

A Supplementary Material

A.1 Learning Structure

We use DDPG + HER for SCAPE and all the baselines in this study. We use the same hyperparameters as OpenAI baselines [32], except for the batch size of 256 (originally 1024).

A.2 Policy Parametrization

In all environments, we use a stiffness variable k to command a desired stiffness in the grasping direction. For the Block environment, the stiffness is applied to the direction of the parallel gripper opening/closing. For the Chip environment, the stiffness is applied to the wrist joint’s pitch rotation, which is the degree of freedom responsible for maintaining the grasp of the chip. For the NuFingers environment, the stiffness is applied to the grasping direction, which coincides with the radial direction of the polar coordinates.

In addition to the stiffness parameter, k_{lim} provides the upper limit for all stiffness controllers in this paper. We have found that having an extra parameter that controls the upper limit of the stiffness helps the policy converge faster to the minimum stiffness. Furthermore, such upper limit can be meaningfully related to the physical passivity of the robot [7].

Therefore, in all environments, SCAPE outputs two additional dimensions of the action space compared to the position control policy, which account for k and k_{lim} .

A.3 Environments

Tables below contain detailed information regarding the environments used in the paper.

Table 1: Reference success rates for each environment.

	Block	Chip	NuFingers
SR_{ref}	0.65	0.85	0.65

Table 2: List of uncertainties included in the training and evaluation.

Measurement Noise	
Property	Adds noise to the measurement <i>Uniform</i> (-1 cm, 1 cm) (Block, Chip) <i>Uniform</i> (-0.02 rad, 0.02 rad) (NuFingers)
Application	3D position of the object (Block, Chip) Rotation of the object (NuFingers)
Occurrence	100%
Random Perturbation	
Property	Adds velocity to the object <i>Uniform</i> (-50 cm/s, 50 cm/s) (Block, Chip) <i>Uniform</i> (-0.5 rad/s, 0.5 rad/s) (NuFingers)
Application	x -dir (Block) x_1 -dir (Chip) θ -dir (NuFingers)
Occurrence	100%
Control Failure	
Property	Repeats the previous action
Application	Entire action
Occurrence	10%

A.3.1 Details for the Block Environment

The Block environment is a modified version of *FetchPickAndPlace* environment from Gym. The grippers are more compliant. Most importantly, the object is now considered broken if the interac-

tion force exceeds a certain threshold called fragility shown in Table 3. In addition to the original observation and action spaces, the new observation space now includes force and stiffness, and the action space includes the changes in stiffness and its limit. The task-related kinematic goal is for the object to reach the goal position, i.e., $\|\mathbf{x}_{o-g}\| < d$, and the force \mathbf{F} is measured from series elasticity of each gripper. The distance threshold is $d = 5cm$, $\alpha = 2e^{-3}$ normalizes the force, and $\beta = 0$. We use a sparse reward function for the task-related kinematic goal to avoid penalizing the agent from necessary exploration [31], and a dense reward function for the safety-related goal to minimize the interaction forces. Also, the target location is always in the air to examine only the grasping solutions and discourage the use of other means of moving the object.

Table 3: List of modifications to model parameters in the Block environment.

<i>Gripper Link</i>	Mass (<i>kg</i>)	Contact Dimension
Original	4	4
Modified	0.4	6
<i>Gripper Actuator</i>	Stiffness (<i>N/m</i>)	Control Range (<i>m</i>)
Original	30000	0.0 – 0.2
Modified	250	-1.0 – 1.0
<i>Gripper Joint</i>	Armature	Damping (<i>N s/m</i>)
Original	100	1000
Modified	1	20
<i>Object</i>	Fragility (<i>N</i>)	Contact Dimension
Original	N/A	4
Modified	300	6

A.3.2 Details for the Chip Environment

The robot in the Chip environment has a compliant wrist. The observation space includes the positions of the object, fingertip, and the goal in Cartesian space. The fingertip velocity is also included. The action space consists of planar movement of the arm, the pitch movement at the wrist, and the changes in the wrist stiffness and its limit. The estimated interaction force is the wrist torque τ , which is calculated from the series elasticity of the wrist actuator.

The task-related kinematic goal is for the chip to rest at the target location, with small velocity, i.e., $\|\mathbf{s}_{o-g}\| < d$, where \mathbf{s} contains the position as well as velocity. Without the velocity goal, the low fidelity of the friction in MuJoCo leads the agent to continuously move the object around the goal position without stopping. This phenomenon is likely due to the fact that the kinetic friction is usually smaller than the static friction. By adding the velocity goal, the agent is penalized from moving and thus able to successfully learn the task. \mathbf{F} is the wrist torque measured from series elasticity, $d = 5cm$, $\alpha = 2e^{-2}$, and $\beta = 0$.

Table 4: List of important model parameters in the Chip environment.

	Stiffness (<i>N/m</i>)	Control Range (<i>m</i>)
<i>Forearm Actuator</i> (x_1)	250	0.0 – 0.2
<i>Forearm Actuator</i> (x_2)	250	0.0 – 0.2
	Stiffness (<i>Nm/rad</i>)	Control Range (<i>rad</i>)
<i>Wrist Actuator</i>	50	-1.0 – 1.0
	Coefficients	Contact Dimension
<i>Friction</i> _{finger-object}	1	6
<i>Friction</i> _{object-wall}	1	3
	Fragility (<i>N</i>)	Mass (<i>kg</i>)
<i>Object</i>	200	0.1

A.3.3 Details for the NuFingers Environment

In the NuFingers environment, the object has an integrated force sensor that directly measures the ground-truth force as well as an orientation sensor using a potentiometer. Also, elastic bands are installed that ground the object to the equilibrium orientation, providing resistance to the rotation.

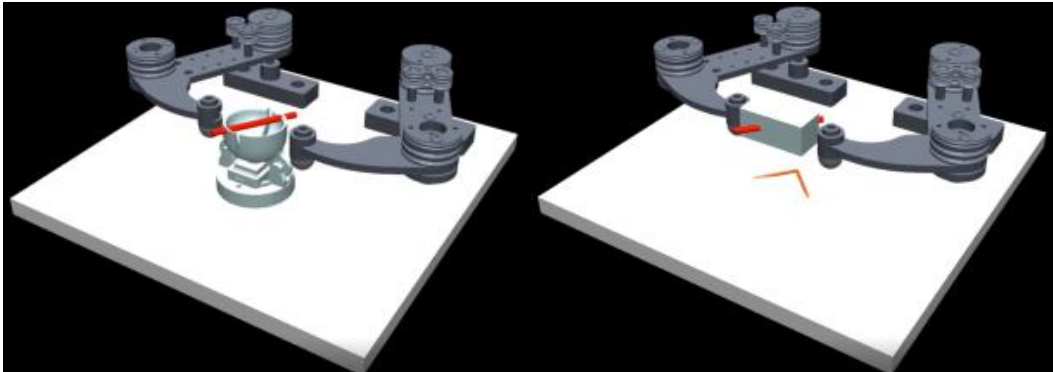
The task-related kinematic goal is to rotate the object to the desired orientation, i.e., $||\theta_{o-g}|| < d$, where $||\theta_{o-g}||$ is the difference between the goal and the current object orientations. The orientation error threshold is $d = \frac{\pi}{16}$. The vector \mathbf{F} contains the forces of each finger measured only in the grasping direction using series elasticity. The vector $\dot{\mathbf{q}}$ contains joint velocities. $\alpha = 4e^{-1}$, and $\beta = 1$ are normalization terms.

Furthermore, we apply domain randomization [2] during training to partially account for model discrepancies. We randomize the position and the width of the object, as well as the elasticity of the rubber bands of the object as shown in Table 5. Note that other important dynamic properties such as backlash or Coulomb friction are not modeled in simulation even though they have considerable effects on the performance of the actual system.

Also, due to the erratic behavior of contact between two concave surfaces, the shape of the object is assumed to be rectangular as shown in Fig. 6.

Table 5: List of parameter variations for domain randomization in the NuFingers Environment.

Stiffness of Elastic Bands	
Variation	<i>Uniform</i> (0 N/m, 100 N/m)
Application	At the object base
Object Width	
Variation	<i>Uniform</i> (15 mm, 25 mm)
Application	Parallel to the grasping direction
Object Location on the Plane	
Variation	Original location + <i>Uniform</i> (-5 mm, 5 mm)
Application	Perpendicular to the grasping direction



(a) NuFingers Environment

(b) Approximated NuFingers Environment

Figure 6: For stable contact between the surfaces in MuJoCo, the object in the NuFingers environment is approximated as a rotating block.

A.4 Safety during Exploration

Although it is evident that the proposed policy is successful in learning a safe policy under uncertainties, it is not yet clear whether the process of acquiring such policy is safe. In existing works, safety assessment of the exploration phase is usually disregarded, but we compare the safety of the different approaches during training and establish that SCAPE is safe and successful both during and after training. To accurately examine the safety during the acquisition of successful policies,

we measure the safety-related success rates during the exploration. Note that during exploration, we add Gaussian noise to the action to improve the policy. The corresponding success rates are shown in Fig. 7, which shows a significant performance gap between SCAPE and position-controlled poli-

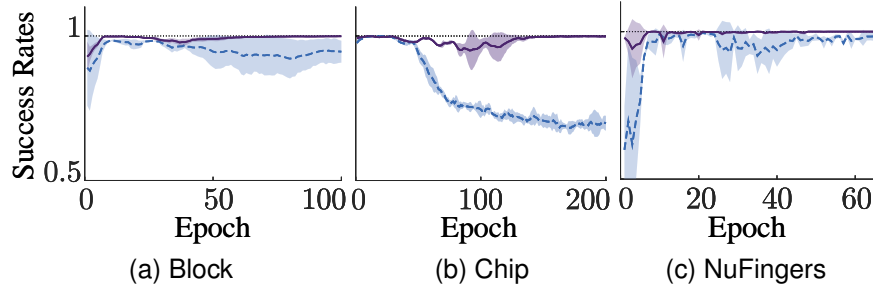


Figure 7: Safety-related success rates of SCAPE (solid) compared to position control (dashed) for each environment during exploration.

cies. SCAPE shows consistently safe performance throughout training, whereas almost half the time the position control policies apply greater force than what the object can withstand in either the beginning or ending phase of learning. Interestingly, these trends are similar to the evaluation results with deterministic policies. Note that the relatively high safety-related success rates of the position control for the Block and NuFingers environments are due to the failure in learning to manipulate the object, resulting in minimal interaction. From these results, we conclude that SCAPE is safer and superior than the existing position control approach both during the exploration and after training.

A.5 Comparison with a Hybrid Approach

The experiments shown in Fig. 3 demonstrate the performance gap between SCAPE and position control. Without the augmented demonstrations, the agent must learn position control to solve the problem or learn stiffness control from scratch as in Fig. 5a. However, if safety during policy improvement is not a concern, the agent can learn position control from demonstrations without the force penalty, and then learn stiffness modulation from scratch on top of the resulting policy. But such hybrid approach requires human input in determining the number of timesteps for each stage of learning, and safety during policy improvement severely deteriorates. Therefore, we include this analysis only in the supplementary material.

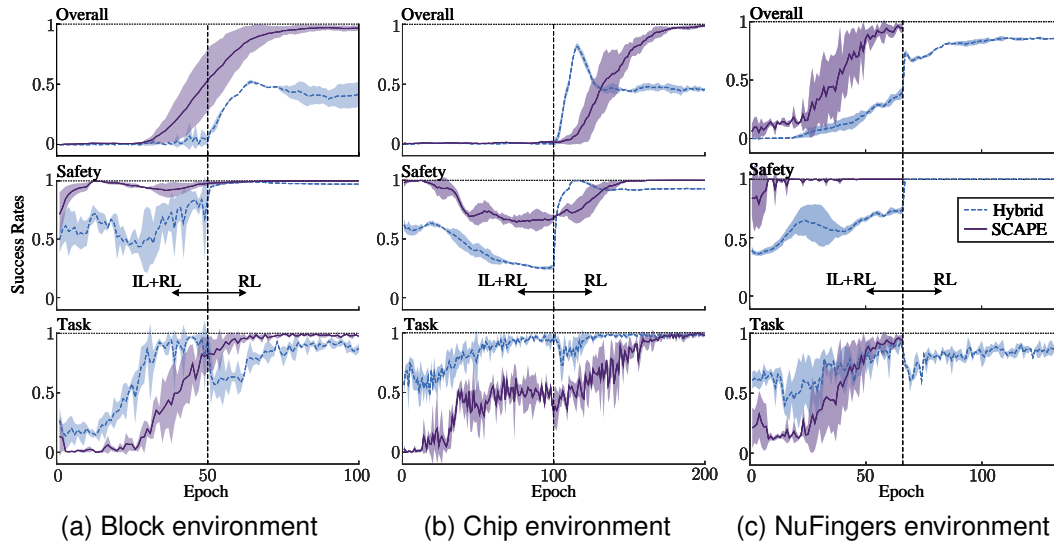


Figure 8: Resulting success rates of SCAPE compared with the hybrid approach. Success rates for task-related goals (e.g., did the object reach the target states?) and safety-related goals (e.g., how often did the object stay intact?) are separately plotted. Overall goals entail both goals.

For the hybrid approach, we first use imitation learning on top of reinforcement learning (IL+RL) until the agent successfully learns to complete the kinematic task (i.e., did the object reach the target states?) without the force penalty. This stage is identical to the work in [11]. Once the agent learns to solve the kinematic task, we switch to reinforcement learning (RL) and use the remaining timesteps to optimize the stiffness parameters with the force penalty included. In this stage, the agent does not have access to the augmented demonstrations as SCAPE does. For the Block and Chip environments, we assign half the total number of timesteps in each stage ($5e4$). For the NuFingers environment, however, we have found that the agent cannot reach the same level of task-related success rate as SCAPE with half the number of timesteps. Therefore, we double the amount of timesteps for the hybrid approach, although this provides an unfair advantage. The results suggest that even though the agents learn to reach 100% task-related success rates for each problem in the first stage, they all fail to optimize the stiffness parameters in the second stage. Ultimately, SCAPE outperforms the proposed hybrid approach in all problems, even in the NuFingers environment, where the hybrid approach is allowed twice the number of timesteps.

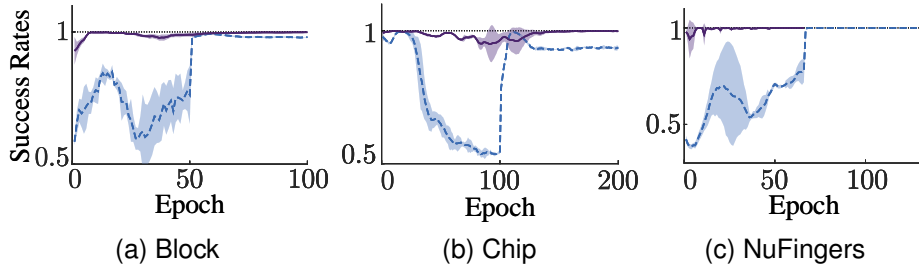


Figure 9: Safety-related success rates of SCAPE (solid) compared to the hybrid approach (dashed) for each environment during exploration.

Once training is completed, the hybrid approach appears to outperform the position control policy shown in Fig. 3. However, this comes at the cost of deteriorated safety during policy improvement as can be seen in Fig. 9. For example, the hybrid approach exerts forces above the breaking threshold almost 60% of the time in the beginning for the NuFingers environment. This is not only significantly more dangerous to use compared to SCAPE, but also compared to the position control approach shown in Fig. 7. The deteriorated safety is mainly due to the absence of the force penalty in the first stage (IL+RL) of the hybrid approach, which disregards the interaction force.