

Hyper-parameters	Values	Obs Type	Dimension	Hyper-parameters	Values
$\lambda_{\text{rot}}$	1.0	$q_t$	$\mathbb{R}^{3 \times 16}$	# environments	48
$\lambda_z$	-1.0	$a_{t-1}$	$\mathbb{R}^{3 \times 16}$	# steps	512
$\lambda_{\text{vel}}$	-0.3	$c_t$	$\mathbb{R}^{32}$	# minibatches	4096
$\lambda_{\text{diff}}$	-0.1	$p_t$	$\mathbb{R}^{4 \times 3}$	# epochs	2000
$\lambda_{\text{ang}}$	-0.3	$w_t$	$\mathbb{R}^7$	learning rate	1e-3
$\lambda_{\text{torque}}$	-0.1	PointCloud	$\mathbb{R}^{100 \times 3}$		
$\lambda_{\text{work}}$	-1.0				

Table 3: Hyper-parameters for the reward function.

Table 4: Dimensions of the inputs of the oracle policy.

Table 5: Hyper-parameters for training the student policy in the simulation.

## 357 A Implementation Details

### 358 A.1 Training Hyper-parameters

359 Our reward function is a combination of  $r_{\text{rot}}, r_z$  and  $r_{\text{energy}}$ . The energy reward consists of  
 360  $r_{\text{vel}}, r_{\text{diff}}, r_{\text{ang}}, r_{\text{torq}}$ , and  $r_{\text{work}}$ . Here,  $r_{\text{vel}}$  penalizes the pen’s linear velocity,  $r_{\text{diff}}$  discourages  
 361 the hand’s pose from deviating much from its initial pose,  $r_{\text{ang}}$  penalizes the pen’s angular velocity  
 362 above a pre-defined threshold to encourage stable rotation,  $r_{\text{torq}}$  penalizes large torques, and  $r_{\text{work}}$   
 363 penalizes the work of the controller. We follow the same definition of reward in [24]. We combine  
 364 the above rewards with weights listed in Table 3.

365 We detail the dimensions of the inputs of our oracle policy in Table 4. We train our oracle policy  
 366 with PPO, and the training hyper-parameters are shown in Table 6. Specifically, we train with 8192  
 367 parallel environments. Each environment gathers # steps data to train in each epoch of PPO. The  
 368 data is split into # minibatches and optimized with PPO loss.  $\gamma$  and  $\lambda$  are used for computing  
 369 generalized advantage estimate (GAE) returns. We use the Adam optimizer to train PPO and adopt  
 370 the gradient clip to stabilize training. We train 5000 epochs in total, which takes less than one day on  
 371 a single GPU. We train our student policy with Behavior Cloning, and the training hyper-parameters  
 372 are shown in Table 5. We collect approximately 50M steps of data in total.

Hyper-parameters	Values
# environments	8192
# steps	12
# minibatches	16384
# epochs	5000
$\gamma$	0.99
$\lambda$	0.95
learning rate	5e-3
clip range	0.2
entropy coefficient	0.0
kl threshold	0.02
max gradient norm	1.0

Table 6: Hyper-parameters for training the oracle policy.