Safety-Aware Unsupervised Skill Discovery

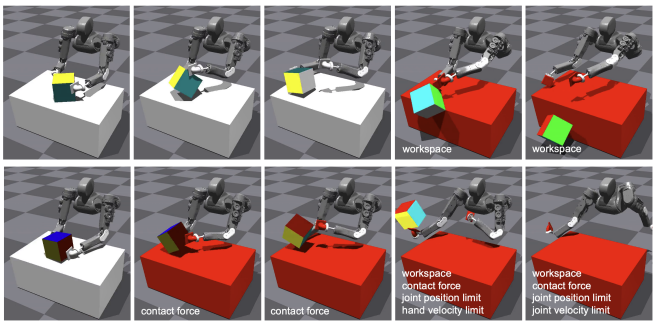
Sunin Kim[1], Jaewoon Kwon[1], Taeyoon Lee[1], Younghyo Park[1] and Julien Perez[2]

***Abstract*— Programming manipulation behaviors can become increasingly difficult with a growing number and complexity of manipulation tasks, particularly in a dynamic and unstructured environment. Recent progress in unsupervised skill discovery algorithms has shown great promise in learning an extensive collection of behaviors without extrinsic supervision. On the other hand, safety is one of the most critical factors for real-world robot applications. As skill discovery methods typically encourage exploratory and dynamic behaviors, it can often be the case that a large portion of learned skills remain too dangerous and unsafe. In this paper, we introduce the novel problem of Safety-Aware Skill Discovery, which aims to learn, in a task-agnostic fashion, a repertoire of reusable skills that are inherently safe to be composed for solving downstream tasks. We present a computationally tractable algorithm that learns a latent-conditioned skill policy that maximizes intrinsic rewards regularized with a safety-critic that can model any user-defined safety constraints. Using the pretrained safe skill repertoire, hierarchical reinforcement learning can solve multiple downstream tasks without the need for explicit consideration of safety during training and testing. We evaluate our algorithm on a collection of force-controlled robotic manipulation tasks in simulation and show promising downstream task performance while satisfying safety constraints.**

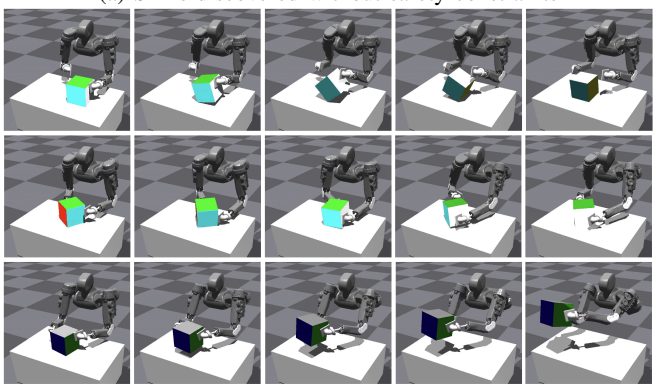
I. Introduction

Safety is a mandatory requirement in the task deployment of real-world robot manipulation systems. Recall that the central behaviors constituting robot manipulation tasks are about changing the state of the surrounding environment by explicitly engaging in a series of physical contact interactions. While a highly performant robot manipulator must be able to actively exploit and sequence diverse contact behaviors to solve the given task, physical contact can raise serious safety issues, e.g., irrecoverable damages to the robot or the surrounding environment. Moreover, various hardware constraints, including self-collisions and actuation limits, should be strictly satisfied, and other task-specific requirements, e.g., an object should not fall down from the table (see Figure 1a), can also relate to safety issues.

The primary objective of this paper is to develop an intrinsically safe and skilled robot manipulation system that can efficiently solve a collection of downstream tasks subject to any given set of safety constraints. We aim to achieve this goal by drawing upon the seemingly unrelated ideas from unsupervised skill discovery and safe reinforcement learning.



(a) Skills discovered without safety constraints



(b) Safety-aware unsupervised skill discovery

Fig. 1: Snapshots of force-controlled bimanual manipulation behaviors of AMBIDEX discovered from scratch. Red colored table indicates that the agent is in unsafe state. Violated safety constraints are listed in each frames. Each row represents a single skill.

A. Related Works

Skill discovery algorithms, also referred to as unsupervised reinforcement learning [1], aim at learning behaviors without relying on extrinsic task rewards. Based only on intrinsic motivations, skill discovery algorithms have shown to be able to learn sufficiently diverse and useful primitive behaviors which can also be leveraged to solve various downstream tasks using hierarchical reinforcement learning. One of the most widely used objective in skill discovery include mutual information between a latent skill variable and some state marginals so as to produce diverse and discriminative behaviors in the state space [2]–[6].

As skill discovery methods typically encourage exploratory and dynamic behaviors owing to the nature of intrinsic motivation, it can often be the case that a large portion of discovered skills turns out to be too dangerous, and hence cannot be reused in solving safety-critical downstream

[1]Equal contribution. Author names are in alphabetical order.

[1]NAVER LABS, Gyeonggi-do, 13561, South Korea. email: ty-lee@naverlabs.com

[2]NAVER LABS Europe, 6 chemin de Maupertuis, Meylan, 38240, France. email: julien.perez@naverlabs.com

tasks. To the best of our knowledge, there are few studies that formally investigate safety issues in the context of unsupervised skill discovery.

There are various approaches to designing a task-specific controller that addresses safety concerns. While designing simple heuristics can be sufficient in some cases, these ad-hoc treatments may not adequately address all risks or may restrict task performance unexpectedly. Constrained optimal control methods [7]–[10] offer various formal ways to synthesize safety-guaranteed controllers. These methods typically assume that the system dynamics are of a particular form and known, and safety is defined by deterministic constraints on the executed trajectory. On the other hand, safe reinforcement learning (RL) methods offer model-free approaches to ensuring probabilistic safety guarantees under unknown, stochastic dynamics [11], [12]. Some of these methods constrain conditional value at risk or probabilistic bounds of rewards and constraints in a stochastic environment [13]–[16], while others focus on constrained Markov Decision Process (CMDP) [17], where the expected sum of (constraint) costs is constrained while maximizing that of rewards. Among them, the Safety-critic-based methods [18]–[20], which we adopt in our work, aim to learn the critic function that estimates the probability of failure in the future events and use it to constrain the task-specific policy.

The concept of addressing safety in a task-agnostic manner is particularly attractive for multi-task applications, where the same set of unsafe behaviors may occur across multiple tasks. For example, control barrier function (CBF)-based approaches [21]–[23] can be applied to any type of task-specific controllers by designing a safety filter that projects the actions from the task-specific controllers. However, the validity of CBF-based safety filter design may depend on the structure of the system dynamics or become computationally intractable for complex, high-dimensional systems [23]. In [18], a safety-critic function is pretrained in some predefined set of tasks, and resued in learning new tasks at test time. In our work, we don't require the user to manually specify the task set to learn safety. Also, we maintain safety in the form of composable low-level skill policies, rather than the safety-critic function. While [24] also aims to learn safe exploratory policies whose safety is implied by matching a state marginal to a particular state distribution, the use of learned policies has not been explored for solving downstream tasks under the hierarchical reinforcement learning framework.

B. Contribution

The contribution of this paper is twofold. First, we define the novel problem of Safety-Aware Skill Discovery (SASD), which aims at learning, in a task-agnostic fashion, a repertoire of reusable skills that is inherently safe to be composed for solving downstream tasks. Second, we propose an algorithm for SASD that learns a latent-conditioned skill-policy maximizing mutual information based intrinsic reward, regularized with safety-critic that can model any user-defined safety constraints. Using the pretrained repertoire of safe skills, one can solve various downstream tasks without the need of additional safety considerations. We evaluate our algorithm on a collection of force-controlled robotic manipulation tasks in simulation and show promising downstream task performance while satisfying safety constraints.

The rest of the paper is organized as follows: Section II gives preliminary elements of safe RL and unsupervised skill discovery. Then, Section III introduces the problem of SASD and a concrete algorithm to solve the problem. Finally, Section IV details a series of comparative quantitative and qualitative results in the domain of force-controlled object manipulation.

II. Preliminaries

In this section, we introduce two key elements that underlie the problem of SASD. We first define the framework of safety-aware Markov Decision Process (MDP) and accompanied safety-critic concept proposed in [18]. Then, we briefly overview unsupervised skill discovery methods that maximize information theoretic objectives.

A. Safety-aware Markov Decision Process

We assume an environment with fully-observable state $s_t \in \mathcal{S}$, action $a_t \in \mathcal{A}$, state transition probability $p(s_{t+1}|s_t, a_t)$, and a scalar reward function $r_t = r(s_t, a_t, s_{t+1})$ which defines a Markov Decision Process (MDP) represented as a tuple $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, p, r \rangle$. As an incremental construction, a safety-aware MDP is defined as,

$$\mathcal{T} = \langle \mathcal{S}, \mathcal{A}, p, r, \mathcal{I} \rangle \tag{1}$$

where a safety-incident binary indicator $\mathcal{I}(s)$ indicates if a given state s is unsafe or not; $\mathcal{S}_{\text{unsafe}} = \{s \mid \mathcal{I}(s) = 1\}$ defines a set of unsafe states.

The goal in safety-aware MDP is to find an optimal stochastic policy that maximizes the expected cumulative reward with a probability of safety constraint violation bounded by ϵ:

$$\begin{aligned} \max_{\pi} \; J(\pi) = \; & \mathbb{E}_{p_\pi(\tau)} \left[\sum_{t=0}^{T} r(s_t, a_t, s_{t+1}) \right] \\ & \text{s.t. } \mathbb{E}_{p_\pi(s)} \left[\mathcal{I}(s) \right] < \epsilon, \end{aligned} \tag{2}$$

where $p_\pi(s)$ and $p_\pi(\tau)$ respectively denote the state marginal and state-action trajectory distribution induced by the policy π. The safety constraint above can be approximated by the safety-critic Q function [18] which is defined as

$$\begin{aligned} & Q^{\pi}_{\text{safe}}(s_t, a_t) = \mathcal{I}(s_t) + \\ & (1 - \mathcal{I}(s_t)) \mathbb{E}_{\substack{s_{t+1} \sim p(\cdot|s_t, a_t) \\ s_{t'} \sim \rho_\pi \text{ for } t' > t+1}} \left[\sum_{t'=t+1}^{T} \gamma_{\text{safe}}^{t'-t} \mathcal{I}(s_{t'}) \right], \end{aligned} \tag{3}$$

where γ_{safe} is a discounting factor. This cumulative discounted probability of failure satisfies the following Bellman equation:

$$Q^{\pi}_{\text{safe}}(s, a) = \mathcal{I}(s) + (1 - \mathcal{I}(s)) \mathbb{E}_{\substack{s' \sim p(\cdot|s,a) \\ a' \sim \pi(\cdot|s')}} \left[\gamma_{\text{safe}} Q^{\pi}_{\text{safe}}(s', a') \right].$$

As in standard Q-learning, it is parameterized by a neural network (with parameter ψ) as $Q^{\pi}_{\text{safe},\psi}$ and obtained by minimizing the following mean squared Bellman error:

$$J_{\text{safe}}(\psi) = \mathbb{E}_{(s,a,s',a')\sim p_\pi}\Big[\big(Q^{\pi}_{\text{safe},\psi}(s,a) - \big(\mathcal{I}(s) + (1-\mathcal{I}(s))\gamma_{\text{safe}}\bar{Q}^{\pi}_{\text{safe},\psi}(s',a')\big)\big)^2\Big], \quad (4)$$

where $\bar{Q}^{\pi}_{\text{safe},\psi}$ corresponds to the delayed target network.

B. Unsupervised Skill Discovery

Unsupervised skill discovery allows the agent to learn diverse behaviors without extrinsic rewards. In the context of mutual information maximizing skill discovery approaches, *skill* is represented as a latent-conditioned policy $\pi(a|s,z)$ where a latent variable $z \in \mathcal{Z}$ is normally drawn from a fixed prior distribution $p(z)$. Executing skill policy $\pi(a|s,z)$ on the initial state sampled from a fixed distribution $s_0 \sim p(s_0)$ induces a skill-conditional trajectory distribution $p_\pi(\tau|z)$.

Maximizing the mutual information between the latent variable z and some marginal over the state trajectory has proven to encourage exploration and produce diverse behaviors. To explain, let us consider the maximization of mutual information objective $\text{MI}(s,s';z)$ between the state transition (s,s') and skill z, which can be approximated with a variational lower bound [2], [3], [5] as follows:

$$\begin{aligned}\text{MI}(z;s,s') &= \mathcal{H}(z) - \mathcal{H}(z|s,s') \\ &= \mathbb{E}_{p(z)}\mathbb{E}_{p_\pi(s,s'|z)}\big[\log p(z|s,s') - \log p(z)\big] \\ &\geq \mathbb{E}_{p(z)}\mathbb{E}_{p_\pi(s,s'|z)}\big[\log q(z|s,s')\big] + (\text{const}),\end{aligned} \quad (5)$$

where $\mathcal{H}(\cdot)$ denotes the entropy and $q(z|s,s')$, so-called skill discriminator, represents the variational approximation for the true posterior $p(z|s,s')$. Parameterizing the skill discriminator with a neural network q_η, typical skill discovery algorithms proceed by alternating between updating the q_η so as to maximize the likelihood of on-policy skill samples, i.e.,

$$\max_\eta J_{\text{disc}}(\eta) = \mathbb{E}_{p(z)}\mathbb{E}_{p_\pi(s,s'|z)}[\log q_\eta(z|s,s')], \quad (6)$$

and optimizing the skill policy, parametrized by a neural network π_θ, using standard RL algorithm,

$$\max_\theta J_{\text{policy}}(\theta) = \mathbb{E}_{p(z)}\mathbb{E}_{p_{\pi_\theta}(\tau|z)}\Big[\sum_{t=0}^{T} r(s_t, s_{t+1}, z)\Big], \quad (7)$$

to maximize the intrinsic reward r_t given as

$$r(s,s',z) = \log q_\eta(z|s,s'). \quad (8)$$

III. Safe Skill Discovery

In this section, we introduce the problem of *safety-aware skill discovery (SASD)*, which aims to discover diverse task-agnostic skills that satisfies user-defined safety constraints. We also propose an algorithm to solve SASD, using safety-critic to model and regulate the probability of failure over the future state visits induced by the skill policy, while maximizing mutual information based intrinsic reward.

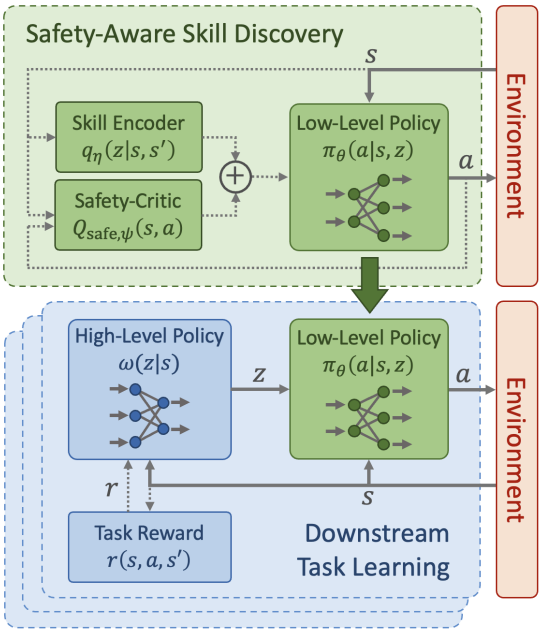


Fig. 2: An overview of our safe skill discovery framework that consists of two stages: pre-training safe skill policies and learning tasks based on the skills. In the first stage, the skill policy π is optimized to minimize the failure probability that is estimated by the safety-critic, while maximizing the skill discovery reward given by the skill encoder. The task policy is then optimized to maximize the task reward using the skill policy as a low-level controller. While the dotted lines denote the computation of the policy losses, the solid lines denote the actual control diagram of the policies.

A. Problem Definition

Our goal is to solve a set of downstream tasks $\{\mathcal{T}_i\}_{i=1}^N$, each of which can be represented as a safety-aware MDP with different task rewards r_i, but with shared safety constraint imposed by a binary indicator $\mathcal{I}$:

$$\mathcal{T}_i = \langle \mathcal{S}, \mathcal{A}, p, r_i, \mathcal{I} \rangle.$$

To remove the necessity of repeatedly considering the same safety constraints for every downstream task $\{\mathcal{T}_i\}_{i=1}^N$ training phase, we aim to pretrain a task-agnostic repertoire of low-level skills, i.e., latent conditioned skill policy $\pi(a|s,z)$, that inherently satisfies the safety constraints imposed by $\mathcal{I}$. By doing so, we can reuse such skill repertoire to solve multiple downstream tasks $\mathcal{T}_i$ in a hierarchical manner, i.e., training high-level task policies $\omega_i(z|s)$, without the need to additionally consider safety constraints during its training.

For this purpose, we require the behaviors generated by low-level latent-conditioned skill-policy π to satisfy the user-defined safety constraints even when a random sequence of skills $(z_0, z_1, \cdots, z_{T-1}) \sim p(\tau_z)$ are temporally composed, where $p(\tau_z)$ denotes the distribution of possible *skill composition schemes*.

Then, we introduce a problem of safety-aware skill discovery, which can be formulated as a constrained optimization

Algorithm 1 Safety-Aware Skill Discovery

$\mathcal{B} \leftarrow$ initialize on-policy buffer
$\pi_\theta \leftarrow$ initialize skill policy
$q_\eta \leftarrow$ initialize discovery reward models
$Q_{\text{safe},\psi} \leftarrow$ initialize safety critic
while not converged **do**
  Sample skill seq. $(z_0, \cdots, z_{T-1}) \sim p(\tau_z)$
  Sample initial state $s_0 \sim p(s_0)$
  Collect transitions $(z_t, s_t, a_t, s_{t+1}, \mathcal{I}(s_t))_{t=0:T}$
  Update buffer $\mathcal{B}$ with collected transitions τ
  Update skill discriminator q_η via SGD (6)
  Compute reward $r(s_t, s_{t+1}, z_t)$ for all transitions (8)
  Update safety-critic $Q_{\text{safe},\psi}$ via SGD (4)
  Update skill policy π_θ via PPO (10)
  Update λ via SGD (10)
end while

$$\max_{\pi} \ \mathrm{MI}(z; s, s') \quad \text{s.t.} \ \ \mathbb{E}_{p_\pi(s)}[\mathcal{I}(s)] < \epsilon \tag{9}$$

bounding the probability of safety constraint violation by ϵ while maximizing the mutual-information based skill discovery objective. ϵ determines the level of safety constraint; one can impose stronger safety constraints by reducing ϵ, i.e., allowable probability of failure.

B. Algorithm

In this section, our main problem (9) is specifically instantiated and solved by a) replacing the mutual information objective with a tractable variational lower bound (5) and b) modeling the probability of safety constraint violation with safety-critic Q_{safe} (3). With a Lagrangian formulation of the constraint, following surrogate objective can be used to update skill policy π_θ:

$$\max_{\theta} \min_{\lambda \geq 0} \ \mathbb{E}_{\substack{s,z\sim\mathcal{B} \\ a\sim\pi_\theta(\cdot|s,z)}} \left[Q^{\pi_\theta}_{\text{skill}}(s,a,z) - \lambda(Q^{\pi_\theta}_{\text{safe},\psi}(s,a) - \epsilon) \right] \tag{10}$$

where $Q^{\pi_\theta}_{\text{skill}}$ is the critic or advantage function for the intrinsic reward (8).

Algorithm 1 provides more detailed overview of the algorithm. We first sample random skill sequence $(z_0, z_1, \cdots, z_{T-1}) \sim p(\tau_z; n, T)$ where the skill composition schemes $p(\tau_z; n, T)$ are designed to repeat the latent z for multiple (n) timesteps before a new latent is sampled from the prior $p(z)$ and again repeated until it reaches the horizon T. We collect state transitions from the on-policy rollouts as well as the safety indicator $\mathcal{I}$. At each timestep t, policy π_θ computes an action conditioned on the current state s_t and skill z_t. Using the collected transitions, we update the skill discriminator network q_η according to (6). Then, the skill discovery reward $r_t = \log q_\eta(z_t|s_t, s_{t+1})$ is calculated using the updated q_η.

IV. Experiments

In this section, we aim to answer the following questions: (1) To what extent does SASD ensure safety while learning a diverse skill repertoire? (2) Is it safe to arbitrarily compose the discovered skills? (3) Can we successfully solve a set of contact-rich manipulation tasks with discovered skills?

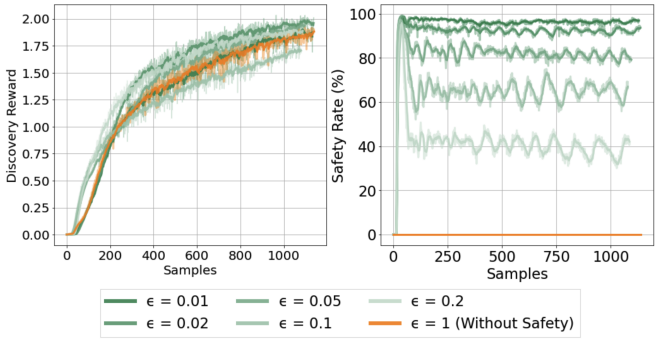


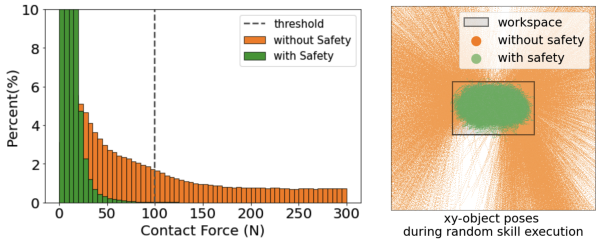
Fig. 3: Reward and safety rate during skill discovery phase.

All experiments are done in a simulated environment using Isaac Gym [25]. During training, 16,000 environments (Fig. 1) each equipped with a table, box, and 14-DoF dual-armed robot AMBIDEX [26] are simulated in parallel with a simulation frequency of 100Hz. During the training, $T_{\text{train}} = 300$ and $n_{\text{train}} = 100$ is used (i.e., composing three skills) to sample skill sequences from $p(\tau_z; n, T)$. We aim to discover diverse bi-manual manipulation skills such as pushing, grasping, flipping, rotating a box using both hands, while ensuring a set of predefined safety constraints. We define following states as *unsafe* in the following experiments: (1) joint position exceeding 95% of its physical limits, (2) joint velocity exceeding 10 rad/s, (3) excessive contact force of 100 N or more applied to the robot, (4) velocity of the robot hands exceeding 2 m/s, (5) the object moving outside of the robot's reachable workspace. If any of the above constraints is violated at least once during an episode, we call that episode unsafe and the *safety rate* simply denotes the ratio of unsafe episodes, i.e., estimate of $1 -$ (probability of failure).

For the skill discovery reward, we use the formulation proposed by Park *et al.* [4] which models the skill discriminator $q_\eta(z|s, s') \sim \mathcal{N}(\phi_\eta(s') - \phi_\eta(s), \mathbf{I})$ under 1-Lipschitz constrained state encoder ϕ_η to better focus on meaningful state differences: reward (8) can thus be calculated $r_t = \log q_\eta(z_t|s_t, s_{t+1}) = (\phi_\eta(s_{t+1}) - \phi_\eta(s_t))^\intercal z_t$. Due to its nature of objective, it encourages the agent to prefer skills with larger traveled distances, learning more diverse and dynamic skills.

A. Discovering Safe Manipulation Skills

In Figure 3, the returned discovery rewards and the safety rates during the skill discovery phase are compared for different values of the safety-critic threshold ϵ. As expected, the safety rate increases as ϵ decreases. Also, it shows that the respective safety rates converge fast at the early stage of training. We believe that this is due in large part to the vast amount of samples that can be collected from the Isaac Gym simulator at each training epoch and also to the growing diversity of exploratory experiences encouraged by the skill discovery objective. It can also be observed that the



(a) Contact force (b) Object position

Fig. 4: Visualization of safety-related statistics.

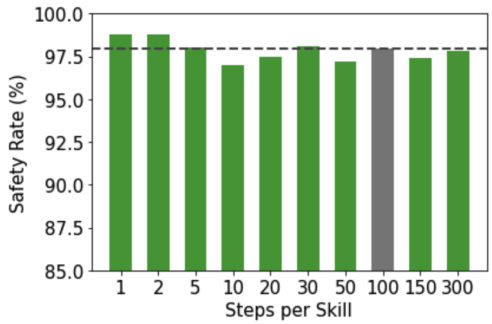


Fig. 5: Different skill composition settings and safety rates. Steps per skill indicates the number of repeated execution of skill z

converged value of skill discovery reward had no significant correlation with that of safety rate. However, we believe this result would not generalize to different experimental conditions, e.g., types of discovery objective, safety constraint and system dynamics. That said, we would like to leave a more comprehensive evaluation of the proposed SASD framework as a future work.

B. Safety Evaluation of The Discovered Skills

We qualitatively analyze how skills discovered with SASD satisfies individual safety constraints. To do so, we analyze the safety constraints while executing random skill sequences τ_z sampled from the same $p_{\text{train}}(\tau_z)$ used during training. As shown in Figure 4, skills discovered without safety constraints (a) constantly pushes the object outside the robot's reachable workspace and (b) applies excessive forces exceeding the predefined threshold. On the other hand, when performing manipulation with skills discovered with SASD, the object is gracefully manipulated without any excessive forces applied to the robot, while at the same time remaining within the reachable workspace.

We also analyze the effect of different skill composition schemes on safety by changing n and T of $p(\tau_z; n, T)$ from the ones used during training. Since the different composition schemes might lead to different state trajectories, it is unsure whether the safety rate during the training phase is also secured for all different n and T.

As the effect of n, Figure 5 shows the safety rates for skill sequences sampled from $p(\tau_z; n, T_{\text{train}})$ where $n \neq n_{\text{train}}$. Almost constant safety rates are observed even with different

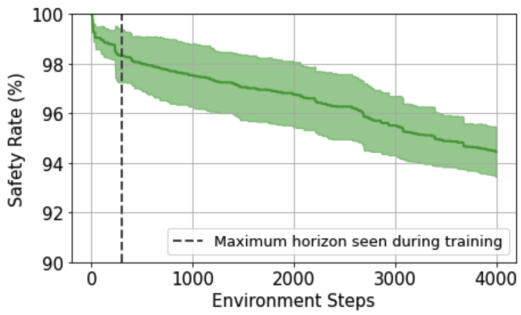
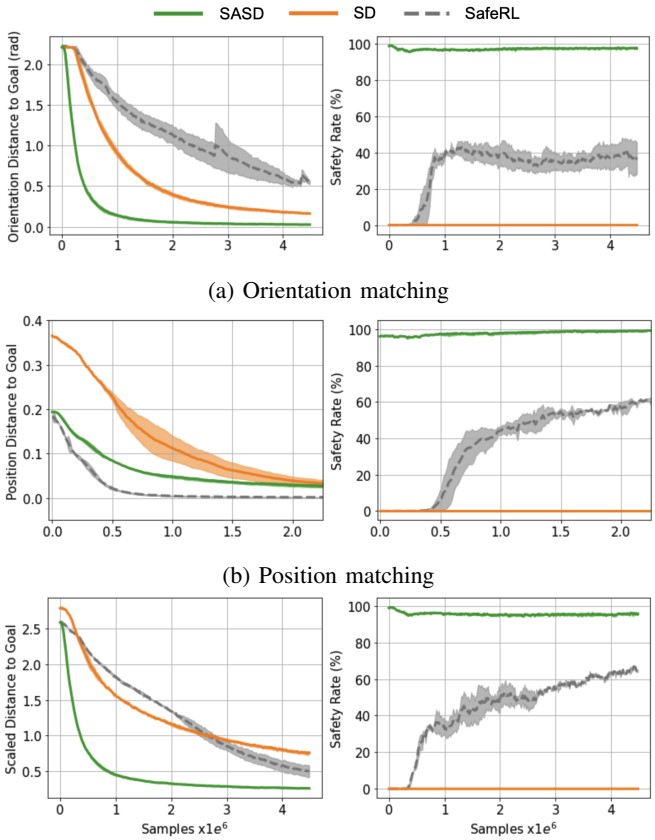


Fig. 6: Safety rates at time steps $t > T_{\text{train}}$ exceeding the horizon seen during skill discovery phase.



(a) Orientation matching

(b) Position matching

(c) Position and orientation matching

Fig. 7: Comparison of downstream task performances.

steps per skill n. Then, as the effect of T, Figure 6 shows the safety rates for extended periods, i.e., $\tau_z \sim p(\tau_z; n_{\text{train}}, T)$ where $T > T_{\text{train}}$. Although the safety rate gradually drops as the horizon increases, we note that it still maintains a high safety rate over 90 percent.

These are in fact useful and necessary properties as the high-level policy $\omega(z|s, g)$ might require higher or lower skill resampling frequency and longer horizon than those used during training.

C. Solving Contact-Rich Downstream Tasks

In this section, we show that our skills discovered by SASD are not only safe, but also diverse and useful enough

to effectively solve various contact-rich downstream manipulation tasks. We consider three downstream tasks: a) *orientation-matching*: reorienting the object to target orientations, b) *position-matching*: moving the object to target positions, and c) *position-and-orientation-matching*: moving the object to target position and orientation at the same time. For each task, we train a high-level task policy $\omega(z|s,g)$ with PPO [27] using negative distance to the target state as a reward.

We compare three methods: solving the tasks via hierarchical RL using the skills discovered a) with safety constraints (**SASD**) and b) without safety constraints (**SD**), and c) solving the tasks from scratch by joint learning the task policy and safety-critic (2) (**SafeRL**). In Figure 7, the learning curves of each method are compared for the tasks. The shading indicates the standard deviation of different seeds, which is taken over the 16,000 parallel episodes (i.e., the variance of the mean). As mentioned earlier, comparing various methods of the safe RL and the skill discovery is out of scope as this paper mainly aims to study the synergies between them.

Upon initial observation, it becomes apparent that SD exhibits faster learning compared to SafeRL, despite the fact that the safety rates remain almost constantly zero across all tasks. On the other hand, SASD outperforms both methods in terms of both performance and safety, and, remarkably, it learns even faster than SD. At first glance, this result may appear counter-intuitive; however, we hypothesize that this is mainly due to the smaller action spaces resulting from the imposition of safety constraints. Such constraints effectively reduce the search space of the optimization process, leading to more efficient learning. This phenomenon has been reported in previous studies [18], highlighting the beneficial effect of safety considerations on learning outcomes.

Although our safety constraint formulation does not require any information regarding the downstream tasks, we have observed that incorporating additional constraints that are applicable to all downstream tasks can enhance learning efficiency (at the cost of not being called "task-agnostic"). To this end, the constraints on the contact force, end-effector speed, joint position and velocity are universal and basic for any robotic tasks, while the restrictions on the object position represent an example of utilizing task-specific information (which may not be relevant for other tasks such as object throwing).

Secondly, we note that the curves' variance is considerably lower for SD and SASD compared to SafeRL. Although it would be unfair to draw direct comparisons between SafeRL and the pre-trained methods, it is noteworthy that SASD exhibits the most stable and efficient learning for multi-task applications, and is highly reproducible across different seeds.

D. Implementation Details

Object pose, represented as 3D position and SO(3) rotation matrix, is concatenated into a 12-dimensional vector as an input to the skill discriminator q_η. Object velocity and robot joint states are additionally concatenated as an input to the skill policy π_θ, safety-critic $Q_{\text{safe},\psi}$, and downstream task policy ω_i. For PPO training, we use clipped objective of PPO with target KL 0.05 and clip ratio 0.2. In addition, to mitigate the issue of sampling out-of-distribution skills during downstream task planning [28], we model the latent space as a hypersphere $\mathcal{Z} = \{z : \|z\| = 1\}$ where $p(z)$ is a uniform distribution on the surface of the sphere. We sample from $p(z)$ by normalizing $\tilde{z}$ sampled from a standard Gaussian distribution,

$$\tilde{z} \sim \mathcal{N}(0, \mathbf{I}),\ z = \tilde{z}/\|\tilde{z}\| \tag{11}$$

Regarding the network size, both the skill policy and skill discriminator comprise four hidden layers, each with a layer size of 256. Meanwhile, the value function and safety-critic consist of four hidden layers, each with a layer size of 512. Notably, the *tanh* function is employed as the output activation for the safety-critic. The skill discovery was performed within 12 hours, utilizing a single A100 GPU.

V. Conclusion

In this paper, we introduce a novel skill discovery formulation called safety-aware unsupervised skill discovery (SASD) that aims to learn intrinsically safe and diverse skills with minimal extrinsic supervision. We achieve this by combining the mutual information-based skill discovery and the safety-critic method. Through extensive simulated experiments with our force-controlled dual arm robot AMBIDEX, we show that our method strongly ensures safety without sacrificing much on the performance of unsupervised skill dicovery. Moreover, our pretrained skill models exhibit strong generalization performance across a variety of initial states, skill composition schemes, and extended execution horizons. Perhaps most significantly, when we apply our method to hierarchical reinforcement learning for downstream task training, it not only guarantees high levels of safety throughout the entire training process, but also leads to faster convergence. As a result, we believe that sim-to-real transfer of skills learned using SASD represents a promising direction for future research, as we envision that it could ultimately free the user from accounting for safety to train robot behaviors in the real-world.

References

[1] M. Laskin, D. Yarats, H. Liu, K. Lee, A. Zhan, K. Lu, C. Cang, L. Pinto, and P. Abbeel, "Urlb: Unsupervised reinforcement learning benchmark," *arXiv preprint arXiv:2110.15191*, 2021.

[2] B. Eysenbach, A. Gupta, J. Ibarz, and S. Levine, "Diversity is all you need: Learning skills without a reward function," in *International Conference on Learning Representations*, 2018.

[3] A. Sharma, S. Gu, S. Levine, V. Kumar, and K. Hausman, "Dynamics-aware unsupervised discovery of skills," in *International Conference on Learning Representations*, 2019.

[4] S. Park, J. Choi, J. Kim, H. Lee, and G. Kim, "Lipschitz-constrained unsupervised skill discovery," in *International Conference on Learning Representations*, 2021.

[5] J. Choi, A. Sharma, H. Lee, S. Levine, and S. S. Gu, "Variational empowerment as representation learning for goal-conditioned reinforcement learning," in *International Conference on Machine Learning*. PMLR, 2021, pp. 1953–1963.

[6] D. Cho, J. Kim, and H. J. Kim, "Unsupervised reinforcement learning for transferable manipulation skill discovery," *IEEE Robotics and Automation Letters*, 2022.

[7] N. Ratliff, M. Zucker, J. A. Bagnell, and S. Srinivasa, "Chomp: Gradient optimization techniques for efficient motion planning," in *2009 IEEE International Conference on Robotics and Automation*. IEEE, 2009, pp. 489–494.

[8] J. F. Fisac, N. F. Lugovoy, V. Rubies-Royo, S. Ghosh, and C. J. Tomlin, "Bridging hamilton-jacobi safety analysis and reinforcement learning," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 8550–8556.

[9] J. Li, D. Lee, S. Sojoudi, and C. J. Tomlin, "Infinite-horizon reach-avoid zero-sum games via deep reinforcement learning," *arXiv preprint arXiv:2203.10142*, 2022.

[10] S. Bansal and C. J. Tomlin, "Deepreach: A deep learning approach to high-dimensional reachability," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 1817–1824.

[11] L. Brunke, M. Greeff, A. W. Hall, Z. Yuan, S. Zhou, J. Panerati, and A. P. Schoellig, "Safe learning in robotics: From learning-based control to safe reinforcement learning," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 5, pp. 411–444, 2022.

[12] J. Garcıa and F. Fernández, "A comprehensive survey on safe reinforcement learning," *Journal of Machine Learning Research*, vol. 16, no. 1, pp. 1437–1480, 2015.

[13] A. Tamar, Y. Glassner, and S. Mannor, "Policy gradients beyond expectations: Conditional value-at-risk," *arXiv preprint arXiv:1404.3862*, 2014.

[14] J. Duan, Y. Guan, S. E. Li, Y. Ren, Q. Sun, and B. Cheng, "Distributional soft actor-critic: Off-policy reinforcement learning for addressing value estimation errors," *IEEE transactions on neural networks and learning systems*, 2021.

[15] X. Ma, L. Xia, Z. Zhou, J. Yang, and Q. Zhao, "Dsac: distributional soft actor critic for risk-sensitive reinforcement learning," *arXiv preprint arXiv:2004.14547*, 2020.

[16] J. Choi, C. Dance, J.-e. Kim, S. Hwang, and K.-s. Park, "Risk-conditioned distributional soft actor-critic for risk-sensitive navigation," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 8337–8344.

[17] J. Achiam, D. Held, A. Tamar, and P. Abbeel, "Constrained policy optimization," in *International conference on machine learning*. PMLR, 2017, pp. 22–31.

[18] K. Srinivasan, B. Eysenbach, S. Ha, J. Tan, and C. Finn, "Learning to be safe: Deep rl with a safety critic," *arXiv preprint arXiv:2010.14603*, 2020.

[19] H. Bharadhwaj, A. Kumar, N. Rhinehart, S. Levine, F. Shkurti, and A. Garg, "Conservative safety critics for exploration," *arXiv preprint arXiv:2010.14497*, 2020.

[20] H. Yu, W. Xu, and H. Zhang, "Towards safe reinforcement learning with a safety editor policy," *arXiv preprint arXiv:2201.12427*, 2022.

[21] A. D. Ames, X. Xu, J. W. Grizzle, and P. Tabuada, "Control barrier function based quadratic programs for safety critical systems," *IEEE Transactions on Automatic Control*, vol. 62, no. 8, pp. 3861–3876, 2017.

[22] R. Cheng, G. Orosz, R. M. Murray, and J. W. Burdick, "End-to-end safe reinforcement learning through barrier functions for safety-critical continuous control tasks," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 3387–3395.

[23] J. J. Choi, D. Lee, K. Sreenath, C. J. Tomlin, and S. L. Herbert, "Robust control barrier–value functions for safety-critical control," in *2021 60th IEEE Conference on Decision and Control (CDC)*. IEEE, 2021, pp. 6814–6821.

[24] L. Lee, B. Eysenbach, E. Parisotto, E. Xing, S. Levine, and R. Salakhutdinov, "Efficient exploration via state marginal matching," *arXiv preprint arXiv:1906.05274*, 2019.

[25] V. Makoviychuk, L. Wawrzyniak, Y. Guo, M. Lu, K. Storey, M. Macklin, D. Hoeller, N. Rudin, A. Allshire, A. Handa *et al.*, "Isaac gym: High performance gpu-based physics simulation for robot learning," *arXiv preprint arXiv:2108.10470*, 2021.

[26] Y.-J. Kim, "Anthropomorphic low-inertia high-stiffness manipulator for high-speed safe interaction," *IEEE Transactions on robotics*, vol. 33, no. 6, pp. 1358–1374, 2017.

[27] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[28] X. B. Peng, Y. Guo, L. Halper, S. Levine, and S. Fidler, "Ase: Large-scale reusable adversarial skill embeddings for physically simulated characters," *ACM Trans. Graph.*, vol. 41, no. 4, Jul. 2022.