Table 6: Hyper-parameters

| **Parameter** | **Value** |
|---|---|
| $\gamma$ | 0.90 |
| $\epsilon$ | 0.1 |
| $\beta$ | 0.1 |
| $\lambda$ | 0.1 |
| $N$ | 32 |
| $T$ | 1000 |
| $T_w$ | 10 |
| $e$ | 128 |
| $\eta_A$ | -1.0 |
| $\eta_C$ | 0.1 |
| $\eta_p$ | 0.1 |
| $\eta_i$ | 1.0 |
| $\eta_1$ | 1.0 |
| $\eta_2$ | 0.00001 |
| $\#attentionhead$ | 2 |
| attention dimension | 1024 |
| layers(Embed$_o$) | $[n, e]$ |
| layers(Embed$_x$) | $[x, e]$ |
| layers(Embed$_a$) | $[m, e]$ |
| layers(Embed$_s$) | $[(K+1)e, e]$ |
| layers(Proj$_s$) | $[e, n_s]$ |
| layers(Proj$_a$) | $[e, n_a]$ |
| layers(Proj$_v$) | $[e, 1]$ |
| learning rate (pretraining) | 0.00001 |
| learning rate (finetuning) | 0.0005 |

# A ADDITIONAL METHODOLOGY DETAILS

During the RL study, the agent generate its action from its policy $\pi(a_t|s_t)$ and push the environment stage forward. This interactive process yields the following episode sequence:

$$\tau_{0:T} = \{s_0, a_0, r_0, s_1, a_1, r_1, \cdots, s_T, a_T, r_T\} \tag{13}$$

in which $T$ means the episode ends or reaching the maximum time length. RL then solves the sequential decision making problem by finding $\pi$ such that $\max_\pi \mathbb{E}_{\tau\sim\pi}(R)$ in which the episode return defined as

$$R := \sum_{t=0}^{T} \gamma^t r_t \tag{14}$$

with $\gamma \in [0, 1)$ as the discounted factor.

The classical PPO methodology inherits from the famous actor-critic framework. The critic generates the state value estimate $V(s)$ , with its loss calculated from Bellman function by bootstrapping the state value function

$$L^{\text{Critic}} = \mathbb{E}_{s\sim d^\pi}[(r_t + \gamma(V_{\theta_{\text{old}}}(s_{t+1}) - V_\theta(s_t))^2]$$

On the other hand, generalized advantage estimation (GAE) (Schulman et al., 2016) is employed to help calculate the action advantage $A_t$ by traversing the episode backward

$$\hat{A}_t = \delta_t + (\gamma\lambda)\delta_{t+1} + \cdots + (\gamma\lambda)^{T-t+1}\delta_{T-1} \tag{15}$$

$$\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t) \tag{16}$$

then the actor is learned by maximizing the surrogate objective of $A_t$ according to the policy-gradient theorem

$$L^{\text{Actor}} = \text{CLIP}(\text{E}_t \frac{\pi_{\theta_{\text{old}}}(a_t|s_t)}{\pi_\theta(a_t|s_t)} \hat{A}_t - \beta\text{KL}[\pi_{\theta_{\text{old}}}(\cdot|s_t), \pi_\theta(\cdot|s_t)]) \tag{17}$$

Table 7: Environment morphology details

| **Environment** | $K$ | $n$ | $m$ | $x$ | $m_s$ | $m_a$ | $n_s$ | $n_a$ |
|---|---|---|---|---|---|---|---|---|
| swimmer | 2 | 2 | 1 | 4 | | | 8 | 2 |
| reacher | 2 | 3 | 1 | 5 | | | 11 | 2 |
| hopper | 3 | 2 | 1 | 5 | | | 11 | 3 |
| halfCheetah | 6 | 2 | 1 | 5 | | | 17 | 6 |
| walker2D | 6 | 2 | 1 | 5 | | | 17 | 3 |
| ant | 8 | 2 | 1 | 95 | | | 111 | 8 |
| humanoid | 9 | 6 | 3 | 342 | 20 | 10 | 376 | 17 |
| walker | 16 | 15 | 4 | 18 | 45 | 25 | 243 | 39 |
| unimal | 12 | 52 | 2 | 1410 | $*$ | $*$ | 624 | 24 |

$*$: varied from each agent morphology.

in which the CLIP function means clipping the object by $[1-\epsilon, 1+\epsilon]\hat{A}_t$, and KL denotes the famous K-L divergence.

A PPO policy update is then conducted by minimizing the objective upon each iteration:

$$L^{\text{PPO}} = -\eta_A L^{\text{Actor}} + \eta_C L^{\text{Critic}}$$

## B Additional Implementation Details

For practical consideration, input and output sequences are truncated by a window length $T_w$, with padding time mask for episodes shorter than $T_w$. The timestep embedding is also considered and concatenated into the latent variable.

To emphasize the instant impact, we further conduct a multi-head attention by querying the target variable and marking the input variable as key and value:

$$\hat{a}_t^p \leftarrow \hat{a}_t^p + \text{Attention}(\text{Q}=\hat{a}_t^p, \text{K}=s_t^p, \text{V}=s_t^p), \quad \hat{s}_t^p \leftarrow \hat{s}_t^p + \text{Attention}(\text{Q}=\hat{s}_t^p, \text{K}=a_{t-1}^p, \text{V}=a_{t-1}^p) \tag{18}$$

Table 6 lists most hyper-parameters in our implementation.

## C Additional Environmental Details

Table 7 exhibit all environments and their the morphological dimensions. Environments are from platforms including gym-mujoco, unimal and unity.

Table 8 shows statistics of all dataset used in the pretraining phase. Pretraining is computed distributed on 4 workers, each with 8 gpu, 4000 cpu and 400M memory. For both pretraining and finetuning, the ADAM optimizer is applied with the decay weight of 0.01.

Table 8: Pretraining dataset details

| Environment | Source | sampling agent | # samples | # episodes | ave ep return | ave ep length |
|---|---|---|---|---|---|---|
| hopper | D4RL | expert | 999,494 | 1,027 | 3511.36±328.59 | 973.22±97.84 |
| | D4RL | medium-expert | 1,999,400 | 3,213 | 2089.88±1039.96 | 622.28±263.22 |
| | D4RL | medium | 999,906 | 2,186 | 1422.06±378.95 | 457.41±110.88 |
| | D4RL | medium-replay | 402,000 | 2,041 | 467.30±511.03 | 196.96±195.15 |
| | D4RL | random | 999,996 | 45,239 | 18.40±17.45 | 22.10±11.99 |
| halfCheetah | D4RL | expert | 1,000,000 | 1,000 | 10656.43±441.68 | 1000.00±0.0 |
| | D4RL | medium-expert | 2,000,000 | 2,000 | 7713.38±2970.24 | 1000.00±0.0 |
| | D4RL | medium | 1,000,000 | 1,000 | 4770.33±355.75 | 1000.00±0.0 |
| | D4RL | medium-replay | 202,000 | 202 | 3093.29±1680.69 | 1000.00±0.0 |
| | D4RL | random | 1,000,000 | 1,000 | -288.80±80.43 | 1000.00±0.0 |
| walker2D | D4RL | expert | 999,214 | 1,000 | 4920.51±136.39 | 999.21±24.84 |
| | D4RL | medium-expert | 1,999,209 | 2,190 | 3796.57±1312.28 | 912.88±194.62 |
| | D4RL | medium | 999,995 | 1,190 | 2852.09±1095.44 | 840.33±240.13 |
| | D4RL | medium-replay | 302,000 | 1,093 | 682.70±895.96 | 276.30±263.13 |
| | D4RL | random | 999,997 | 48,907 | 1.87±5.81 | 20.45±8.46 |
| ant | D4RL | expert | 999,877 | 1,034 | 4620.73±1409.06 | 967.00±140.87 |
| | D4RL | medium-expert | 1,999,823 | 2,236 | 3776.93±1509.93 | 894.38±243.20 |
| | D4RL | medium | 999,946 | 1,202 | 3051.06±1180.59 | 831.90±290.71 |
| | D4RL | medium-replay | 302,000 | 485 | 976.05±1005.71 | 622.68±140.87 |
| | D4RL | random | 999,930 | 5,821 | -58.07±97.76 | 171.78±281.25 |
| walker | self | ppo | 1,001,861 | 7,928 | 262.74±281.18 | 126.37±119.17 |
| unimal | self | metamorph | 1,638,400 | 3,710 | 2.52±5.21 | 410.44±410.57 |