Appendix

A. Environment Details

Below, we provide environment details for each of the four experimental domains.

A.1. Sawyer Reaching

In this environment, which is based on the simulated Sawyer reaching task in the Meta-World suite (Yu et al., 2019), the goal is to reach a particular position. The target position, which is unobserved throughout, moves after each episode.

The episodes are 150 timesteps long, and the state is the position of the end-effector in (x, y, z) coordinate space. The actions correspond to changes in end-effector positions. The reward is defined as

$$r(\mathbf{s}, \mathbf{a}) = -\|\mathbf{s} - \mathbf{s}_g\|_2,$$

where $\mathbf{s}_g$ at episode i is defined as

$$\mathbf{s}_g = \begin{bmatrix} 0.1 \cdot \cos(0.5 \cdot i) \\ 0.1 \cdot \sin(0.5 \cdot i) \\ 0.2 \end{bmatrix}.$$

In other words, the sequence of goals is defined by a circle in the xy-plane. For the oracle comparison, the sequence of goals and the reward function are the same, except the state observation here is the concatenation of the end-effector position and the goal position $\mathbf{s}_g$.

A.2. Half-Cheetah WindVel

In this variant of Half-Cheetah from OpenAI Gym (Brockman et al., 2016), the agent needs to reach a target velocity in the x-direction, which varies across episodes, while subjected to varying wind forces. That is, the reward is

$$r(\mathbf{s}, \mathbf{a}) = -\|v_s - v_g\|_2 - 0.05 \cdot \|\mathbf{a}\|_2,$$

where v_s is the observed velocity of the agent. The state consists of the position and velocity of the agent's center of mass and the angular position and angular velocity of each of its six joints, and actions correspond to torques applied to each of the six joints. The target velocity v_g varies according to a sine function, i.e., the target velocity for episode i is

$$v_g = 1.5 + 1.5 \sin(0.5 \cdot i).$$

For the oracle comparison, the target velocity v_g is appended to the state observation. The force for each episode is defined by

$$f_w = 10 + 10 \sin(0.2 \cdot i)$$

and is applied constantly along the x-direction throughout the episode. Each episode, across all comparisons, is 50 timesteps long.

A.3. Minitaur Mass

We use the simulated Minitaur environment developed by Tan et al. (2018). We induce non-stationarity by varying the mass of the agent between episodes akin to increasing and decreasing payloads. Specifically, the mass at each episode is

$$m = 1.0 + 0.75 \sin(0.3 \cdot i).$$

The reward is defined by

$$r(\mathbf{s}_t, \mathbf{a}_t) = 0.3 - |0.3 - \mathbf{s}_{t,v}| - 0.01 \cdot \|\mathbf{a}_t - 2\mathbf{a}_{t-1} + \mathbf{a}_{t-2}\|_1,$$

where the first two terms correspond to the velocity reward, which encourages the agent to run close to a target velocity of 0.3 m/s, and the last term corresponds to an acceleration penalty defined by the last three actions taken by the agent. The state includes the angles, velocities, and torques of all eight motors, and the action is the target motor angle for each motor. Each episode is 100 timesteps long.

A.4. 2D Open World

Finally, we design an infinite, non-episodic environment in which the agent's goal is to collect red pellets and avoid blue pellets and other obstacles. Simultaneously, the agent is subjected non-stationary dynamics, in particular directed wind forces f_w and action re-scaling c. These dynamics change after every 100 timesteps (these boundaries of change are known to the agent), according to the following equations:

$$f_w = \begin{bmatrix} 0.015 \cos(0.2 \cdot \lfloor t/100 \rfloor) \\ 0.015 \sin(0.2 \cdot \lfloor t/100 \rfloor) \end{bmatrix}$$
$$c = 0.03 + 0.015 \sin(0.125 \cdot \lfloor t/100 \rfloor)$$

B. Hyperparameter Details

In this section, we provide the hyperparameter values used for each method.

B.1. LILAC (Ours)

Latent space. For our method, we use a latent space size of 8 in Sawyer Reaching and 2D Open World, and size of 40 in the other experiments: Half-Cheetah Vel, Half-Cheetah Wind+Vel, and Minitaur Mass.

Inference and decoder networks. The inference and decoder networks are MLPs with 2 fully-connected layers of size 64 in Sawyer Reaching and 2D Open World; 1 fully-connected layer of size 512 in Half-Cheetah Vel and Half-Cheetah Wind+Vel; and 2 fully-connected layers of size 512 in Minitaur Mass.

Policy and critic networks. The policy and critic networks are MLPs with 3 fully-connected layers of size 256 in the Sawyer Reaching experiment; and 2 fully-connected layers of size 256 in the other experiments.

For the Sawyer Reaching experiment, β_1 and β_2 are

$$\beta_1 = \begin{cases} 0, & \text{iter} < 10000 \\ 1, & \text{iter} \geq 10000 \end{cases}$$

$$\beta_2 = \begin{cases} 0, & \text{iter} < 10000 \\ \min(\text{1e-6} \cdot (\text{iter} - 10000), 1), & \text{iter} \geq 10000 \end{cases}$$

For Half-Cheetah WindVel, β_1 and β_2 are

$$\beta_1 = \begin{cases} 0, & \text{iter} < 50000 \\ 1, & \text{iter} \geq 50000 \end{cases}$$

$$\beta_2 = \begin{cases} 0, & \text{iter} < 10000 \\ \min(\text{1e-6} \cdot (\text{iter} - 10000), 1), & \text{iter} \geq 10000 \end{cases}$$

For the Minitaur Mass experiment, β_1 and β_2 are

$$\beta_1 = \begin{cases} 0, & \text{iter} < 10000 \\ 1, & \text{iter} \geq 10000 \end{cases}$$

$$\beta_2 = \begin{cases} 0, & \text{iter} < 10000 \\ \text{1e-6}, & \text{iter} \geq 10000 \end{cases}$$

Finally, for 2D Open World, β_1 and β_2 are

$$\beta_1 = 0$$

$$\beta_2 = \begin{cases} \text{1e-5}, & \text{iter} < 10000 \\ \min(\text{1e-5} \cdot (\text{iter} - 10000), 1), & \text{iter} \geq 10000 \end{cases}$$

B.2. Stochastic Latent Actor-Critic

Latent space. SLAC factorizes its per-timestep latent variable $\mathbf{z}_t$ into two stochastic layers $\mathbf{z}_t^1$ and $\mathbf{z}_t^2$, i.e. $p(\mathbf{z}_t) = p(\mathbf{z}_t^2|\mathbf{z}_t^1)p(\mathbf{z}_t^1)$. In the Sawyer Reaching and 2D Open World experiments, the size of $\mathbf{z}_t^1$ is 16 and the size of $\mathbf{z}_t^2$ is 8. In all other experiments, the size of $\mathbf{z}_t^1$ is 64 and the size of $\mathbf{z}_t^2$ is 32.

Inference and decoder networks. The inference and decoder networks are MLPs with 2 fully-connected layers of size 64 in Sawyer Reaching and 2D Open World; 1 fully-connected layer of size 512 in Half-Cheetah Vel and Half-Cheetah Wind+Vel; and 2 fully-connected layers of size 512 in Minitaur Mass.

Policy and critic networks. The policy and critic networks are MLPs with 3 fully-connected layers of size 256 in the Sawyer Reaching experiment; and 2 fully-connected layers of size 256 in the other experiments.

B.3. Soft Actor-Critic

Policy and critic networks. The policy and critic networks are MLPs with 3 fully-connected layers of size 256 in the Sawyer Reaching experiment; and 2 fully-connected layers of size 256 in the other experiments.

C. Significance Testing

Recognizing the challenges of reproducibility in deep RL (Henderson et al., 2018; Colas et al., 2018), we performed bootstrap analysis on our main experimental results to test their statistical significance, as recommended by Henderson et al. (2018) and Colas et al. (2018). The bootstrap confidence interval test recommended additional seeds for the SAC algorithm in the Minitaur task and no additional seeds for the other algorithms or environments, so we ran additional trials for a total of 5 seeds for SAC in the Minitaur task. Running the bootstrap test again with the additional trials showed that no more are needed for significant comparison. In Table 1, we report the results of the bootstrap confidence interval test, with 10000 bootstrap iterations, at significance level 0.05.

Task	Algorithm	CI for $\mu_1 - \mu_2$
Sawyer	SAC	[6.995, 16.608]
	SLAC	[11.861, 20.660]
	PPO	[46.729, 49.840]
HC WindVel	SAC	[14.094, 30.647]
	SLAC	[44.944, 65.299]
	PPO	[56.213, 72.702]
Minitaur	SAC	[0.181, 14.207]
	SLAC	[15.208, 24.571]
	PPO	[2.845, 7.900]
2D Open World	SAC	[0.158, 0.808]
	SLAC	[3.808, 4.842]
	PPO	[1.063, 4.128]

Table 1. The estimated confidence intervals for $\mu_1 - \mu_2$, where μ_1 is the average final return achieved by LILAC and μ_2 corresponds to that of the comparisons.