

352 Appendix

353 A Reward Definitions

354 A.1 Opening Rewards

The opening reward r_o is composed of the reward for opening the door to the target angle r_{od} along with rewards for manipulating the door handle denoted by r_{hm} . For pull doors, we additionally include a reward for encouraging the robot to move its base and end-effector around the door panel denoted by r_{adp} . These rewards together compose the opening reward

$$r_o = 3r_{od} + \begin{cases} r_{hm} & \theta < 30^\circ \\ \bar{r}_{hm} + 0.5r_{adp} & \text{otherwise} \end{cases}$$

355 When the door has been opened enough, set as $\theta > 30^\circ$, r_{hm} is set to its maximum value \bar{r}_{hm} as it
 356 is no longer necessary for the policy to interact with the handle to open the door further. r_{adp} is only
 357 applied once the door has been opened enough.

The handle manipulation reward is composed of the following components

$$r_{hm} = r_{ehd} + r_{th} + r_{eho} + 0.5r_{hg} + r_{plg}$$

358 All individual reward terms in r_o are defined as follows.

- r_{ehd} (end-effector to handle): Minimizes the distance between the end-effector point \mathbf{e} and the handle point \mathbf{h} :

$$r_{ehd} = \exp(-\|\mathbf{e} - \mathbf{h}\|_2)$$

- r_{th} (turn handle): Rewards increasing the handle turning angle ϕ :

$$r_{th} = \phi / \phi_{\max}$$

359 where ϕ_{\max} is the maximum the handle can be turned.

- r_{eho} (end-effector grasp orientation): Rewards the end-effector for tracking a desired orientation for grasping the handle.

$$r_{eho} = 1 - \frac{|e_o|}{\pi}$$

360 where e_o angular error between the end-effector orientation and the desired end-effector orienta-
 361 tion.

- r_{hg} (handle in end-effector grasp): Give a binary reward when the handle point \mathbf{h} is within the grasp zone \mathcal{G} of the end-effector.

$$r_{hg} = \begin{cases} \mathbf{1}_{\mathcal{G}}(\mathbf{h}) & \|\mathbf{e} - \mathbf{h}\|_2 \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

362 $\mathbf{1}_{\mathcal{A}}(x)$ is the indicator function of value 1 if $x \in \mathcal{A}$ and 0 otherwise. For the hook-end effector
 363 used in this work, we defined the grasp zone \mathcal{G} as the region along the opening of the hook. This
 364 reward is only active when the end-effector point \mathbf{e} is close enough to the handle (within 1 m).

- r_{plg} (penalize lost grasp): Give a binary penalty when if the handle point \mathbf{h} is in the grasp zone at step $t - 1$ and leaves the grasp zone at t .

$$r_{plg} = \begin{cases} -\mathbf{1}_{\mathcal{G}}(\mathbf{h}_{t-1})(1 - \mathbf{1}_{\mathcal{G}}(\mathbf{h}_t)) & \|\mathbf{e} - \mathbf{h}\|_2 \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

365 Similar to r_{hg} , r_{plg} is only active when the end-effector is close enough to the handle.

- r_{od} (open door to target angle): Rewards opening the door to the target opening angle $\bar{\theta}$

$$r_{od} = 1 - \frac{|\theta - \hat{\theta}|}{\hat{\theta}}$$

366 where θ is the door hinge joint angle. This reward can also be used to train an opening only policy
 367 that opens the door to θ . For training the opening and passing through policy θ is set to 75° .

- r_{adp} (move around the door panel): This reward is only applied for pull doors. Given zones \mathcal{Z}_1 and \mathcal{Z}_2 defined relative to the door panel as shown in Fig 4, r_{adp} is computed based on the locations of the base \mathbf{b} and the end-effector \mathbf{e} as

$$r_{adp} = \begin{cases} 1 & \mathbf{b} \in \mathcal{Z}_1 \\ 2 & \mathbf{b} \in \mathcal{Z}_2 \\ 0 & \text{otherwise} \end{cases} + \begin{cases} 1 & \mathbf{e} \in \mathcal{Z}_1 \\ 2 & \mathbf{e} \in \mathcal{Z}_2 \\ 0 & \text{otherwise} \end{cases}$$

369 A.2 Passing Rewards

- r_p (passing progress): Rewards the base velocity \mathbf{v}_B for moving along the unit progress vector \mathbf{p}

$$r_p = \max \left(1, \frac{\mathbf{p} \cdot \mathbf{v}_B}{\|\mathbf{v}_B\|_{\max}} \right)$$

370 where $\|\mathbf{v}_B\|_{\max}$ is the the max allowable commanded velocity of the locomotion controller.

371 A.3 Shaping Rewards

The shaping reward r_s is defined as

$$r_s = 0.3r_{ma} + 0.5r_{pbt} + r_{psa} + 0.1r_{pcl} + 2r_{pc}$$

372 Individual terms of r_s are defined as:

- r_{ma} (minimize arm motion): Rewards minimizing the arm joint velocities and accelerations

$$r_{ma} = \sum_{i=1}^6 \exp(0.01\dot{q}_i^2) + \exp(0.000001\ddot{q}_i^2)$$

373 where \dot{q}_i and \ddot{q}_i are the joint velocity and acceleration for the i^{th} arm joint respectively.

- r_{pbt} (penalize base tilt): Penalizes large tilt of the robot base. The base tilt angle ψ can be computed from the projected gravity vector expressed in the robot base frame \mathbf{g}_B and expressed in the world frame \mathbf{g}_W as follows

$$\psi = \arccos \left(\frac{\mathbf{g}_W \cdot \mathbf{g}_B}{\|\mathbf{g}_W\| \|\mathbf{g}_B\|} \right)$$

374 Then $r_{pbt} = -1$ if $\psi > \bar{\psi}$, where $\bar{\psi}$ is a tilt threshold, and 0 otherwise. We set $\bar{\psi}$ as 8° .

- r_{psa} (penalize stretched arm): Penalize the arm from reaching out too far to prevent singular arm configurations.

$$r_{psa} = -\text{clip} \left(\frac{\|\mathbf{e} - \mathbf{s}\| - (0.7 - 0.1)}{0.1}, 0, 1 \right)$$

375 where \mathbf{e} and \mathbf{s} are the locations of the end-effector and shoulder joint.

- r_{pcl} (penalize command out of limits): As the arm PD target and locomotion commands are clipped within certain bounds we penalize the policy for commands that exceed these bounds.

$$r_{pcl} = -\sum_{i=1}^9 \text{clip} \left(\frac{|a_i| - \bar{a}_i}{\sigma_i}, 0, 1 \right)$$

376 where a_i , \bar{a}_i , and σ_i corresponding to the action, action limit, and penalty ramp up speed for the
377 i^{th} component of the policy's output action. The action limits are discussed in Sec. 3.0.2.

- r_{pc} (penalize collisions): Penalizes robot collisions.

$$r_{pc} = -\sum_{c \in \mathcal{C}} \left(\begin{cases} 1 & \|\lambda_c\| > 0 \\ 0 & \text{otherwise} \end{cases} \right)$$

379 where $\lambda_{(\cdot)}$ is the contact force on robot link (\cdot) and \mathcal{C} is the set of robot links where collisions are
380 penalized including the base, thighs, and arm.

B Domain Randomization Parameters

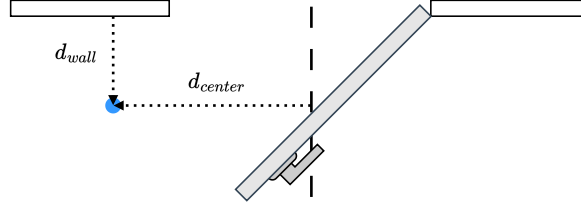


Figure 11: The initial robot location relative to the door is determined at the beginning of each episode by sampling d_{wall} and d_{center} .

The following randomizations are resampled for each new episode:

- Initial Base Location: Set relative to the doorway by the distances d_{wall} and d_{center} as shown in Fig. 11. d_{wall} and d_{center} are sampled uniformly from $[1, 2]$ m and $[-2, 2]$ m respectively.
- Initial Base Yaw: We define a yaw of 0° as the robot facing forwards along the direction of the doorway. The initial yaw is sampled uniformly from $[-180, 180]^\circ$.
- Initial Base Velocity: The initial base velocity components v_x and v_y are sampled uniformly from $[-0.5, 0.5]$ m/s.
- Door Panel Mass: Sampled uniformly from $[15, 75]$ kg.
- Door Hinge Resistance Torque: Sampled uniformly from $[0, 30]$ Nm, set to 0 with probability 0.2.
- Door Handle Resistance Torque: Sampled uniformly from $[0, 3]$ Nm, set to 0 with probability 0.2.
- Door Hinge Damping Torques: The hinge damping torque comprises of the air resistance given by $K_d^{ar}\dot{\theta}^2$ and the door closer mechanism damping given by $K_d^{dc}\dot{\theta}$. We sample K_d^{ar} uniformly from $[0, 4]$ Nms². For most doors, the door closer's damping is tuned to prevent the door from closing too quickly. To model this, we set K_d^{dc} to be some multiple α of the hinge resistance torque, where α is sampled uniformly from $[1.5, 3]$ s. The hinge damping torque is set to 0 with probability 0.4.
- Maximum Handle Turning Angle: Sampled uniformly from $[15, 90]^\circ$.
- Arm Joint Proportional Gain: Sampled uniformly from $[40, 60]$.
- Arm Joint Damping Gain: Sampled uniformly from $[3, 6]$.

We generated door models with different dimensions that are loaded into the simulation during initialization. The randomized door dimensions are shown in Fig. 12.

- d_W : Sampled uniformly from $[0.8, 1.0]$ m.
- d_T : Sampled uniformly from $[0.02, 0.06]$ m.
- h_L : Sampled uniformly from $[0.08, 0.12]$ m.
- h_H : Sampled uniformly from $[0.7, 1.3]$ m.
- h_O : Sampled uniformly from $[0.03, 0.12]$ m.

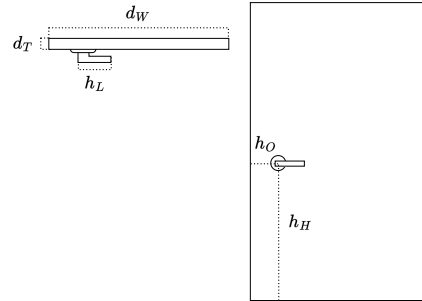


Figure 12: Randomized dimensions of the door and handle.