A Supplementary Illustration

A.1 Notations

Table 3: Notations.

SYMBOL	DEFINITION
s_t	THE ENVIRONMENT STATE AT TIMESTEP t
a_t	THE ACTION INTERACTING WITH THE ENVIRONMENT AT TIMESTEP t
r_t	THE REWARD FEEDBACK FROM THE ENVIRONMENT AT TIMESTEP t
$\tilde{s}_t$	THE STATE OBSERVED BY THE AGENT AT TIMESTEP t
$\tilde{a}_t$	THE ACTION TAKEN BY THE AGENT AT TIMESTEP t
$\tilde{r}_t$	THE REWARD RECEIVED BY THE AGENT AT TIMESTEP t
p^a	ACTION DELAY DISTRIBUTION
p^o	OBSERVATION DELAY DISTRIBUTION
p^{joint}	THE DISTRIBUTION OF THE SUM OF ACTION DELAY AND OBSERVATION DELAY
δ_a	MAXIMUM VALUE OF THE ACTION DELAY DISTRIBUTION
δ_o	MAXIMUM VALUE OF THE OBSERVATION DELAY DISTRIBUTION
δ_{joint}	MAXIMUM VALUE OF p^{joint}
d_p	WASSERSTEIN METRIC
$\bar{d}_p$	THE MAXIMAL FORM OF THE WASSERSTEIN METRIC

A.2 Interaction Process in Random Delay Environments

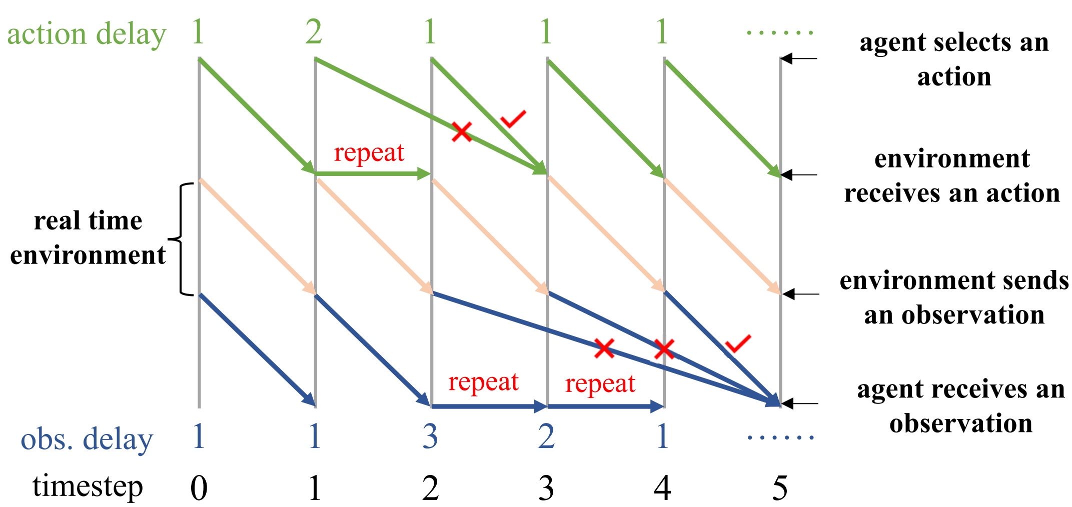


Figure 4: Interaction Process in Random Delay Environments.

B Proof

B.1 Proposition 1 Derivation

Proof. Under pure action delay, the stochastic delay representation $Z(\tilde{s}_t, \tilde{a}_t)$ at timestep t is expressed as:

$$Z(\tilde{s}_t, \tilde{a}_t) = \sum_{i=1}^{\delta_a} p_i^a \cdot (\tilde{r}_{t+i} + \gamma Z(\tilde{s}_{t+i+1}, \tilde{a}_{t+i+1})) \,. \tag{9}$$

Under pure observation delay, it takes the form:

$$Z(\tilde{s}_t, \tilde{a}_t) = \sum_{j=1}^{\delta_o} p_j^o \cdot (\tilde{r}_{t+j} + \gamma Z(\tilde{s}_{t+j+1}, \tilde{a}_{t+j+1})) \,. \tag{10}$$

In the presence of both delays, assuming independence of i and j as $\mathbb{P}(i,j) = p_i^a \cdot p_j^o$, the combined version of stochastic delay representation becomes:

$$Z(\tilde{s}_t, \tilde{a}_t) = \sum_{i=1}^{\delta_a} \sum_{j=1}^{\delta_o} p_i^a \cdot p_j^o \cdot \left(\tilde{r}_{t+(i+j)} + \gamma Z(\tilde{s}_{t+(i+j)+1}, \tilde{a}_{t+(i+j)+1})\right). \tag{11}$$

Let $k = i + j$ and define the joint delay distribution as:

$$p_k^{joint} = \sum_{\substack{i+j=k \\ i \in [1,\delta_a], j \in [1,\delta_o]}} p_i^a \cdot p_j^o, \tag{12}$$

with the total maximum delay denoted as $\delta_{joint} = \delta_a + \delta_o$. To simplify notation and unify bounds across formulations, we allow k to range from 1 to δ_{joint} and define $p_1^{joint} = 0$ to cover the case where no such (i,j) exists with $i + j = 1$. This does not affect the value of the summation but allows consistent indexing.

Substituting into Eq. (11) yields the unified form:

$$Z(\tilde{s}_t, \tilde{a}_t) = \sum_{k=1}^{\delta_{joint}} p_k^{joint} \cdot (\tilde{r}_{t+k} + \gamma Z(\tilde{s}_{t+k+1}, \tilde{a}_{t+k+1})). \tag{13}$$

Comparing Eq. (13) with Eqs. (9) and (10), we conclude that $Z(\tilde{s}_t, \tilde{a}_t)$ depends only on the aggregated delay k and its probability p_k^{joint}, and not on whether the delay originates from action or observation. Therefore, $Z(\tilde{s}_t, \tilde{a}_t)$ is invariant under the decomposition of k into i and j, establishing the equivalence of delay types in the stochastic delay representation.

B.2 Theorem 1 Derivation

In section B.2, we provide the proof of Theorem 1, which builds upon existing theoretical results in distributional reinforcement learning (Bellemare, Dabney, and Munos 2017; Duan et al. 2021; Nam, Kim, and Park 2021).

The stochastic delay representation with distributional returns operator derived from the Eq. (4) can be defined as

$$\mathcal{T}^\pi Z(\tilde{s}, \tilde{a}) = \sum_{i=1}^{\delta_a} p_i^a \sum_{j=1}^{\delta_o} p_j^o \left(\tilde{R}_{i+j}(\tilde{s}_{i+j}) + \gamma P^\pi Z(\tilde{s}_{i+j}, \tilde{a}_{i+j})\right) \tag{14}$$

where $P^\pi : \mathcal{Z} \to \mathcal{Z}$ is a state transition operator under policy π, with $\mathcal{Z}$ defined as the distributional value function space, and $P^\pi Z(\tilde{s}, \tilde{a}) \stackrel{D}{:=} Z(\tilde{s'}, \tilde{a'})$, where

$$\tilde{S'} \sim \mathop{\mathbb{E}}_{\substack{i \sim p_i^a \\ j \sim p_j^o}} P(\cdot \mid s_{-j}, \tilde{a}_{-i-j}) \tag{15}$$

and $\tilde{a}' \sim \pi(\cdot|\tilde{s}')$. The subscripts indicate the temporal relationship of states and actions relative to the current timestep. Specifically, $(\tilde{s}_{i+j}, \tilde{a}_{i+j})$ represents the state-action pair $i+j$ timesteps after the current pair $(\tilde{s}, \tilde{a})$, s_{-j} denotes the environment state i timesteps prior to the current pair, and $\tilde{a}_{-i-j}$ refers to the action taken $i + j$ timesteps prior to the current pair. We view the reward function as a random vector $\tilde{R}_{i+j} \in \mathcal{Z}$.

Lemma 1*: (Stochastic Delay Representation with Distributional Returns Policy Evaluation)*: Consider the stochastic delay representation with distributional returns backup operator $\mathcal{T}^\pi$, and a state–action distribution function $\mathcal{Z}_0(Z_0(\tilde{s}, \tilde{a})|\tilde{s}, \tilde{a}) : \mathcal{S} \times \mathcal{A} \to \mathcal{P}(Z_0(\tilde{s}, \tilde{a}))$, which maps a state–action pair $(\tilde{s}, \tilde{a})$ to a distribution over random state–action returns $Z_0(\tilde{s}, \tilde{a})$. Define $Z_{i+1}(\tilde{s}, \tilde{a}) = \mathcal{T}^\pi Z_i(\tilde{s}, \tilde{a})$, where $Z_{i+1}(\tilde{s}, \tilde{a}) \sim \mathcal{Z}_{i+1}(\cdot|\tilde{s}, \tilde{a})$. Then, the sequence $\mathcal{Z}_i$ will converge to $\mathcal{Z}^\pi$ as $i \to \infty$.

Proof. As proved in (Bellemare, Dabney, and Munos 2017), $\bar{d}_p$ is a metric over value distributions, which can be formulated as

$$\bar{d}_p(Z_1, Z_2) = \sup_{s,a} d_p(Z_1(s,a), Z_2(s,a)) \tag{16}$$

where d_p is the Wasserstein metric, which has the following properties:

$$d_p(aU, aV) \leq |a| d_p(U, V) \tag{P1}$$

$$d_p(A + U, A + V) \leq d_p(U, V) \tag{P2}$$

$$d_p\left(\sum_{i=1}^n U_i, \sum_{i=1}^n V_i\right) \leq \sum_{i=1}^n d_p(U_i, V_i) \tag{P3}$$

where a is a scalar, U and V are random variables, while U_i and V_i are independent.

Consider $Z_1, Z_2 \in \mathcal{Z}$. By definition,

$$\bar{d}_p(\mathcal{T}^\pi Z_1, \mathcal{T}^\pi Z_2) = \sup_{\tilde{s},a} d_p(\mathcal{T}^\pi Z_1(\tilde{s},\tilde{a}), \mathcal{T}^\pi Z_2(\tilde{s},\tilde{a})). \tag{17}$$

By the properties of d_p, we can obtain that

$$\begin{aligned}
&d_p(\mathcal{T}^\pi Z_1(\tilde{s},\tilde{a}), \mathcal{T}^\pi Z_2(\tilde{s},\tilde{a})) \\
&= d_p(\sum_{i=1}^{\delta_a} p_i^a \sum_{j=1}^{\delta_o} p_j^o \left(\tilde{R}_{i+j}(\tilde{s}_{i+j}) + \gamma P^\pi Z_1(\tilde{s}_{i+j}, \tilde{a}_{i+j})\right), \sum_{i=1}^{\delta_a} p_i^a \sum_{j=1}^{\delta_o} p_j^o \left(\tilde{R}_{i+j}(\tilde{s}_{i+j}) + \gamma P^\pi Z_2(\tilde{s}_{i+j}, \tilde{a}_{i+j})\right)) \\
&\leq d_p(\sum_{i=1}^{\delta_a} p_i^a \sum_{j=1}^{\delta_o} p_j^o \gamma P^\pi Z_1(\tilde{s}_{i+j}, \tilde{a}_{i+j}), \sum_{i=1}^{\delta_a} p_i^a \sum_{j=1}^{\delta_o} p_j^o \gamma P^\pi Z_2(\tilde{s}_{i+j}, \tilde{a}_{i+j})) \\
&\leq \sum_{i=1}^{\delta_a} p_i^a \sum_{j=1}^{\delta_o} p_j^o \gamma d_p(P^\pi Z_1(\tilde{s}_{i+j}, \tilde{a}_{i+j}), P^\pi Z_2(\tilde{s}_{i+j}, \tilde{a}_{i+j})) \\
&\leq \sum_{i=1}^{\delta_a} p_i^a \sum_{j=1}^{\delta_o} p_j^o \gamma \sup_{\tilde{s}'_{i+j}, \tilde{a}'_{i+j}} d_p(Z_1(\tilde{s}'_{i+j}, \tilde{a}'_{i+j}), Z_2(\tilde{s}'_{i+j}, \tilde{a}'_{i+j}))
\end{aligned} \tag{18}$$

By utilizing the definition of Eq. (17) and the results of Eq. (18), we obtain

$$\begin{aligned}
\bar{d}_p(\mathcal{T}^\pi Z_1, \mathcal{T}^\pi Z_2) &= \sup_{\tilde{s},\tilde{a}} d_p(\mathcal{T}^\pi Z_1(\tilde{s},\tilde{a}), \mathcal{T}^\pi Z_2(\tilde{s},\tilde{a})) \\
&\leq \sum_{i=1}^{\delta_a} p_i^a \sum_{j=1}^{\delta_o} p_j^o \gamma \sup_{\tilde{s}'_{i+j}, \tilde{a}'_{i+j}} d_p(Z_1(\tilde{s}'_{i+j}, \tilde{a}'_{i+j}), Z_2(\tilde{s}'_{i+j}, \tilde{a}'_{i+j})) \\
&= \sum_{i=1}^{\delta_a} p_i^a \sum_{j=1}^{\delta_o} p_j^o \gamma \bar{d}_p(Z_1, Z_2)
\end{aligned} \tag{19}$$

The weights p_i^o and p_j^a are normalized: $\sum_{i=1}^{\delta_a} p_i^a = 1, \quad \sum_{j=1}^{\delta_o} p_j^o = 1$. Combining these weights with the discount factor, we define an effective discount factor:

$$\gamma_{\text{eff}} = \sum_{i=1}^{\delta_a} p_i^a \sum_{j=1}^{\delta_o} p_j^o \gamma \in (0,1). \tag{20}$$

For any Z_1, Z_2, the operator satisfies:

$$\bar{d}_p(\mathcal{T}^\pi Z_1, \mathcal{T}^\pi Z_2) \leq \gamma_{\text{eff}} \cdot \bar{d}_p(Z_1, Z_2), \tag{21}$$

By Banach's fixed-point theorem, $\mathcal{T}^\pi$ has a unique fixed point Z^π, and the sequence $Z_i = \mathcal{T}^\pi Z_{i-1}$ will converges to it as $i \to \infty$, i.e., $\mathcal{Z}_i$ will converge to $\mathcal{Z}^\pi$ as $i \to \infty$.

Lemma 2*: (Policy Improvement)*: Let π_{new} be the optimal solution of maximizing cumulative returns. Then, $Q^{\pi_{\text{new}}}(\tilde{s},\tilde{a}) \geq Q^{\pi_{\text{old}}}(\tilde{s},\tilde{a})$ for $\forall(\tilde{s},\tilde{a}) \in \mathcal{S} \times \mathcal{A}$.

Proof. We can update the policy by maximizing the objective in terms of Q value

$$\pi_{new}(\cdot|\tilde{s}) = \arg\max_\pi \mathop{\mathbb{E}}_{\tilde{a}\sim\pi} [Q^{\pi_{old}}(\tilde{s},\tilde{a})] \quad \forall \tilde{s} \in \mathcal{S} \tag{22}$$

and the expected stochastic delay representation operator $\mathcal{T}_\pi$ can be written as

$$T^\pi Q^\pi(\tilde{s},\tilde{a}) = \sum_{i=1}^{\delta_a} p_i^a \cdot \sum_{j=1}^{\delta_o} p_j^o \cdot \left[\mathop{\mathbb{E}}_{\tilde{r}_{i+j}\sim\tilde{R}_{i+j}} [\tilde{r}_{i+j}(\tilde{s}_{i+j})] + \gamma \mathop{\mathbb{E}}_{\substack{\tilde{s}_{i+j+1}\sim p \\ \tilde{a}_{i+j+1}\sim\pi}} [Q^\pi(\tilde{s}_{i+j+1}, \tilde{a}_{i+j+1})] \right]. \tag{23}$$

Then, we can obtain that

$$\mathbb{E}_{\tilde{a}\sim\pi_{new}} [Q^{\pi_{old}}(\tilde{s},\tilde{a})] \geq \mathbb{E}_{\tilde{a}\sim\pi_{old}} [Q^{\pi_{old}}(\tilde{s},\tilde{a})] \quad \forall \tilde{s} \in \mathcal{S}. \tag{24}$$

By utilizing Eq. (24) and repeatedly expanding $Q^{\pi_{old}}$ on the right side with Eq. (23), we can derive the following results:

$$
\begin{aligned}
Q^{\pi_{old}}(\tilde{s},\tilde{a}) &= \sum_{i=1}^{\delta_a} p_i^a \cdot \sum_{j=1}^{\delta_o} p_j^o \cdot \left[\mathop{\mathbb{E}}_{\tilde{r}_{i+j}\sim\tilde{R}_{i+j}} [\tilde{r}_{i+j}(\tilde{s}_{i+j})] + \gamma \mathop{\mathbb{E}}_{\substack{\tilde{s}_{i+j+1}\sim p \\ \tilde{a}_{i+j+1}\sim\pi_{old}}} [Q^{\pi_{old}}(\tilde{s}_{i+j+1},\tilde{a}_{i+j+1})] \right] \\
&\le \sum_{i=1}^{\delta_a} p_i^a \cdot \sum_{j=1}^{\delta_o} p_j^o \cdot \left[\mathop{\mathbb{E}}_{\tilde{r}_{i+j}\sim\tilde{R}_{i+j}} [\tilde{r}_{i+j}(\tilde{s}_{i+j})] + \gamma \mathop{\mathbb{E}}_{\substack{\tilde{s}_{i+j+1}\sim p \\ \tilde{a}_{i+j+1}\sim\pi_{new}}} [Q^{\pi_{old}}(\tilde{s}_{i+j+1},\tilde{a}_{i+j+1})] \right] \\
&\dots \\
&\le Q^{\pi_{new}}(\tilde{s},\tilde{a}) \;\; \forall(\tilde{s},\tilde{a}) \in \mathcal{S}\times\mathcal{A}
\end{aligned} \tag{25}
$$

Thus, through the policy improvement, we can obtain $Q^{\pi_{\text{new}}}(\tilde{s},\tilde{a}) \ge Q^{\pi_{\text{old}}}(\tilde{s},\tilde{a})$ for $\forall(\tilde{s},\tilde{a}) \in \mathcal{S}\times\mathcal{A}$.

We define π_k as the policy at iteration k. From Lemma 1, we can obtain the $\mathcal{Z}^{\pi_k}$ for $\forall\pi_k$ through stochastic delay representation with distributional returns policy evaluation process. Since the expectation of distributional value function $Z(\tilde{s},\tilde{a})$ is the value $Q(\tilde{s},\tilde{a})$, i.e. $Q(\tilde{s},\tilde{a}) = \mathbb{E}[Z(\tilde{s},\tilde{a})]$, we can obtain $Q^{\pi_k}(\tilde{s},\tilde{a}) = \mathbb{E}[\mathcal{Z}^{\pi_k}(\tilde{s},\tilde{a})]$. By Lemma 2, $Q^{\pi_k}(\tilde{s},\tilde{a})$ is monotonically increasing for $\forall(\tilde{s},\tilde{a}) \in \mathcal{S}\times\mathcal{A}$. Since Q^{π} is bounded everywhere for $\forall\pi$, the policy sequence π_k converges to some $\pi^\dagger$ as $k\to\infty$, and it follows that

$$\mathbb{E}_{\tilde{a}\sim\pi^\dagger}\left[Q^{\pi^\dagger}(\tilde{s},\tilde{a})\right] \ge \mathbb{E}_{\tilde{a}\sim\pi}\left[Q^{\pi^\dagger}(\tilde{s},\tilde{a})\right] \quad \forall\pi\ \forall\tilde{s}\in\mathcal{S}. \tag{26}$$

Utilizing the results of Lemma 2, we can obtain

$$Q^{\pi^\dagger}(\tilde{s},\tilde{a}) \ge Q^{\pi}(\tilde{s},\tilde{a}) \quad \forall\pi\ \forall(\tilde{s},\tilde{a})\in\mathcal{S}\times\mathcal{A}. \tag{27}$$

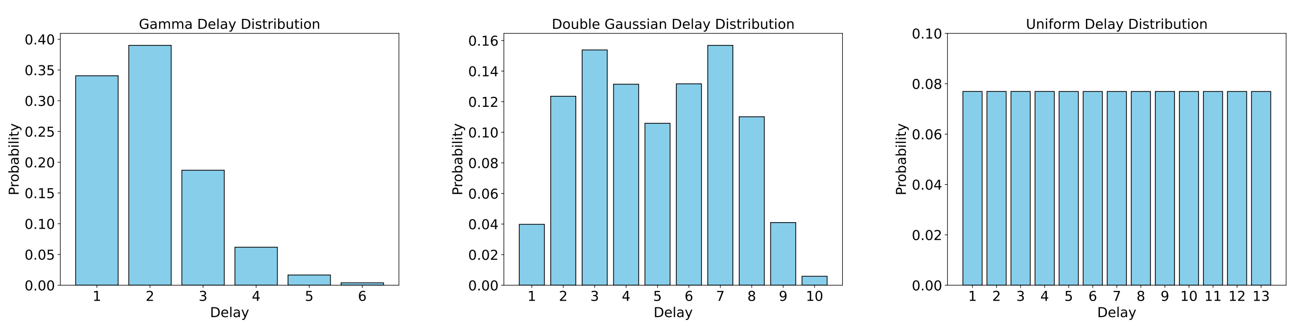
Therefore, $\pi^\dagger$ is optimal, i.e., $\pi^\dagger = \pi^*$.

It is worth mentioning that Eq. (4) is applicable to environments where both observation and action delays are present and can be simplified to Eq. (5) when only one type of delay is present in the environment. Thus, Eq. (5) exhibits the same convergence.

C Experimental Setup

C.1 Random Delay Environments

To better evaluate the algorithm's performance, we design three delay distributions to simulate random delays in real application scenarios, as shown in Figure 5. The gamma delay distribution has a range of 1 to 6 with an expectation of 2, and the double Gaussian delay distribution has a range of 1 to 10 with an expectation of 5, while the uniform delay distribution has a range of 1 to 13 with an expectation of 6.



(a) Gamma Distribution. (b) Double Gaussian Distribution. (c) Uniform Distribution.

Figure 5: The different distributions of random delays.

C.2 Hyperparameters

Table 4: Hyperparameter Settings

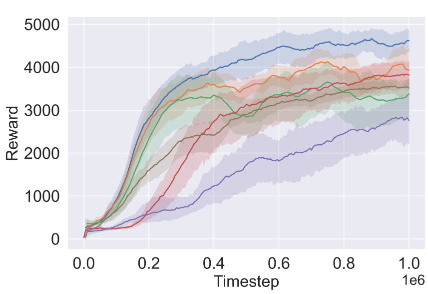
Hyperparameter	Setting
Network	[256, 256, 256]
Batch Size	256
Total Timesteps	1,000,000
Learning Rate	0.0001
Learning Rate for α	0.0003
Hidden Activation	GELU
Output Activation	Linear
γ	0.99
Optimizer	Adam
Initial α	0.2

C.3 Compute Resources

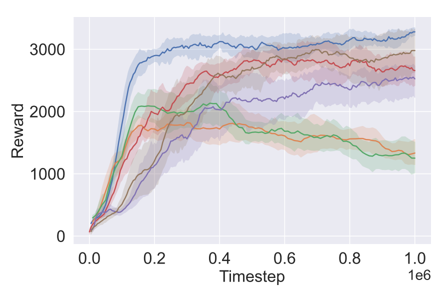
All experimental results across our experiments are obtained on servers equipped with Intel(R) Xeon(R) CPU E5-2678 v3 @ 2.50GHz and NVIDIA GeForce RTX3090.

D Experimental Results

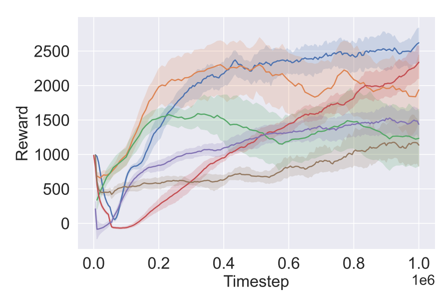
D.1 Comparative Experimental Results in Random Delay Environments



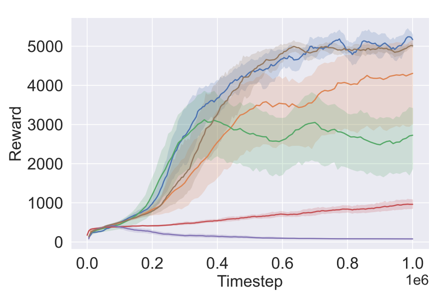
(a) Walker2d-v4



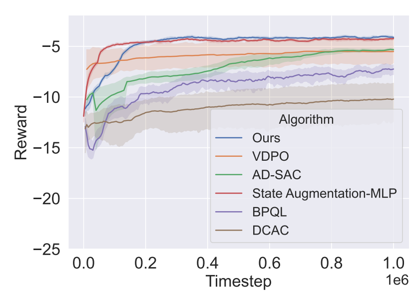
(b) Hopper-v4



(c) Ant-v4

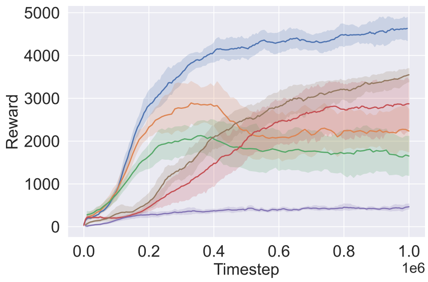


(d) Humanoid-v4

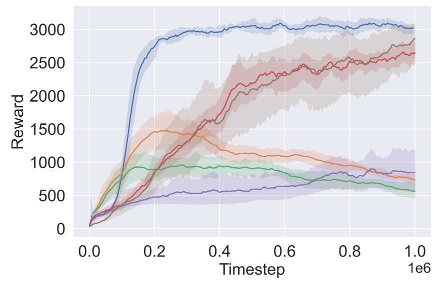


(e) Reacher-v4

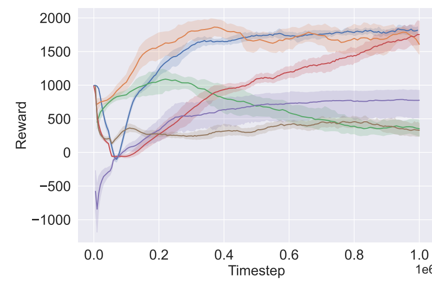
Figure 6: Comparison results in the gamma delay environment.



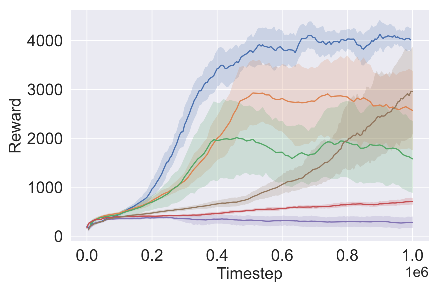
(a) Walker2d-v4



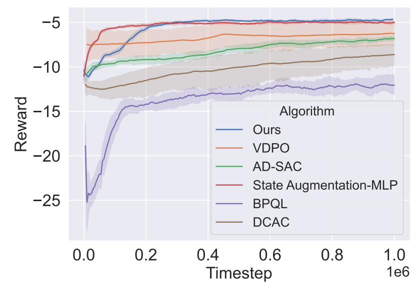
(b) Hopper-v4



(c) Ant-v4

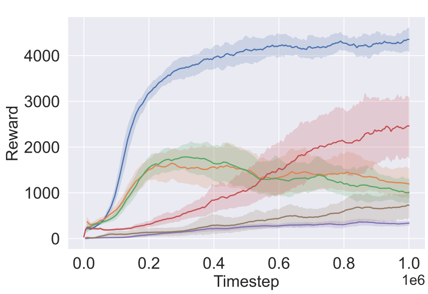


(d) Humanoid-v4

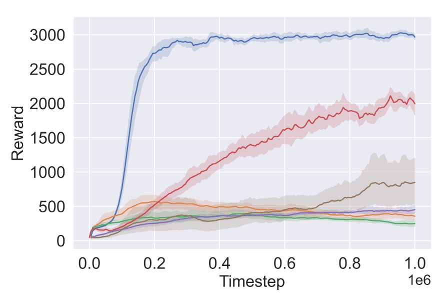


(e) Reacher-v4

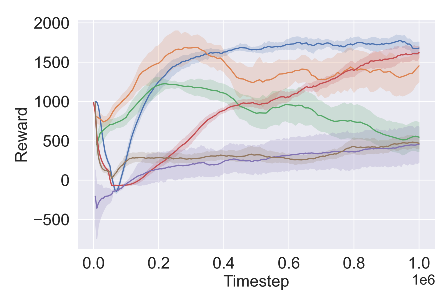
Figure 7: Comparison results in the double Gaussian delay environment.



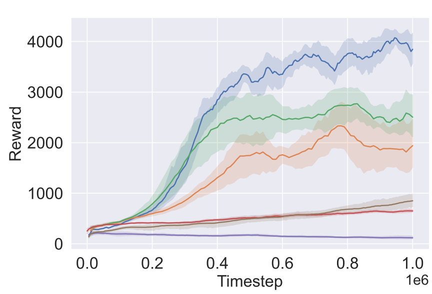
(a) Walker2d-v4



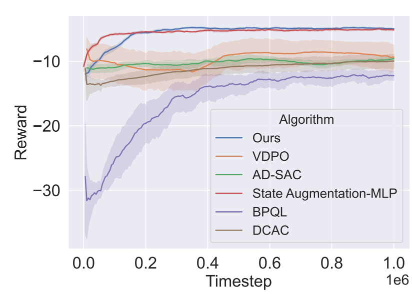
(b) Hopper-v4



(c) Ant-v4



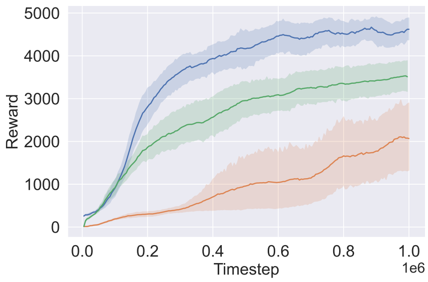
(d) Humanoid-v4



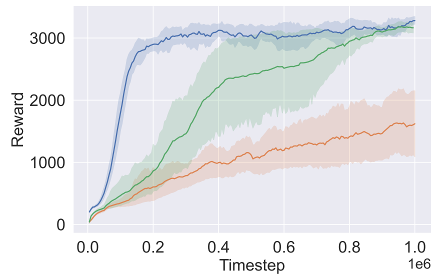
(e) Reacher-v4

Figure 8: Comparison results in the uniform delay environment.

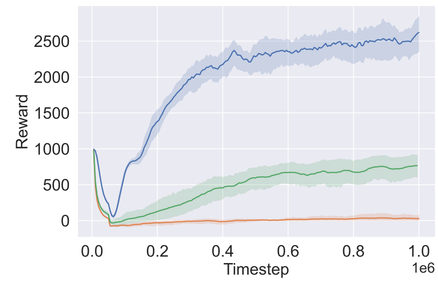
D.2 Ablation Experimental Results



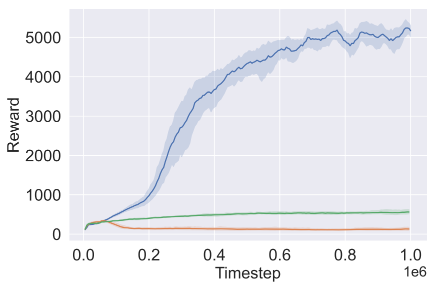
(a) Walker2d-v4



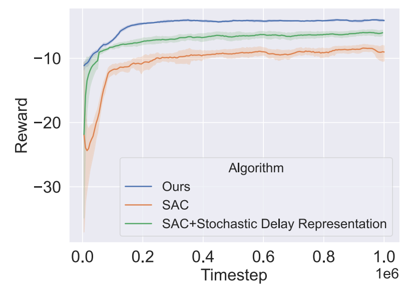
(b) Hopper-v4



(c) Ant-v4

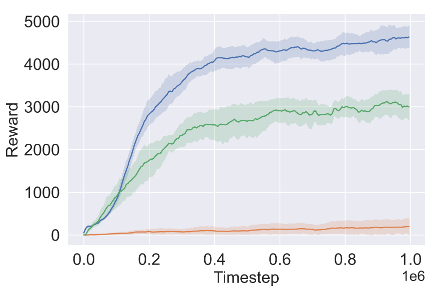


(d) Humanoid-v4

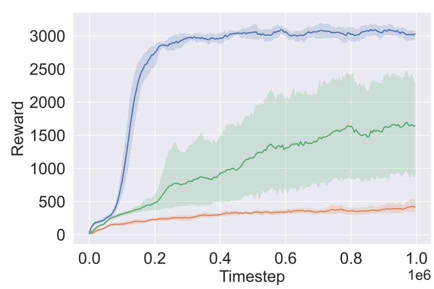


(e) Reacher-v4

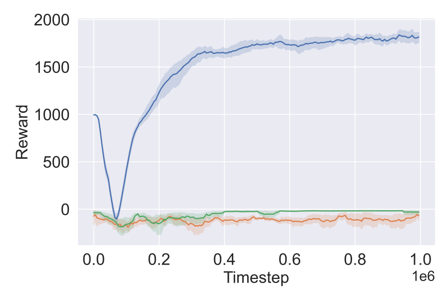
Figure 9: Ablation results in the gamma delay environment.



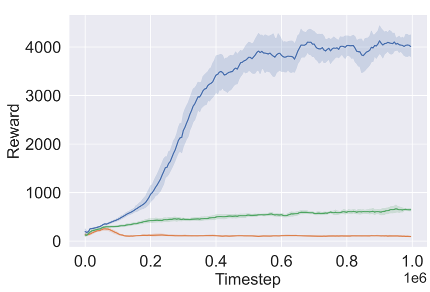
(a) Walker2d-v4



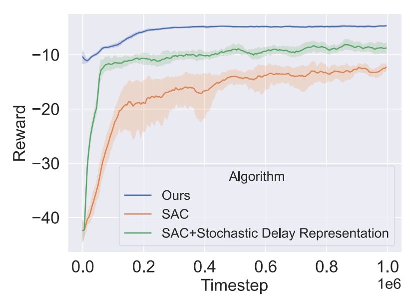
(b) Hopper-v4



(c) Ant-v4



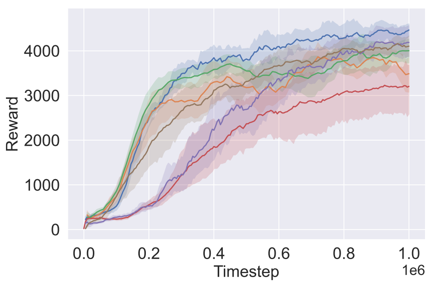
(d) Humanoid-v4



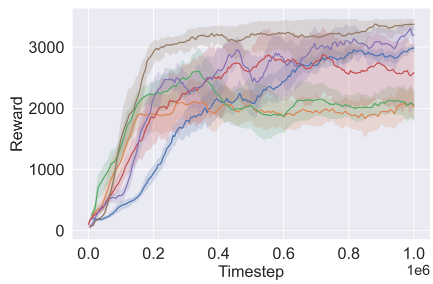
(e) Reacher-v4

Figure 10: Ablation results in the double Gaussian delay environment.

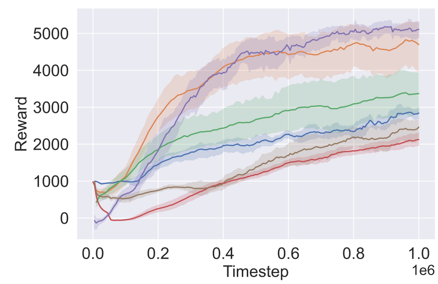
D.3 Comparative Experimental Results in a Constant Delay Environment



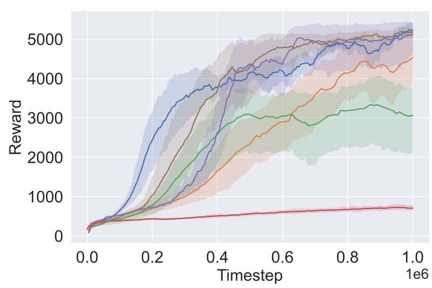
(a) Walker2d-v4



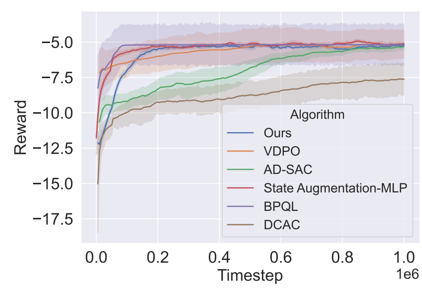
(b) Hopper-v4



(c) Ant-v4



(d) Humanoid-v4



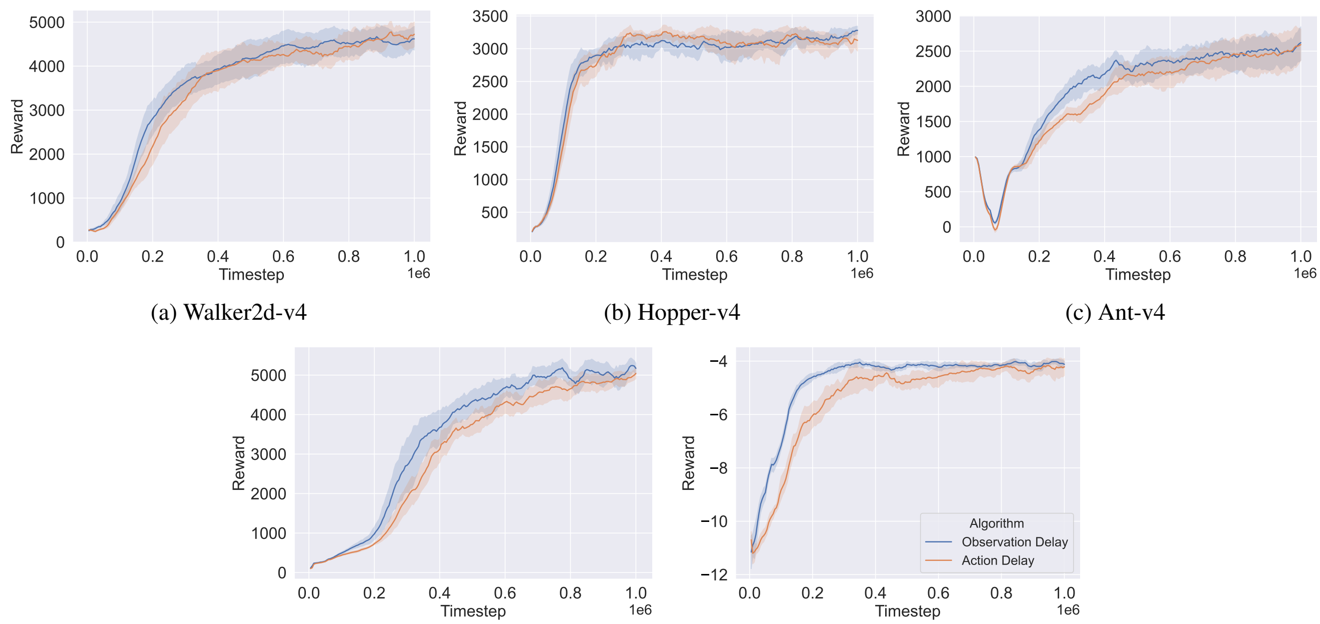
(e) Reacher-v4

Figure 11: Comparison results in the constant-5 delay environment.

Table 5: Results of comparative experiments within 1M global steps in a constant delay environment. Data are presented as mean ± standard error of the mean (S.E.M). The best performing methods, including those within the range of S.E.M of the best, are highlighted in bold.

Delay	Method	Walker2d-v4	Hopper-v4	Ant-v4	Humanoid-v4	Reacher-v4
Constant 5	VDPO(Wu et al. 2024a)	3589.2$_{\pm1458.7}$	1912.5$_{\pm439.5}$	4057.2$_{\pm2041.7}$	4950.0$_{\pm1201.4}$	-5.1$_{\pm1.5}$
	AD-SAC(Wu et al. 2024b)	3713.2$_{\pm943.1}$	2129.7$_{\pm635.5}$	3391.2$_{\pm1014.4}$	3650.0$_{\pm1546.9}$	**-5.1$_{\pm0.5}$**
	State Aug-MLP(Wang et al. 2024)	3355.7$_{\pm793.7}$	2828.4$_{\pm802.7}$	2171.1$_{\pm185.4}$	708.1$_{\pm60.6}$	**-5.1$_{\pm0.5}$**
	BPQL(Kim et al. 2023)	4250.3$_{\pm409.9}$	3224.4$_{\pm607.9}$	**5458.4$_{\pm167.5}$**	5181.4$_{\pm519.0}$	-5.2$_{\pm2.1}$
	DCAC(Bouteiller et al. 2021)	4320.3$_{\pm432.9}$	**3364.6$_{\pm163.8}$**	2629.8$_{\pm417.9}$	5213.0$_{\pm199.8}$	-7.7$_{\pm2.0}$
	D^{2}AC(ours)	**4564.2$_{\pm377.4}$**	2947.8$_{\pm235.8}$	3002.2$_{\pm245.6}$	**5280.1$_{\pm517.2}$**	-5.2$_{\pm0.6}$

D.4 Results of Evaluation of Handling Simultaneous Observation and Action Delays



(a) Walker2d-v4

(b) Hopper-v4

(c) Ant-v4

(d) Humanoid-v4

(e) Reacher-v4

Figure 12: Results of Evaluation of Handling Simultaneous Observation and Action Delays.