

Hyper-parameters	Values
λ_{rot}	1.0
λ_z	-1.0
λ_{vel}	-0.3
λ_{diff}	-0.1
λ_{ang}	-0.3
λ_{torque}	-0.1
λ_{work}	-1.0

Table 3: Hyper-parameters for the reward function.

Obs Type	Dimension
\mathbf{q}_t	$\mathbb{R}^{3 \times 16}$
\mathbf{a}_{t-1}	$\mathbb{R}^{3 \times 16}$
\mathbf{c}_t	\mathbb{R}^{32}
\mathbf{p}_t	$\mathbb{R}^{4 \times 3}$
\mathbf{w}_t	\mathbb{R}^7
PointCloud	$\mathbb{R}^{100 \times 3}$

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

Hyper-parameters	Values
# environments	48
# steps	512
# minibatches	4096
# epochs	2000
learning rate	1e-3

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

A Implementation Details

A.1 Training Hyper-parameters

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

We detail the dimensions of the inputs of our oracle policy in Table 4. We train our oracle policy with PPO, and the training hyper-parameters are shown in Table 6. Specifically, we train with 8192 parallel environments. Each environment gathers # steps data to train in each epoch of PPO. The data is split into # minibatches and optimized with PPO loss. γ and λ are used for computing generalized advantage estimate (GAE) returns. We use the Adam optimizer to train PPO and adopt the gradient clip to stabilize training. We train 5000 epochs in total, which takes less than one day on a single GPU. We train our student policy with Behavior Cloning, and the training hyper-parameters 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
γ	0.99
λ	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.