VI. Appendix

A. Experiment Details

In this section, we elaborate on the details of the tasks that we used for evaluating our method.

1) Simulated Table Tennis Task: We run our experiments in a custom MuJoCo environment, simulating the muscles with a Hill-type muscle model [2]. Similar to Büchler et al. [1], we replay ball trajectories until contact and simulate afterwards. The average episode length is 37-39 time steps, depending on the policy. The episode ends when the ball touches the racket or gets out of reach of the racket.

For HiS, parallel episodes continue after the main episode ended. The problem is that we sample the policy distribution conditioned on the state containing the main virtual object. In these cases, we sample the policy with one of the other virtual objects, until the episode is finished for all virtual objects.

2) Real Robot Table Tennis Task: We keep most of the task details similar to the original experiment in Büchler et al. [1] and list our adaptations.

- We employ our policy on a new iteration of the robot hardware.
- The initialization procedure is different: Instead of using a position controller, we send fixed target pressures to get the robot to an initial position.
- We send actions to the robot at a frequency of 25 Hz instead of 100 Hz because we found this to lead to significantly better results with SAC.

Collecting 15k episodes takes about 24 hours on the real robot.

3) FetchPush and FetchSlide: Each episode has a fixed length of 50 time steps. To apply HiS, we sample additional virtual objects according to the same distribution that is used to sample the main object.

B. Hyperparameters

The hyperparameters for the three tasks are shown in Tables I, II, and III. The SAC hyperparameters for the table tennis task were tuned on the simulated task without HiS. The gym robotics experiments incorporate the hyperparameters found by [3].

C. Ablation Study: Selection Criterion

Using the reward criterion, which selects mainly successful trajectories, works well for HiS on all our tasks. Because HER relabels trajectories as successful, it is less relevant to select successful HiS trajectories when applying HER and HiS together. Therefore, picking trajectories where the virtual object is moved, works even better in this case. In this work, we focused on these criteria to showcase our method. Our experiments indicat that the TD error criterion also helps the learning progress but not as much as the other two criteria. Table IV shows an evaluation of the different selection criteria. Note that the real table tennis task was evaluated for only one training run. The simulated table tennis task was averaged over 10 training runs with different random seeds, and the sliding task over 5 training runs. The table tennis task, both real and

TABLE I: Hyperparameters table tennis

	hyperparameter	value
SAC	gamma	0.9999
	ent_coef	0
	learning_rate	0.0003
	batch_size	256
	policy_network	MLP
	num_layers	1
	num_hidden	200
	gradient_steps	500
	train_freq	1
	train_freq_unit	episode
	buffer_size	5000000
	learning_starts	10000
HiS	criterion	reward per episode
	k_c	3
	ψ_c	0.5

TABLE II: Hyperparameters FetchPush

	hyperparameter	value
SAC	gamma	0.95
	ent_coef	auto
	learning_rate	0.001
	batch_size	256
	policy_network	MLP
	num_layers	2
	num_hidden	64
	gradient_steps	1
	train_freq	1
	train_freq_unit	step
	buffer_size	5000000
	learning_starts	1000
HER	goal_selection_strategy	future
	n_sampled_goal	4
HiS	criterion	Δx virt. object
	k_c	3
	ψ_c	0.02

TABLE III: Hyperparameters FetchSlide

	hyperparameter	value
SAC	gamma	0.95
	ent_coef	auto
	learning_rate	0.001
	batch_size	2048
	policy_network	MLP
	num_layers	3
	num_hidden	512
	gradient_steps	1
	train_freq	1
	train_freq_unit	step
	buffer_size	5000000
	learning_starts	1000
HER	goal_selection_strategy	future
	n_sampled_goal	4
HiS	criterion	Δx virt. object
	k_c	3
	ψ_c	0.02

simulated was evaluated after 15000 episodes, and the sliding task was evaluated after 5000 episodes.

TABLE IV: Evaluation of different selection criteria for HiS and HER+HiS on final success rate.

	Task		
	Sim. Table Tennis	Real Table Tennis	Sliding
van. SAC	41.2 %	34.5 %	9.9 %
HiS reward	70.6 %	44.7 %	33.7 %
HiS Δx virt. object			25.5 %
HiS TD error	50.8 %	34.2 %	13.2 %
HER+HIS reward			53.8 %
HER+HIS Δx virt. object			64.5 %

References Appendix

[1] Dieter Büchler, Simon Guist, Roberto Calandra, Vincent Berenz, Bernhard Schölkopf, and Jan Peters. Learning to play table tennis from scratch using muscular robots. *IEEE Transactions on Robotics*, 38(6):3850–3860, 2022.

[2] D. F. B. Haeufle, M. Günther, A. Bayer, and S. Schmitt. Hill-type muscle model with serial damping and eccentric force–velocity relation. *Journal of Biomechanics*, 47(6): 1531–1536, April 2014. ISSN 0021-9290. doi: 10.1016/j.jbiomech.2014.02.009. URL http://www.sciencedirect.com/science/article/pii/S0021929014001018.

[3] Antonin Raffin. RL Baselines3 Zoo. https://github.com/DLR-RM/rl-baselines3-zoo, 2020.