Appendix

A *ROBEL* task details

In this section, we outline details of the benchmark task presented in section 4.

A.1 *D'Claw* tasks

The action space of all *D'Claw* tasks (subsection 4.1) is a 1D vector of 9 *D'Claw* joint positions.

(i) **Pose**: This task involves posing *D'Claw* by driving its joints $\boldsymbol{\theta}_t$ to a desired joint angles $\boldsymbol{\theta}_{goal}$ sampled randomly from the feasible joint angle space at the beginning of the episode. The observation space s_t is a 36-size 1D vector that consists of the current joint angles $\boldsymbol{\theta}_t$, the joint velocities $\dot{\boldsymbol{\theta}}_t$, the error between the goal and current joint angles, and the last action. The reward function is defined as:

$$r_t = -\|\boldsymbol{\theta}_{goal} - \boldsymbol{\theta}_t\| - 0.1\left\|\dot{\boldsymbol{\theta}}_t * \mathbb{1}(|\dot{\boldsymbol{\theta}}| > 0.5)\right\|$$

Three variants of this task are provided:

(a) *DClawPoseFixed*: a static variant where the desired joint angles remain constant for the episode

(b) *DClawPoseRandom*: a dynamic variant where the desired joint angle is time-dependent and oscillates between two goal positions that are sampled at the beginning of the episode.

(c) *DClawPoseRandomDynamic*: same as previous. The joint damping, and the joint friction loss are randomized at the beginning of every episode.

Success evaluator metric $\phi_{se}(\pi)$ of policy π is defined using the mean absolute tracking error being within the threshold $\beta = 10^\circ$

$$\phi_{se}(\pi) = \mathbb{E}_{\tau\sim\pi}\Big[\frac{1}{T}\sum_{t=0}^{T} mean\Big\|\theta^{(\tau)}_{goal} - \theta^{(\tau)}_{t}\Big\| < \beta\Big]$$

(ii) **Turn**: This task involves rotating an object from an initial angle $\theta_{0,obj}$ to a goal angle $\theta_{goal,obj}$. The observation space is a 21-size 1D vector of the current joint angles $\boldsymbol{\theta}_t$, the joint velocities $\dot{\boldsymbol{\theta}}_t$, the sine and cosine values of the object's angle $\theta_{t,obj}$, the last action, and the error between the goal and the current object angle $\Delta\theta_{t,obj} = \theta_{t,obj} - \theta_{goal,obj}$. The reward function is defined as

$$r_t = -5|\Delta\theta_{t,obj}| - \|\boldsymbol{\theta}_{nominal} - \boldsymbol{\theta}_t\| - \left\|\dot{\boldsymbol{\theta}}_t\right\| + 10\mathbb{1}(|\Delta\theta_{t,obj}| < 0.25) + 50\mathbb{1}(|\Delta\theta_{t,obj}| < 0.1)$$

Three variants of this task are provided:

(a) *DClawTurnFixed*: constant initial angle (0°) and constant goal angle (180°).

(b) *DClawTurnRandom*: random initial and goal angle.

(c) *DClawTurnRandomDynamics*: same as previous. The position of the *D'Claw* relative to the object, the object's size, the joint damping, and the joint friction loss are randomized at the beginning of every episode.

Success evaluator metric $\phi_{se}(\pi)$ of policy π is defined using the error in last step of the episode $(t = T)$ being within the goal threshold $\beta = 0.1$ as:

$$\phi_{se}(\pi) = \mathbb{E}_{\tau\sim\pi}[\Delta\theta^{(\tau)}_{T,obj} < \beta]$$

(iii) **Screw**: This task involves rotating an object at a desired velocity $\dot{\theta}_{desired}$ from an initial angle. This is represented by a $\theta_{t,goal}$ that is updated every step as $\theta_{t,goal} = \theta_{t-1,goal} + \dot{\theta}_{desired} * dt$. Screw tasks have the same observation space and reward definitions as the Turn tasks. Three variants of this task are provided:

(a) *DClawScrewFixed*: constant initial angle (0°) and velocity ($0.5\frac{\text{rad}}{\text{sec}}$)

(b) *DClawScrewRandom*: random initial angle ($[-180^\circ, 180^\circ]$) and desired velocity ($[-0.75\frac{\text{rad}}{\text{sec}}, 0.75\frac{\text{rad}}{\text{sec}}]$)

(c) *DClawScrewRandomDynamics*: same as previous. The position of the *D'Claw* relative to the object, the object's size, the joint damping, and the joint friction loss are randomized at the beginning of every episode.

Success evaluator metric $\phi_{se}(\pi)$ of policy π is defined using the mean absolute tracking error being within the threshold $\beta = 0.1$

$$\phi_{se}(\pi) = \mathbb{E}_{\tau\sim\pi}\Big[\frac{1}{T}\sum_{t=0}^{T}|\Delta\theta^{(\tau)}_{t,obj}| < \beta\Big]$$

A.2 *D'Kitty* tasks

The action space of all of the *D'Kitty* tasks is a 1D vector of 12 joint positions. The observation space shares 49 common entries: the Cartesian position (3), Euler orientation (3), velocity (3), and angular velocity (3) of the *D'Kitty* torso, the joint positions $\boldsymbol{\theta}$ (12) and velocities $\dot{\boldsymbol{\theta}}$ (12) of the 12 joints, the previous action (12), and 'uprightness' $u_{t,kitty}$ (1). The uprightness $u_{t,kitty}$ of the *D'Kitty* is measured as it's orientation projected over the global vertical axis:

$$u_{t,kitty} = \boldsymbol{R}_{\hat{z},t,kitty} \cdot \hat{\boldsymbol{Z}}$$

The *D'Kitty* tasks share a common term in the reward function $r_{t,upright}$ regarding uprightness defined as:

$$r_{t,upright} = \alpha_{upright}\frac{u_{t,kitty} - \beta}{1-\beta} + \alpha_{falling}(u_{t,kitty} < \beta)$$

where β is the cosine similarity threshold with the global z-axis beyond which we consider the *D'Kitty* to have fallen. When perfectly upright $\alpha_{t,upright}$ reward is collected, when alignment ($u_{t,kitty}$) falls below the threshold β, the episode terminates early and $\alpha_{falling}$ is collected.

(i) **Stand**: This task involves *D'Kitty* coordinating its 12 joints $\boldsymbol{\theta}_t$ to stand upright maintaining a pose specified by $\boldsymbol{\theta}_{goal}$. The observation space is a 61-size 1D vector of the shared observation space entries and pose error $\boldsymbol{e}_{t,pose} = (\boldsymbol{\theta}_{goal} - \boldsymbol{\theta}_t)$. The reward function is defined as:

$$r_t = r_{t,upright} - 4\bar{e}_{t,pose} - 2||\boldsymbol{p}_{t,kitty}||_2 + 5u_{t,kitty}\mathbb{1}(\bar{e}_{t,pose} < \frac{\pi}{6}) + 10u_{t,kitty}\mathbb{1}(\bar{e}_{t,pose} < \frac{\pi}{12})$$

where $\bar{e}_{t,pose}$ is mean absolute pose error, $\boldsymbol{p}_{t,kitty}$ is the cartesian position of *D'Kitty* on the horizontal plane and the shared reward function constants are $\alpha_{upright} = 2$, $\alpha_{falling} = -100$, $\beta = \cos(90^\circ)$.

Three variants of this task are provided:

(a) *DKittyStandFixed*: constant initial pose.

(b) *DKittyStandRandom*: random initial pose.

(c) *DKittyStandRandomDynamics*: same as previous. The joint gains, damping, friction loss, geometry friction coefficients, and masses are randomized. In addition, a randomized height field is generated with heights up to $0.05\mathrm{m}$

The successor evaluator indicates success if the mean pose error is within the goal threshold $\beta = \frac{\pi}{12}$ and the *D'Kitty* is sufficiently upright at the last step ($t = T$) of the episode:

$$\phi_{se}(\pi) = \mathbb{E}_{\tau\sim\pi}[\mathbb{1}(\bar{e}^{(\tau)}_{T,pose} < \beta) * \mathbb{1}(u^{(\tau)}_{T,kitty} > 0.9)]$$

(ii) **Orient**: This task involves *D'Kitty* matching its current facing direction $\boldsymbol{\omega}_t$ with a goal facing direction $\boldsymbol{\omega}_{goal}$, thus minimizing the facing angle error $e_{t,facing}$ between $\boldsymbol{\omega}_{desired}$ and $\boldsymbol{\omega}_t$. The observation space is a 53-size 1D vector of the shared observation space entries, $\boldsymbol{\omega}_t$ and $\boldsymbol{\omega}_{goal}$ represented as unit vectors on the (X,Y) plane, and angle error $e_{t,facing}$. The reward function is defined as:

$$\begin{aligned} r_t &= r_{t,upright} - 4e_{t,facing} - 4||\boldsymbol{p}_{t,kitty}||_2 + r_{bonus\_small} + r_{bonus\_big} \\ r_{bonus\_small} &= 5(e_{t,facing} < 15^\circ \text{ or } u_{t,kitty} > \cos(15^\circ)) \\ r_{bonus\_big} &= 10(e_{t,facing} < 5^\circ \text{ and } u_{t,kitty} > \cos(15^\circ)) \end{aligned}$$

where the shared reward function constants are $\alpha_{upright} = 2$, $\alpha_{falling} = -500$, $\beta = \cos(25^\circ)$.

Three variants of this task are provided:

(a) *DKittyOrientFixed*: constant initial facing ($0°$) and goal facing ($180°$).
(b) *DKittyOrientRandom*: random initial facing ($[-60°, 60°]$) and goal facing ($[120°, 240°]$)
(c) *DKittyOrientRandomDynamics*: same as previous. The joint gains, damping, friction loss, geometry friction coefficients, and masses are randomized. In addition, a randomized height field is generated with heights up to 0.05m

The successor evaluator indicates success if the facing angle error is within the goal threshold and the *D'Kitty* is sufficiently upright at the last step ($t = T$) of the episode:

$$\phi_{se}(\pi) = \mathbb{E}_{\tau\sim\pi}[\mathbb{1}(e^{(\tau)}_{T,facing} < 5°) * \mathbb{1}(u^{(\tau)}_{T,kitty} > \cos(15°)]$$

(iii) **Walk**: This task has the *D'Kitty* move its current Cartesian position $\boldsymbol{p}_{t,kitty}$ to a desired Cartesian position $\boldsymbol{p}_{goal}$, minimizing the distance $d_{t,goal} = ||\boldsymbol{p}_{goal} - \boldsymbol{p}_{t,kitty}||_2$. Additionally, the *D'Kitty* is incentivized to face towards the goal. The heading alignment is calculated as $h_{t,goal} = \boldsymbol{R}_{\hat{y},t,kitty} \cdot \frac{\boldsymbol{p}_{goal} - \boldsymbol{p}_{t,kitty}}{d_{t,goal}}$. The observation space is a 52-size 1D vector of the shared observation space entries, $h_{t,goal}$ and $\boldsymbol{p}_{goal} - \boldsymbol{p}_{t,kitty}$.
The reward function is defined as:

$$\begin{aligned} r_t &= r_{t,upright} - 4d_{t,goal} + 2h_{t,goal} + r_{bonus\_small} + r_{bonus\_big} \\ r_{bonus\_small} &= 5(d_{t,goal} < 0.5 \text{ or } h_{t,goal} > \cos(25°)) \\ r_{bonus\_big} &= 10(d_{t,goal} < 0.5 \text{ and } h_{t,goal} > \cos(25°)) \end{aligned}$$

and the shared reward function constants are $\alpha_{upright} = 1$, $\alpha_{falling} = -500$, $\beta = \cos(25°)$.
Three variants of this task are provided:

(a) *DKittyWalkFixed*: constant distance (2m) towards $0°$.
(b) *DKittyWalkRandom*: random distance ($[1, 2]$) towards random angle ($[-60°, 60°]$)
(c) *DKittyWalkRandomDynamics*: same as previous. The joint gains, damping, friction loss, geometry friction coefficients, and masses are randomized. In addition, a randomized height field is generated with heights up to 0.05m

The successor evaluator indicates success if the goal distance is within a threshold and the *D'Kitty* is sufficient upright at the last step of the episode:

$$\phi_{se}(\pi) = \mathbb{E}_{\tau\sim\pi}[\mathbb{1}(d^{(\tau)}_{T,goal} < 0.5) * \mathbb{1}(u^{(\tau)}_{T,kitty} > \cos(25°))]$$

A.3 Safety metrics

The following safety scores are shared between all tasks.

(i) **Position violations**: This score indicates that the joint positions are near their operating bounds. For the N joints of the robot, this is defined as:

$$s_{position} = \sum_{i=1}^{N} \Big(\mathbb{1}(|\theta_i - \beta_{i,lower}| < \epsilon) + \mathbb{1}(|\theta_i - \beta_{i,upper}| < \epsilon)\Big)$$

where $\beta_{i,lower}$ and $\beta_{i,upper}$ is the respective lower and upper joint position bound for the ith joint, and ϵ is the threshold within which the joint position is considered to be near the bound.

(ii) **Velocity violations**: This score indicates that the joint velocities are exceeding a safety limit For the N joints of the robot, this is defined as:

$$s_{velocity} = \sum_{i=1}^{N} \mathbb{1}(|\dot{\theta}_i| > \alpha_i)$$

where α_i is the speed limit for the ith joint.

(iii) **Current violations**: This score indicates that the joints are exerting forces that exceed a safety limit. For the N joints of the robot, this is defined as:

$$s_{current} = \sum_{i=1}^{N} \mathbb{1}(|k_i| > \gamma_i)$$

where γ_i is the current limit for the ith joint.

B Locomotion benchmark performance on *D'Kitty*

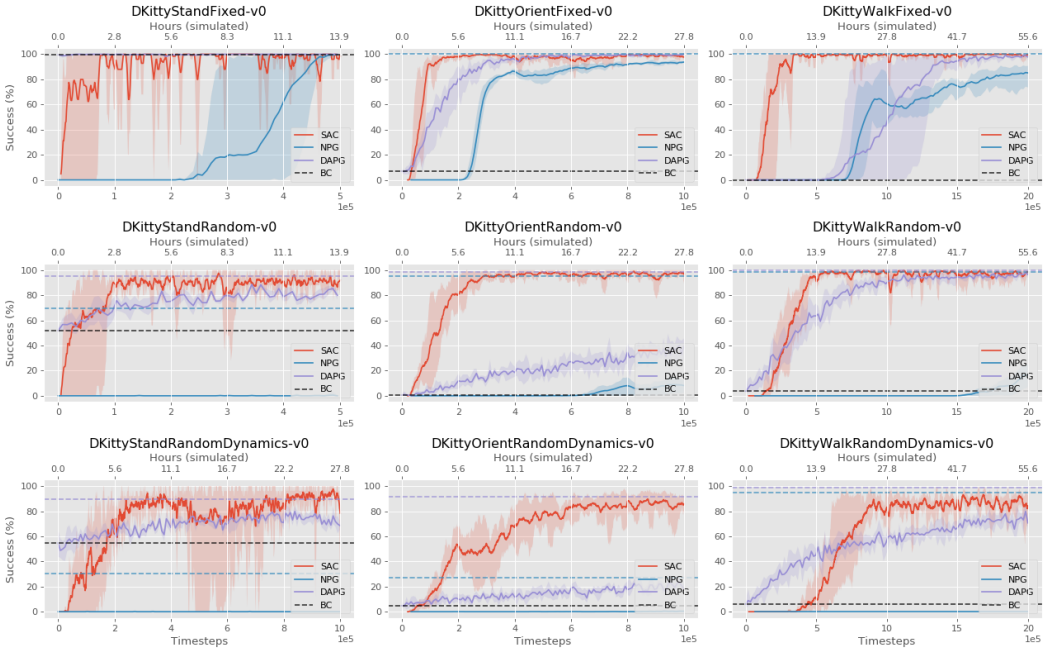

Figure 11: Success percentage (3 seeds) for all *D'Kitty* tasks trained on a simulated *D'Kitty* robot using Soft Actor Critic (SAC), Natural Policy Gradient (NPG), Demo-Augmented Policy Gradient (DAPG), and Behavior Cloning (BC) over 20 trajectories. Each timestep corresponds to 0.1 simulated seconds.

C *ROBEL* reproducibility

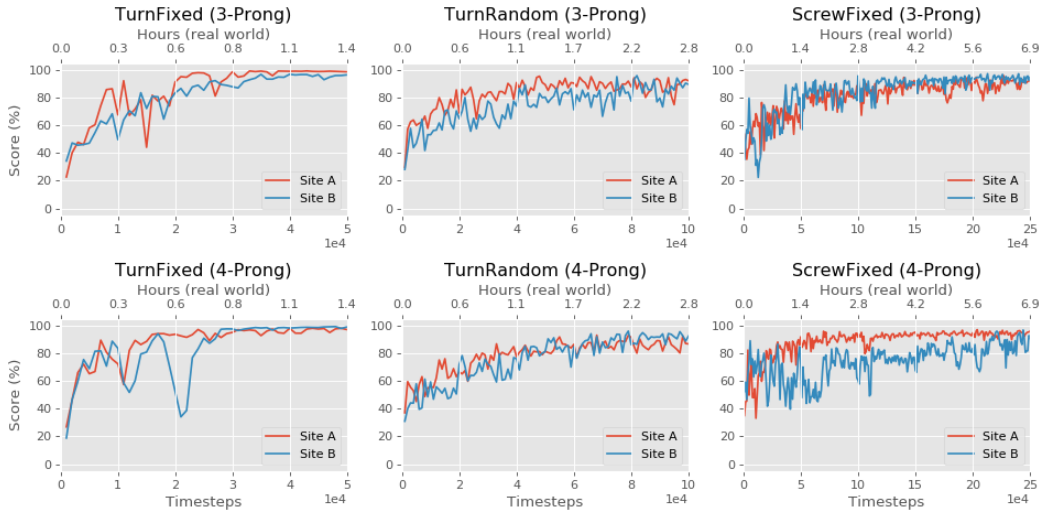

Figure 12: SAC training performance of *D'Claw* tasks on two real *D'Claw* robots each at different laboratory locations. Score denotes the closeness to the goal. Each timestep corresponds to 0.1 simulated seconds. Each task is trained over two different task objects: a 3-prong valve and a 4-prong valve.