Safety-Guaranteed Skill Discovery for Robot Manipulation Tasks

Sunin Kim*
NAVER LABS
Gyeonggi-do, 13561, South Korea
`sunin.kim@naverlabs.com`

Jaewoon Kwon*
NAVER LABS
Gyeonggi-do, 13561, South Korea
`jaewoon.kwon@naverlabs.com`

Taeyoon Lee*
NAVER LABS
Gyeonggi-do, 13561, South Korea
`ty-lee@naverlabs.com`

Younghyo Park*
NAVER LABS
Gyeonggi-do, 13561, South Korea
`younghyo.park@naverlabs.com`

Julien Perez
NAVER LABS Europe
6 chemin de Maupertuis, Meylan, 38240, France
`julien.perez@naverlabs.com`

Abstract

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 remains too dangerous and unsafe. In this paper, we introduce the novel problem of safe skill discovery, 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. We propose *Safety-Guaranteed Skill Discovery* (SGSD), an algorithm that learns a latent-conditioned skill-policy, regularized with a safety-critic modeling a user-defined safety definition. Using the pretrained safe skill repertoire, hierarchical reinforcement learning can solve downstream tasks without the need of 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 with safety guarantees. Please find `https://sites.google.com/view/safe-skill` for supplementary videos.

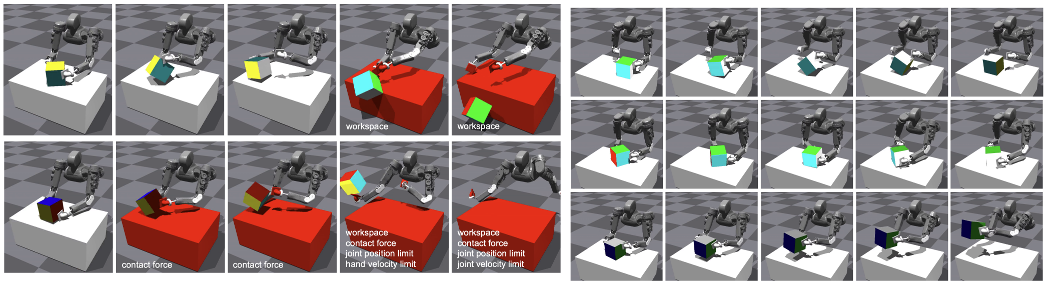
1 Introduction

Safety remains 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

---

*equal contribution

NeurIPS 2022 Workshop on Robot Learning: Trustworthy Robotics, Virtual, Virtual



(a) Skills discovered without safety constraints

(b) Safety-Guaranteed Skill Discovery

Figure 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.

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 a 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.

The contribution of this paper is twofold. First, we define the novel problem of safe skill discovery (SSD), 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 introduce SGSD, *Safety-Guarantee Skill Discovery*, an algorithm that learns a latent-conditioned skill-policy, regularized with safety-critic modeling of any user-defined safety definition. Using the pretrained safe skill repertoire, hierarchical reinforcement learning can solve downstream tasks without 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 with safety guarantees.

2 Safe Skill Discovery

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 [1], $\mathcal{T}_i = \langle \mathcal{S}, \mathcal{A}, p, r_i, \mathcal{I} \rangle$, with different task reward r_i; please refer to Appendix B for a formal definition of safety-aware MDP and related notions. The low-level skill-policy $\pi_\theta(a|s,z)$ pretrained in the skill discovery phase is reused to solve all the downstream tasks $\{\mathcal{T}_i\}_{i=1}^N$ by training task-specific high-level policies $\omega_i(z|s)$.

For this purpose, we require that the low-level skill-policy should not only be trained in such a way to generate diverse behaviors but also strictly guarantee safety criteria imposed by the binary indicator $\mathcal{I}$ even when a random sequence of skills $(z_1, z_2, \cdots, z_T),\ z_t \sim p(z|t)$ are temporally sequenced. Here $p(z|t)$ is determined by the resampling rate of skills from the fixed prior distribution $p(z)$.

The ensuing constrained optimization problem for our safety-guaranteed skill discovery approach is formulated as follows:

$$\max_{\pi_\theta}\ \mathrm{MI}(z; s, s') \quad \text{s.t.} \ \ \mathbb{E}_{a\sim\pi_{\text{comp}}}[Q_{\text{safe}}^{\pi_{\text{comp}}}(s,a)] < \epsilon_{\text{safe}}. \tag{1}$$

Replacing the mutual information objective with a tractable variational lower bound

$$\begin{aligned}\mathrm{MI}(z; s, s') = \mathcal{H}(z) - \mathcal{H}(z|s,s') &= \mathbb{E}_{z\sim p(z),(s,s')\sim p_{\pi_z}}\big[\log p(z|s,s') - \log p(z)\big] \\ &\geq \mathbb{E}_{z_t,s,s'}\big[\log q_\eta(z|s,s')\big] + (\text{const}),\end{aligned}$$

and adopting a Lagrangian formulation for the safety-critic bound constraint, the surrogate objective used for skill policy update is given as follows:

$$\max_{\pi_\theta} \min_{\lambda\geq 0} \mathbb{E}_{\substack{z\sim p(z) \\ s,a,s'\sim p_{\pi_z}}} [\log q_\eta(z|s,s') - \lambda(Q_{\text{safe}}(s,a) - \epsilon_{\text{safe}})] \tag{2}$$

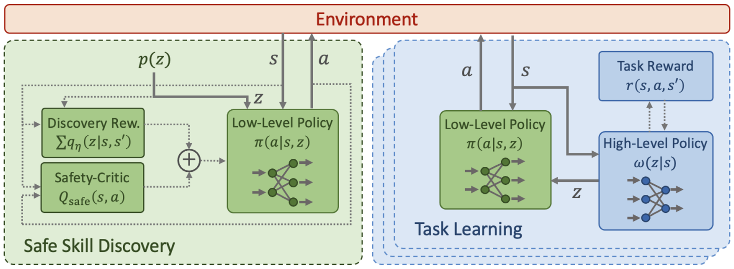


Figure 2: An overview of our safe skill discovery framework that consists of two stages: pre-training safe skill policies and learning tasks based on the skills. In the first stage, the skill policy π is optimized to minimize risks estimated by the safety-critic, while maximizing the skill discovery 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.

Here is how our Safety-Guaranteed Skill Discovery (SGSD) algorithm proceeds. We first sample random skill sequence $(z_1, z_2, \cdots, z_T)$, $z_t \sim p(z|t)$, and collect state transitions from the on-policy rollouts as well as the safety indicator $\mathcal{I}$. At each timestep t, policy π computes an action conditioned by the current state s_t and skill z_t. Using the collected transitions, we then update the skill discriminator network via $\max_\eta J_{\text{skill}}(\eta) = \mathbb{E}_{z,s,s'}[\log q_\eta(z|s,s')]$.

Then, the skill discovery reward $r_t = \log q_\eta(z_t|s_t, s_{t+1})$ is calculated using the updated q_η. To update the skill policy considering safety constraints, we first update safety-critic $Q_{\text{safe}}(s,a)$ to minimize the Bellman error [1–3]:

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

Then, the skill policy π_θ is updated by maximizing the surrogate objective (2) using any standard RL algorithm. We used clipped objective of proximal policy optimization(PPO) [4] with target KL 0.05 and clip ratio 0.2. In addition to the advantage function estimates, we also use a clipped version of the safety critic $Q_{\text{safe}}(s,a)$ to prevent over-exploitation of the current belief about safety. We then optimize the Lagrangian multiplier λ. The overall algorithm is summarized in Algorithm 1 and depicted in Figure 2.

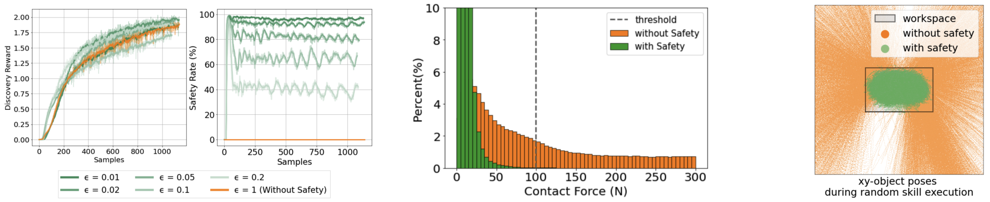
3 Experiments

We aim to answer the following questions: (1) To what extent does SGSD ensure safety while learning a skill repertoire? (2) Can any sequential composition of our discovered safe skill repertoire, i.e., random exploration on latent space $\mathcal{Z}$, ensure safety? (3) Can we successfully solve a set of contact-rich downstream manipulation tasks while ensuring safety, leveraging our discovered safe skill repertoire?

All environments are simulated using Isaac Gym[5]. During training, 16,000 environments (Fig. 1) each equipped with a table, box, and 14-DoF dual-armed robot AMBIDEX [6] are simulated in parallel with a simulation frequency of 100Hz. We aim to discover diverse bi-manual manipulation skills such as pushing, grasping, flipping, rotating a box using both hands, while ensuring safety defined with a set of predefined 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 one of the constraints is violated at least once during an episode, the episode is defined as unsafe. Safety rate is defined as the proportion of safe episodes among all episodes.

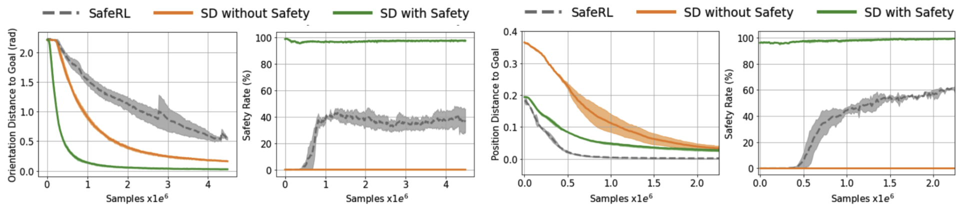
3.1 Safe Manipulation Skill Discovery

We choose to maximize the Lipschitz-constrained Skill Discovery (LSD) objective proposed by Park *et al.* [7] for safe manipulation skill discovery. LSD objective encourages the agent to prefer dynamic



(a) Reward and safety rate (b) Contact force (c) Object position

Figure 3: Reward, safety rate, and safety violations during skill discovery phase.



(a) Orientation matching (b) Position matching

Figure 4: Comparison of downstream task performances. Position-and-orientation matching task is given in Figure 6 in the Appendix. Safety rate indicates the ratio of safe episodes out of every 8192 number of randomly reset episodes.

skills with large traveled distances. In practice, we observe that this can lead to the violation of safety constraints. Figure 3a shows the discovery rewards and safety rates during skill discovery phase. It can clearly be observed that almost every behaviors generated by the agent trained without safety constraints are unsafe. On the other hand, skills discovered with our method SGSD show high safety rate. Surprisingly, even with tight safety regulations with small ϵ_{safe}, the agent receives similar scale of discovery rewards compared to that with no safety constraints while maintaining high safety rate.

In addition, we qualitatively analyze how skills discovered with SGSD satisfies individual safety constraints by executing random skills z sampled from the prior. The skills discovered without safety constraints constantly pushes the object outside the robot's reachable workspace (Figure 3c) and applies excessive forces to the environment (Figure 3b). On the other hand, with skills discovered by SGSD, the object is gracefully manipulated without any excessive forces applied to the robot, while at the same time remaining within the reachable workspace[2]. From these results, we could validate that diverse and useful safe skills can be successfully learned entirely from scratch using SGSD.

3.2 Solving Contact-Rich Downstream Tasks

In this section, we show that skills discovered by SGSD can be temporarily composed to solve various contact-rich downstream manipulation tasks while satisfying the safety constraints. Here we highlight again that we do not require training of any safety-related information during downstream task training phase. We consider three downstream tasks: a) *orientation matching*: reorienting the object to various target orientations, b) *position matching*: moving the object to various target positions, and c) *position-and-orientation matching*: moving and reorienting the object to target position and orientation at the same time. For each task, we train a high-level task policy $\omega(z|s, g)$ using per-step difference in negative distances to the target state used as a reward function. We compare our framework with two baselines: 'skill discovery (SD) without safety' that follows the original formulation of LSD without considering any safety constraints and a safe RL method denoted by 'SafeRL' that jointly learns the task policy and the safety-critic; the policy is constrained so that the safety-critic values of the policy output is kept under 0.01. Our method is denoted by 'SD with Safety'. As shown in Figure 4, it can be seen that, skills discovered with safety can effectively solve downstream tasks: it not only is faster at solving the task compared to SafeRL, but also zero-shot attains high safety rate without further finetuning of the safety-critic or the low-level skill policy.

---

[2] See also Appendix D for evaluation of safety over the extended length of random skill composition.

References

[1] 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.

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

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

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

[5] 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.

[6] 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.

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

[8] 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.

[9] 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.

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

[11] 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.

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

[13] 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.

[14] M. Posa, C. Cantu, and R. Tedrake, "A direct method for trajectory optimization of rigid bodies through contact," *The International Journal of Robotics Research*, vol. 33, no. 1, pp. 69–81, 2014.

[15] 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.

[16] A. R. Ansari and T. D. Murphey, "Sequential action control: Closed-form optimal control for nonlinear and nonsmooth systems," *IEEE Transactions on Robotics*, vol. 32, no. 5, pp. 1196–1214, 2016.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] N. Fazeli, R. Kolbert, R. Tedrake, and A. Rodriguez, "Parameter and contact force estimation of planar rigid-bodies undergoing frictional contact," *The International Journal of Robotics Research*, vol. 36, no. 13-14, pp. 1437–1454, 2017.

[22] 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.

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

[24] 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.

[25] 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.

[26] 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.

[27] B. Eysenbach, S. Gu, J. Ibarz, and S. Levine, "Leave no trace: Learning to reset for safe and autonomous reinforcement learning," *arXiv preprint arXiv:1711.06782*, 2017.

[28] B. Thananjeyan, A. Balakrishna, S. Nair, M. Luo, K. Srinivasan, M. Hwang, J. E. Gonzalez, J. Ibarz, C. Finn, and K. Goldberg, "Recovery rl: Safe reinforcement learning with learned recovery zones," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 4915–4922, 2021.

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

[30] 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.

A Related work

Skill discovery algorithms, also referred to as unsupervised reinforcement learning [8], 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. 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 [9, 10, 7, 11, 12].

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 tasks. To the best of our knowledge, there are few studies that formally investigate safety issues in the context of unsupervised skill discovery [8, 13].

Safety can be addressed in various ways depending on the required level of task performance. In some cases, it could often be sufficient to manually design, e.g., event-based safety control strategies at known risky states. However, these ad-hoc safety treatments can essentially restrict task performances or even fail to address some risks adequately; a gripper that avoids collision cannot grasp anything.

Given that an accurate model of the dynamics and the safety constraints are known, constrained optimal control methods can formally address safety [14–16]. Furthermore, reachability analysis offers a more general concept and control synthesis methods for various safety-critical systems [17–20]. However, model-based control methods pose another challenge in accurately modeling and estimating the contact dynamics and the contact states in practice [21].

Safe reinforcement learning (RL) offers model-free methods for ensuring different concepts of safety [22, 13]. While some of them constrain conditional value at risk or probabilistic bounds of rewards and constraints in a stochastic environment [23–26], others have studied reversibility and learning to reset [27, 28]. Most of safe RL methods however assume constrained Markov Decision Process (CMDP) [29] in which the expected sum of (constraint) cost is minimized while maximizing that of reward. Among them, the Safety-critic-based methods are the most relevant to our works [1–3]. They directly learn to estimate the probability of failure in order to guide their robots away from the actions that are likely to fail. Nevertheless, most of the safe RL methods are dedicated to a given set of tasks (i.e., safety value conditioned on their task policies) and often tend to generalize poorly for different tasks.

B 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,

$$\mathcal{T} = \langle \mathcal{S}, \mathcal{A}, p, r, \mathcal{I} \rangle \tag{3}$$

where a safety-incident 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_{safe}:

$$\begin{aligned} \max_{\pi} \; J(\pi) = \; & \mathbb{E}_{p_\pi(\tau)} \left[\sum_{t=0}^{T} r(s_t, a_t, s_{t+1}) \right] \\ & \text{s.t. } \mathbb{E}_{p_\pi(s)} \left[\mathcal{I}(s) \right] < \epsilon_{\text{safe}}, \end{aligned} \tag{4}$$

where $p_\pi(\tau) = p(s_0) \prod_{t=0}^{T-1} \pi(a_t|s_t) p(s_{t+1}|s_t, a_t)$ denotes the distribution over state-action trajectories and $p_\pi(s)$ denotes the state marginal distribution induced by the policy π.

For a given policy π, the safety-critic $Q^{\pi}_{\text{safe}}(s_t, a_t)$ is defined by the discounted cumulative failure probability in the future if starting at state s_t and it takes the action a_t:

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

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

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

C Implementation Details

For the above experiments, 3D position and SO(3) rotation matrix of the object is concatenated into a 12-dimensional vector as an input to the LSD for skill discovery. We reset every environment after executing 3 different skills, where single skill consists of 100 environment steps. We have empirically found that sequentially executing multiple skills before manual reset enables reliable sequential composition of random skills. In addition, to mitigate the issue of sampling out-of-distribution skills during downstream task planning [30], we use uniform distribution on the unit hypersphere as our latent prior $p(z)$, which has a compact support. In the case of the safety-critic, the position and velocity of the object and the position and velocity of the robot joint are used as inputs to require that

Algorithm 1 Safety-Guaranteed Skill Discovery

```
$\mathcal{B} \leftarrow$ initialize on-policy buffer
$\pi_\theta \leftarrow$ initialize skill policy
$q_\psi \leftarrow$ initialize discovery reward models
$Q_{\text{safe}} \leftarrow$ initialize safety critic
while not converged do
    Sample skill sequence $(z_1, \cdots, z_T)$, $z_t \sim p(z|t)$
    Sample initial state $s_0 \sim p(s_0)$
    Collect transitions $\tau = (z_t, s_t, a_t, s_{t+1}, \mathcal{I}(s_t))_{t=0:T}$
    Update buffer $\mathcal{B}$
    for $i = 1 : N_{skill}$ do
        Update skill discriminator network $q_\eta$
    end for
    Compute reward $r(s_t, s_{t+1}, z_t)$
    for $i = 1 : N_{policy}$ do
        Update safety-critic $Q_{\text{safe}}$
        Update skill policy $\pi_\theta$
        Update Lagrangian multiplier $\lambda$
    end for
end while
```

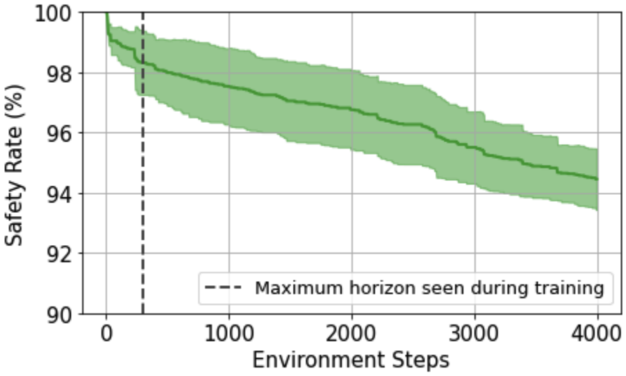
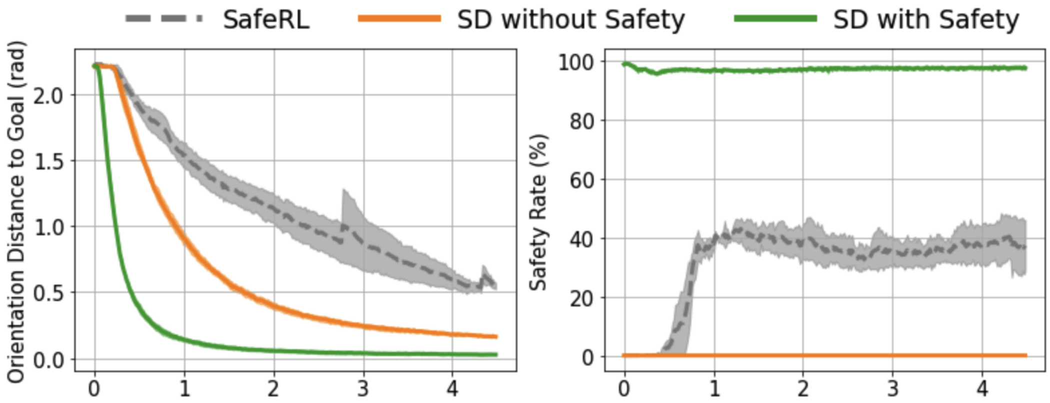


Figure 5: Safety rates at time steps exceeding the horizon seen during skill discovery phase.

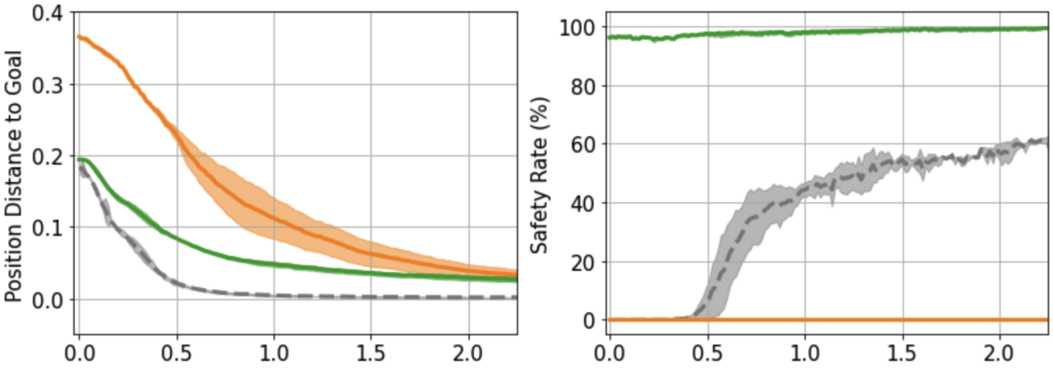
the states are fully observable for evaluating safety. We also experimentally found that the Q-value was better estimated with tanh than sigmoid as the output activation of the safety critic. As for the size of each network, skill policy and skill discriminator have 4 hidden layers with a layer size of 256, and value function and safety critic have 4 hidden layers with a layer size of 512. Skill discovery was done in 12hrs using a single A100 GPU. Algorithm 1 shows the pseudo code of our SGSD.

D Generalization over Extended Horizon

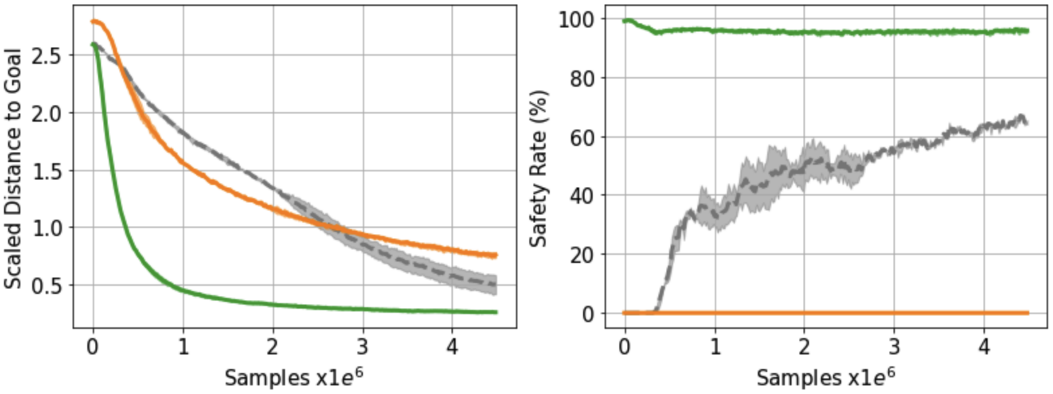
Figure 5 shows the safety rate measured while sequentially executing random skills discovered with SGSD for an extended length of time horizon compared to that used during the skill discovery training phase. Although the safety rate gradually drops as time progresses, we note that it still maintains a high safety rate over 90 percent. The decrease in safety rate can be attributed to the agent visiting out-of-distribution state space on which the safety-critic nor the skill policy has been learned to explore safely.



(a) Orientation matching



(b) Position matching



(c) Position and orientation matching

Figure 6: Comparison of downstream task performances.