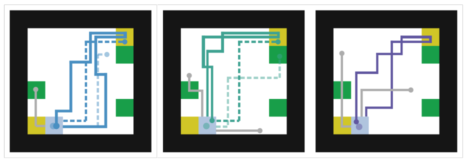

(c) ForageWorld

Figure 1: Skill trajectories for VIC (Blue), DIAYN (green) and LSD (purple) algorithms. Solid lines are trajectories from focused skills, dashed lines are from DUSDi skills and grey lines are from the baseline algorithms. Skills start in the blue square and terminate at the circles. Focused skills learn to collect objects and return to the start state in order to minimize side effects.

skills. The difference in the side effect penalties between focused skill discovery and DUSDi leads to significantly different skills learned from the same baseline algorithms. When side effects must be minimized to accomplish the task, our skills outperform DUSDi, and are the only ones that can accomplish the task in our most challenging environment. When agents optimize a proxy reward which does not penalize side effects, all focused skill discovery algorithms automatically avoid making unnecessary changes, while all other methods fall short.

We apply focused skill discovery to three baseline algorithms: VIC (Gregor et al., 2017), DIAYN (Eysenbach et al., 2018) and LSD (Park et al., 2022), obtaining three focused skill discovery algorithms: Focused VIC, Focused DIAYN and Focused LSD. These baselines were selected because they are have similarities with a broad range of other skill discovery algorithms, while each being unique from one another. In particular, the difference between DIAYN and VIC is largely whether the skill reward is sparse or dense, which allows us to examine the effect of sparse and dense focused skill rewards. LSD takes an entirely different approach to discovering skills. We compare these methods to DUSDi VIC and DUSDi DIAYN, the algorithms obtained by applying DUSDi VIC and DIAYN. There is no analog of DUSDi for LSD since it only works for mutual-information-based skill discovery algorithms. We compare these focused skills with the original VIC, DIAYN and LSD algorithms.

Environments. We learn skills in the FourRooms, ForageWorld and MudWorld environments shown in Figure 1. Each environment has four "primitive" actions that move the agent in each of the cardinal directions. When an agent selects an action to move in one direction, it moves in one of the other three directions with probability 0.1.

FourRooms has the same map as Sutton et al. (1999), with four added tools that the agent can pick up (shown in pink). It contains five state variables: one for the agent's position and one binary variable for each tool indicating whether it has been picked up. This environment is quite challenging to explore completely: a sixteenth of the state space (i.e. the set of states where all tools have been collected) is only accessible once agent has navigated through all four rooms.

ForageWorld contains yellow resources for an agent to collect, as well as delicate plants (green cells) that should be avoided. If the agent walks over plant, the cell is destroyed and does not regenerate. The states in ForageWorld contain six state variables: one variable for the agent's position, two integer-valued variables for the quantity of each resource collected and three binary variables indicating which plants have been destroyed. The challenge for skills in this environment is not exploration, but how to control the position and resource variables while *avoiding* damage to the plants.

The MudWorld environment contains a patch of mud (brown), a piece of treasure (yellow), and a puddle (teal). The agent becomes muddy if it moves to the mud patch and, when it is muddy, it tracks mud onto the vacant (white) cells. The agent can become clean again if it moves into the puddle. We simplify the agent state in the MudWorld environment to contain four state variables: one for the agent's position, one indicating whether the agent is muddy, one variable that indicates whether the treasure has been collected and a variable for the number of muddy cells. The agent has only partial observability on its environment it has sufficient information to find an optimal policy in the downstream task. Unlike FourRooms and ForageWorld, the state variables in MudWorld are *entangled*; it is impossible for the agent to collect treasure without becoming muddy. Once in the mud-patch, the only way to become less muddy is to move to the puddle, which increases the number of muddy cells.

Implementation. We trained all policies using tabular Q-learning with ϵ-greedy exploration and a discount factor of 0.99. We trained a set of skill policies $\mathcal{Z}$ with $|\mathcal{Z}| = 16$ skills in each skill discovery algorithm. We assign each focused skill a single target variable (i.e. $|\mathcal{V}_z| = 1$). For the focused skill discovery algorithms and DUSDi algorithms, we assign two skills to target each tool, resource or treasure and use the remaining skills (8 in FourRooms, 12 in ForageWorld, 14 in MudWorld) to control the agent's position. All skills were allowed to execute for a maxmium of 40 steps in FourRooms and 20 steps in ForageWorld and MudWorld.

For the side effects penalty of focused skill rewards, we apply a weighted 2-norm $\|\cdot\|_{\boldsymbol{\lambda}}$, where $\boldsymbol{\lambda} \in \mathbb{R}^{d_1} \times \cdots \times \mathbb{R}^{d_N}$, to penalize changes on non-target variables. More precisely, for history $h_t = (s_0, a_0, \ldots, a_{t-1}, s_t)$ and target variables $\mathcal{V}$, $\ell(h, \mathcal{V}_z^{\mathsf{c}}) = \|s_0 - s_t\|_{\boldsymbol{\lambda}|_{\mathcal{V}=0}}$, where $\boldsymbol{\lambda}_{\mathcal{V}=0}$ is the weight matrix obtained by setting the values of $\boldsymbol{\lambda}$ on target variables equal to zero, thus avoiding penalties on target variables. Using a weight matrix to control the penalty strength is useful since each state variable may have a different range of values. For each focused skill discovery algorithm, we chose a *single hyperparemeter* $\lambda > 0$ and set λ_{ij} equal to λ divided by the maximum 2-norm between any two values on variable j. So, for a state variable whose values range from 0 to k, the weight was λ/k. This ensures that the strength of the side effect penalty for each variable is between 0 and λ. We used $\lambda = 10$ for Focused VIC and Focused DIAYN and $\lambda = 2$ for Focused LSD, discussing the effects for different values of λ in Section 5.5.

Additional training details are available in Appendix B.

5.1 Qualitative Analysis of Learned Skills

Figure 1 shows skill trajectories sampled from each skill discovery algorithm. As expected, all focused skill discovery algorithms find skills that change control individual state variables while minimizing side effects. The DUSDi skills are less effective at avoiding side effects; in ForageWorld and MudWorld, the skills that are designed to control the resource and treasure state variables do so while damaging nearby green cells and without cleaning off after stepping into the mud patch. This is consistent with the differences between focused skill discovery and DUSDi described in Section 4.1.

5.2 Exploration Efficiency

One major benefit of focusing skills is that they can provide the agent with a way to structure its exploration of the environment. Equipped with skills for each tool in the FourRooms environment, agents are capable of traversing across the four different rooms in just four skill execution steps. This drastically improves exploration efficiency. To measure exploration efficiency, we used the Area Under the Curve (AUC) of the State Coverage Fraction vs. Skill Chain Length graph, shown in Figure 2. This measures how many states a set of skills is able to reach for a given number of skill executions. To compute the state coverage fraction of a skill chain of length l, we measured the

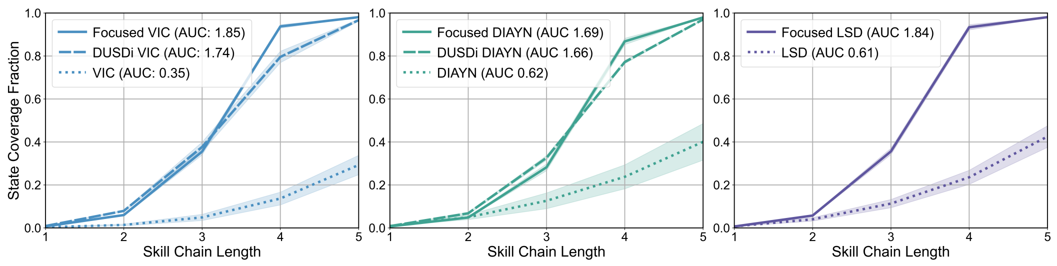


Figure 2: State coverage in the FourRooms environment. Focused skills explore three times more efficiently than un-focused skills, as measured by the Area Under the Curve (AUC).

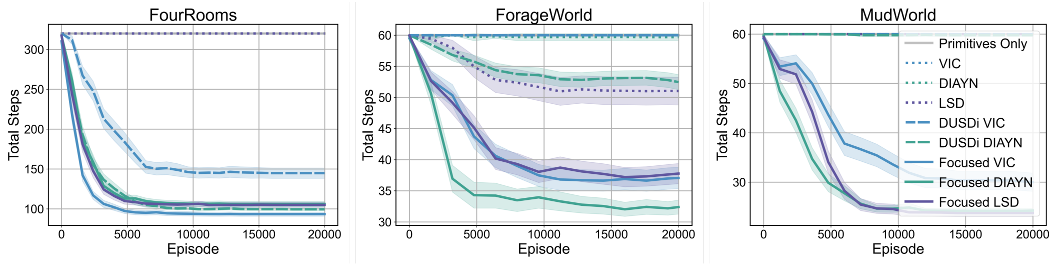


Figure 3: Learning performance in downstream tasks. Focused skills (solid lines) lead to faster learning and are the only ones which can accomplish the task in MudWorld.

fraction of unique final states that could be reached after executing all possible skill chain combinations of length l from a given start state s_0. We plotted the mean state coverage fraction and 90% confidence intervals over 10 random start states. In all cases, the exploration benefits of focusing skills are remarkable: exploration improves by 5.3× for VIC (AUC 1.86 vs AUC 0.35), 2.7× for DIAYN (AUC 1.69 vs AUC 0.62) and 3.0× for LSD (AUC 1.84 vs. 0.61). This corresponds to an average improvement in exploration efficiency of 3.67× across these three methods.

The exploration efficiency of the skills learned with DUSDi and our method are comparable. This was expected, since both objectives explicitly learn skills that can pick up each tool, facilitating the agent's exploration of hard-to-reach states. Because DUSDi is limited to mutual-information-based skill discovery algorithms, there is no DUSDi version of LSD.

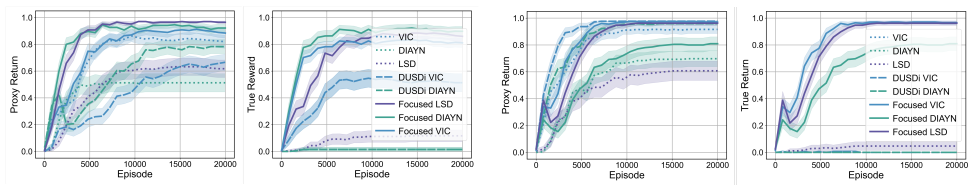
5.3 Performance on Downstream Tasks

With a set of focused skills that facilitate structured exploration while minimizing side effects, an agent is much more capable to solve downstream tasks. We define downstream tasks as follows. In all environments, the agent's goal is to collect all of the tools or resources and navigate to the bottom-right corner. In the FourRooms environment, the agent's goal is to pick up all four tools and navigate to the bottom-right corner. In the ForageWorld environment, the agent must collect two units of both resources without destroying any delicate plants. In the MudWorld environment, the agent must collect the treasure while tracking fewer than five mud cells. In all tasks, the agent starts in the top-left corner and receives a (sparse) reward of +1 for accomplishing the task. The agent can take up to 320 steps in FourRooms (a maximum of 8 skill execution steps) and up to 60 steps in ForageWorld and MudWorld (a maximum of 3 skill execution steps). We conduct 50 independent training runs for each of the agents we consider, plotting the mean and 90% confidence intervals of our results in Figure 3.

Without focused skills, agents fail to accomplish their goals. Moreover, while the performance of DUSDi and focused skills are comparable in FourRooms, focused skills lead to significantly better results in ForageWorld and MudWorld, where an agent must minimize side effects in order to accomplish its goal. This is consistent with our observations from Figure 1, where DUSDi skills seem to have the *same effects* on non-target state variables, while the focused skills learn to minimize side effects. In the MudWorld environment, only the agents with focused skills can accomplish the task.

5.4 Mitigating Reward Hacking

In addition to exploration and learning benefits, focused skill discovery is an effective pretraining step for minimizing reward hacking. Figure 4 shows the results of downstream tasks in ForageWorld and MudWorld, where the agent is not given the true task rewards from Section 5.3, but instead rewards proxy which do not penalize side effects. In ForageWorld, the proxy reward gives a score of +1 for reaching the bottom-right corner after picking up all of the resources, regardless of the number of plants destroyed. In the MudWorld task, the proxy reward assigns a score of +1 for reaching the bottom-right corner after picking up the treasure and cleaning itself off, regardless of the number of tracked mud cells. Across both environments and all baseline algorithms, focused skill discovery illustrates a striking ability to accomplish the true task even if it is only given the proxy reward. In contrast, the DUSDi and un-focused skills maximize the proxy reward at the expense of the true reward.



(a) ForageWorld (b) MudWorld

Figure 4: Reward hacking behaviours in the ForageWorld and MudWorld when agents are trained with a proxy reward instead of the true reward. Focused skills (solid lines) are the only ones which mitigate reward hacking in MudWorld.

5.5 Ablation Studies

The tendency for focused skills to avoid side effects in our experiments was controlled by the side effects hyperparameter λ. To better understand impact of the penalty strength, we re-ran the MudWorld task for focused skills learned with varying penalty strengths. We chose values of λ that were half and four times the values of λ that were used in our previous experiments, comparing these to focused skill discovery algorithms without a side effects penalty ($\lambda = 0$). In general, higher values of λ tend to lead to better performance and the methods were relatively robust to the penalty

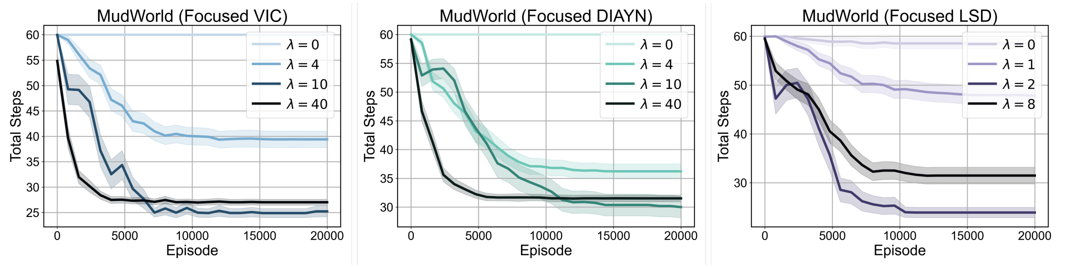


Figure 5: Impact of the side effect penalty strength on focused skills in the MudWorld domain. Skills are not effective when there is no side effects penalty (i.e. when $\lambda = 0$)..

strength. Without the side effects penalty, agents with focused skills are not able to accomplish the MudWorld task.

6 Discussion & Conclusion

We presented focused skill discovery, a general approach that allows skill discovery algorithms to learn focused skills. Focused skills are useful not just for minimizing side effects but also for improving exploration and learning efficiency. We showed that focused skill can dramatically improve both an agent's exploration and learning capabilities, while also providing a natural buffer for reward hacking when downstream rewards are under-specified. Compared to a recently proposed approach to discovering focused skills, our method showed substantial improvement, particularly in an environment where states were entangled. We are excited to keep exploring the benefits of focused skills.

There are a number of interesting avenues of future work to consider. Empirically, it would be interesting to scale these experiments up to larger environments and test our method for sets of continous skills. Since the baseline methods we considered scale well to these settings, we are confident that focused skill discovery will also scale. While pre-defined state variables play a key role in our approach, we hope to extend this idea to include other kinds of state abstractions. There is also interesting theoretical territory to explore at the intersection of reward hacking and focused skill discovery. It seems plausible that in some cases, focused skill discovery could be guaranteed to lead to improvements an agent's capabilities while at the same time mitigating reward hacking.

References

Joshua Achiam, Harrison Edwards, Dario Amodei, and Pieter Abbeel. Variational option discovery algorithms. *arXiv preprint arXiv:1807.10299*, 2018.

Cameron Allen, Michael Katz, Tim Klinger, George Konidaris, Matthew Riemer, and Gerald Tesauro. Efficient black-box planning using macro-actions with focused effects. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pp. 4024–4031, 2021.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety, 2016. URL https://arxiv.org/abs/1606.06565.

David Barber and Felix Agakov. The im algorithm: a variational approach to information maximization. In *Proceedings of the 17th International Conference on Neural Information Processing Systems*, NIPS'03, pp. 201–208, Cambridge, MA, USA, 2003. MIT Press.

Víctor Campos, Alexander Trott, Caiming Xiong, Richard Socher, Xavier Giró-i Nieto, and Jordi Torres. Explore, discover and learn: Unsupervised discovery of state-covering skills. In *International conference on machine learning*, pp. 1317–1327. PMLR, 2020.

Jongwook Choi, Sungtae Lee, Xinyu Wang, Sungryull Sohn, and Honglak Lee. Unsupervised object interaction learning with counterfactual dynamics models. In *Workshop on Reincarnating Reinforcement Learning at ICLR 2023*, 2023. URL https://openreview.net/forum?id=dYjH8Nv81K.

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*, 2018.

Karol Gregor, Danilo Jimenez Rezende, and Daan Wierstra. Variational intrinsic control. *arXiv*, art. 1611.07507, 2017.

Steven Hansen, Will Dabney, Andre Barreto, David Warde-Farley, Tom Van de Wiele, and Volodymyr Mnih. Fast task inference with variational intrinsic successor features. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=BJeAHkrYDS.

Jiaheng Hu, Zizhao Wang, Peter Stone, and Roberto Martín-Martín. Disentangled unsupervised skill discovery for efficient hierarchical reinforcement learning. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 76529–76552. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/8c263f70550cc7d69dba3fc170a23e77-Paper-Conference.pdf.

Xing Hu, Rui Zhang, Ke Tang, Jiaming Guo, Qi Yi, Ruizhi Chen, Zidong Du, Ling Li, Qi Guo, Yunji Chen, et al. Causality-driven hierarchical structure discovery for reinforcement learning. *Advances in neural information processing systems*, 35:20064–20076, 2022.

Jaekyeom Kim, Seohong Park, and Gunhee Kim. Unsupervised skill discovery with bottleneck option learning. In *International Conference on Machine Learning*, pp. 5572–5582. PMLR, 2021.

Sunin Kim, Jaewoon Kwon, Taeyoon Lee, Younghyo Park, and Julien Perez. Safety-aware unsupervised skill discovery. In *2023 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022.

Cassidy Laidlaw, Shivam Singhal, and Anca Dragan. Correlated proxies: A new definition and improved mitigation for reward hacking. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=msEr27EejF.

Youngwoon Lee, Jingyun Yang, and Joseph J. Lim. Learning to coordinate manipulation skills via skill behavior diversification. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ryxB21BtvH.

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

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

Seohong Park, Jongwook Choi, Jaekyeom Kim, Honglak Lee, and Gunhee Kim. Lipschitz-constrained unsupervised skill discovery. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=BGvt0ghNgA.

Seohong Park, Kimin Lee, Youngwoon Lee, and Pieter Abbeel. Controllability-aware unsupervised skill discovery. In *International Conference on Machine Learning*, pp. 27225–27245. PMLR, 2023.

Seohong Park, Oleh Rybkin, and Sergey Levine. METRA: Scalable unsupervised RL with metric-aware abstraction. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=c5pwL0Soay.

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

Joar Skalse, Nikolaus HR Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward hacking. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pp. 9460–9471, 2022.

Richard S. Sutton, Doina Precup, and Satinder Singh. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112(1):181–211, 1999. ISSN 0004-3702. DOI: https://doi.org/10.1016/S0004-3702(99)00052-1. URL https://www.sciencedirect.com/science/article/pii/S0004370299000521.

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. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, volume 37, 2024.

Heng Zhang, Gokhan Solak, Gustavo J. G. Lahr, and Arash Ajoudani. Srl-vic: A variable stiffness-based safe reinforcement learning for contact-rich robotic tasks. *IEEE Robotics and Automation Letters*, 9(6):5631–5638, 2024. DOI: 10.1109/LRA.2024.3396368.

Jesse Zhang, Haonan Yu, and Wei Xu. Hierarchical reinforcement learning by discovering intrinsic options. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=r-gPPHEjpmw.

A Algorithms for Focused DIAYN and LSD

In this section, we describe the additional algorithms used in our experiments.

Algorithm 2 Focused Diversity is All You Need

```
for episode = $1, M$ do
    Sample $s_0$ from the initial state distribution $\mu$
    Sample skill $z$ from $\nu$
    for $t = 1, t_{\text{max}}$ do
        Select action $a_t \sim \pi_z(\cdot|s_t)$
        Observe $s_{t+1} \sim p(\cdot|s_t, a_t)$
        Calculate the skill reward $r_{\text{focused-DIAYN}}(s_0, z, s_{t+1})$ using Equation 10.
        Update policy $\pi_z$
        for $i \in \mathcal{V}_z$ do
            Update the skill discriminators $d_i$ from $(z^i, s^i_{t+1})$
        end for
    end for
end for
```

Algorithm 3 Focused Lipschitz Constrained Skill Discovery

```
for episode = $1, M$ do
    Sample $s_0$ from the initial state distribution $\mu$
    Sample skill $z$ from $\nu$
    Follow policy $\pi_z$ until termination state $s_T$
    for $i \in \mathcal{V}_z$ do
        Update $\phi_i$ using $(s^i_0, z^i, s^i_T)$
    end for
    Calculate the reward $r_{\text{focused-LSD}}(s_0, z, s_T)$ using Equation 11.
    Update $\pi_z$ to maximize $r_{\text{focused-LSD}}$
end for
```

B Additional Training Details

Following prior work (Gregor et al., 2017; Eysenbach et al., 2018; Park et al., 2022), the skill distribution ν way uniform over all skills and held constant during training. The skill discriminators for all mutual-information-based skill discovery algorithms made predictions using an exponentially weighted moving average of the previous samples.

For LSD, the state representation function ϕ was a linear function of the state trained with stochastic gradient descent. Following Park et al., we used spectral normalization (Miyato et al., 2018) to satisfy the Lipschitz constraint during training.

For the DUSDi algorithms, we used a penalty strength of 0.1 as in Hu et al. (2024). We also experimented with penalty strengths of 0.01, 0.2, and 0.5, finding that a penalty strength 0.1 lead to the best performance. Higher values of λ did not effect the performance of DUSDi in the MudWorld environment.