A ALGORITHM

A.1 VALUE-BASED EPISODIC MEMORY CONTROL

Algorithm 1 Value-based Episodic Memory Control

Initialize critic networks $V_{\theta_1}, V_{\theta_2}$ and actor network π_ϕ with random parameters θ_1, θ_2, ϕ
Initialize target networks $\theta_1' \leftarrow \theta_1, \theta_2' \leftarrow \theta_2$
Initialize episodic memory $\mathcal{M}$
for $t = 1$ **to** T **do**
  for $i \in \{1, 2\}$ **do**
    Sample N transitions $\left(s_t, a_t, r_t, s_t, R_t^{(i)}\right)$ from $\mathcal{M}$
    Update $\theta_i \leftarrow \min_{\theta_i} N^{-1} \sum \left(R_t^{(i)} - V_{\theta_i}(s_t)\right)^2$
    Update $\phi \leftarrow \max_\phi N^{-1} \sum \nabla \log \pi_\phi(a_t|s_t) \cdot f\left(\min_i R_t^{(i)} - \text{mean}_i V_{\theta_i}(s_t)\right)$
  end for
  if t mod u **then**
    $\theta_i' \leftarrow \kappa\theta_i + (1-\kappa)\theta_i'$
    Update Memory
  end if
end for

Algorithm 2 Update Memory

for trajectories τ in buffer $\mathcal{M}$ **do**
  for s_t, a_t, r_t, s_{t+1} in reversed(τ) **do**
    for $i \in \{1, 2\}$ **do**
      Compute $R_t^{(i)}$ with Equation 8 and save into buffer $\mathcal{M}$
    end for
  end for
end for

A.2 AN APPROACH FOR AUTO-TUNING τ

When we have a good estimation of V^*, for example, when there is some expert data in the dataset, we can auto-tune τ such that the value learned by EVL is close to the estimation of V^*. This can be done by calculating the Monte-Carlo return estimates of each state and selecting good return values as the estimation of optimal value $\tilde{V}^*$. Based on this target, we develop a method for auto-tuning τ.

By parameterizing $\tau = \text{sigmoid}(\xi)$ with a differentiable parameter $\xi \in \mathbb{R}$, we can auto-tune τ by minimizing the following loss $\mathcal{J}(\xi) = \xi(\mathbb{E}\hat{V}(s) - \tilde{V}^*)$. If $(\mathbb{E}\hat{V}(s) - \tilde{V}^*) < 0$, the differentiable parameter ξ will become larger and the value estimation $\mathbb{E}\hat{V}(s)$ will become larger accordingly. Similarly, ξ and $\mathbb{E}\hat{V}(s)$ will become smaller if $(\mathbb{E}\hat{V}(s) - \tilde{V}^*) > 0$. The experimental results in Figure 10 in Appendix D.1 show that auto-tuning can lead to similar performance compared with manual selection.

B THEORETICAL ANALYSIS

B.1 COMPLETE DERIVATION.

The expectile regression loss (Rowland et al., 2019) is defined as

$$\mathrm{ER}(q;\varrho,\tau)=\mathbb{E}_{Z\sim\varrho}\left[\left[\tau\mathbb{I}_{Z>q}+(1-\tau)\mathbb{I}_{Z\leq q}\right](Z-q)^2\right], \tag{13}$$

where ϱ is the target distribution and the minimiser of this loss is called the τ-expectile of ϱ. the corresponding loss in reinforcement learning is

$$\begin{aligned}\mathcal{J}_V(\theta)&=\mathbb{E}_\mu\left[\tau(r(s,a)+\gamma V_{\theta'}(s')-V_\theta(s))_+^2+(1-\tau)(r(s,a)+\gamma V_{\theta'}(s')-V_\theta(s))_-^2\right]\\&=\mathbb{E}_\mu\left[\tau(y-V_\theta(s))_+^2+(1-\tau)(y-V_\theta(s))_-^2\right].\end{aligned} \tag{14}$$

Then, taking the gradient of the value objective:

$$\begin{aligned}\nabla\mathcal{J}_V(\theta)&=\sum\mu(a\mid s)\left[-2\tau(y-V_\theta)_+\mathbb{I}(y-V_\theta)-2(1-\tau)(y-V_\theta)_+\mathbb{I}(y-V_\theta)\right]\\&=\sum\mu(a\mid s)\left[-2\tau(y-V_\theta)_+-2(1-\tau)(y-V_\theta)_+\right]\\&=\sum\mu(a\mid s)\left[-2\tau(\delta)_+-2(1-\tau)(\delta)_+\right].\end{aligned} \tag{15}$$

Therefore,

$$\begin{aligned}\hat{V}(s)&=V_\theta(s)-\alpha\nabla\mathcal{J}_V(\theta)\\&=V_\theta(s)+2\alpha\mathbb{E}_{a\sim\mu}\left[\tau[\delta(s,a)]_+ +(1-\tau)[\delta(s,a)]_-\right]\end{aligned} \tag{16}$$

B.2 PROOF OF LEMMA 1

Lemma 1. *For any $\tau\in[0,1)$, $\mathcal{T}_\tau^\mu$ is a γ_τ-contraction, where $\gamma_\tau=1-2\alpha(1-\gamma)\min\{\tau,1-\tau\}$.*

Proof. Note that $\mathcal{T}_{1/2}^\mu$ is the standard policy evaluation Bellman operator for μ, whose fixed point is V^μ. We see that for any V_1,V_2,

$$\begin{aligned}&\mathcal{T}_{1/2}^\mu V_1(s)-\mathcal{T}_{1/2}^\mu V_2(s)\\&=V_1(s)+\alpha\mathbb{E}_{a\sim\mu}[\delta_1(s,a)]-(V_2(s)+\alpha\mathbb{E}_{a\sim\mu}[\delta_2(s,a)])\\&=(1-\alpha)(V_1(s)-V_2(s))+\alpha\mathbb{E}_{a\sim\mu}[r(s,a)+\gamma V_1(s')-r(s,a)-\gamma V_2(s')]\\&\leq(1-\alpha)\|V_1-V_2\|_\infty+\alpha\gamma\|V_1-V_2\|_\infty\\&=(1-\alpha(1-\gamma))\|V_1-V_2\|_\infty.\end{aligned} \tag{17}$$

We introduce two more operators to simplify the analysis:

$$\begin{aligned}\mathcal{T}_+^\mu V(s)&=V(s)+\mathbb{E}_{a\sim\mu}[\delta(s,a)]_+,\\\mathcal{T}_-^\mu V(s)&=V(s)+\mathbb{E}_{a\sim\mu}[\delta(s,a)]_-.\end{aligned} \tag{18}$$

Next we show that both operators are non-expansion (e.g., $\|\mathcal{T}_+^\mu V_1-\mathcal{T}_+^\mu V_2\|_\infty\leq\|V_1-V_2\|_\infty$). For any V_1,V_2, we have

$$\begin{aligned}\mathcal{T}_+^\mu V_1(s)-\mathcal{T}_+^\mu V_2(s)&=V_1(s)-V_2(s)+\mathbb{E}_{a\sim\mu}[[\delta_1(s,a)]_+-[\delta_2(s,a)]_+]\\&=\mathbb{E}_{a\sim\mu}[[\delta_1(s,a)]_+ +V_1(s)-([\delta_2(s,a)]_+ +V_2(s))].\end{aligned} \tag{19}$$

The relationship between $[\delta_1(s,a)]_+ +V_1(s)$ and $[\delta_2(s,a)]_+ +V_2(s)$ exists in four cases, which are

- $\delta_1\geq0,\delta_2\geq0$, then $[\delta_1(s,a)]_+ +V_1(s)-([\delta_2(s,a)]_+ +V_2(s))=\gamma(V_1(s')-V_2(s'))$.
- $\delta_1<0,\delta_2<0$, then $[\delta_1(s,a)]_+ +V_1(s)-([\delta_2(s,a)]_+ +V_2(s))=V_1(s)-V_2(s)$.
- $\delta_1\geq0,\delta_2<0$, then

$$\begin{aligned}&[\delta_1(s,a)]_+ +V_1(s)-([\delta_2(s,a)]_+ +V_2(s))\\&=(r(s,a)+\gamma V_1(s'))-V_2(s)\\&<(r(s,a)+\gamma V_1(s'))-(r(s,a)+\gamma V_2(s'))\\&=\gamma(V_1(s')-V_2(s')),\end{aligned} \tag{20}$$

  where the inequality comes from $r(s,a)+\gamma V_2(s')<V_2(s)$.

- $\delta_1 < 0, \delta_2 \geq 0$, then

$$\begin{aligned}
&[\delta_1(s,a)]_+ + V_1(s) - ([\delta_2(s,a)]_+ + V_2(s)) \\
&= V_1(s) - (r(s,a) + \gamma V_2(s')) \\
&\leq V_1(s) - V_2(s),
\end{aligned} \tag{21}$$

  where the inequality comes from $r(s,a) + \gamma V_2(s') \geq V_2(s)$.

Therefore, we have $\mathcal{T}_+^\mu V_1(s) - \mathcal{T}_+^\mu V_2(s) \leq \|V_1 - V_2\|_\infty$. With the $\mathcal{T}_+^\mu, \mathcal{T}_-^\mu$, we rewrite $\mathcal{T}_\tau^\mu$ as

$$\begin{aligned}
\mathcal{T}_\tau^\mu V(s) &= V(s) + 2\alpha \mathbb{E}_{a\sim\mu}[\tau[\delta(s,a)]_+ + (1-\tau)[\delta(s,a)]_-] \\
&= (1-2\alpha)V(s) + 2\alpha\tau(V(s) + \mathbb{E}_{a\sim\mu}[\delta(s,a)]_+) + 2\alpha(1-\tau)(V(s) + \mathbb{E}_{a\sim\mu}[\delta(s,a)]_-) \\
&= (1-2\alpha)V(s) + 2\alpha\tau\mathcal{T}_+^\mu V(s) + 2\alpha(1-\tau)\mathcal{T}_-^\mu V(s).
\end{aligned} \tag{22}$$

And

$$\begin{aligned}
\mathcal{T}_{1/2}^\mu V(s) &= V(s) + \alpha \mathbb{E}_{a\sim\mu}[\delta(s,a)] \\
&= V(s) + \alpha(\mathcal{T}_+^\mu V(s) + \mathcal{T}_-^\mu V(s) - 2V(s)) \\
&= (1-2\alpha)V(s) + \alpha(\mathcal{T}_+^\mu V(s) + \mathcal{T}_-^\mu V(s)).
\end{aligned} \tag{23}$$

We first focus on $\tau < \frac{1}{2}$. For any V_1, V_2, we have

$$\begin{aligned}
&\mathcal{T}_\tau^\mu V_1(s) - \mathcal{T}_\tau^\mu V_2(s) \\
&= (1-2\alpha)(V_1(s) - V_2(s)) + 2\alpha\tau(\mathcal{T}_+^\mu V_1(s) - \mathcal{T}_+^\mu V_2(s)) + 2\alpha(1-\tau)(\mathcal{T}_-^\mu V_1(s) - \mathcal{T}_-^\mu V_2(s)) \\
&= (1-2\alpha-2\tau(1-2\alpha))(V_1(s) - V_2(s)) + 2\tau\left(\mathcal{T}_{1/2}^\mu V_1(s) - \mathcal{T}_{1/2}^\mu V_2(s)\right) + \\
&\qquad\qquad 2\alpha(1-2\tau)\left(\mathcal{T}_-^\mu V_1(s) - \mathcal{T}_-^\mu V_2(s)\right) \\
&\leq (1-2\alpha-2\tau(1-2\alpha))\|V_1 - V_2\|_\infty + 2\tau(1-\alpha(1-\gamma))\|V_1 - V_2\|_\infty + 2\alpha(1-2\tau)\|V_1 - V_2\|_\infty \\
&= (1-2\alpha\tau(1-\gamma))\|V_1 - V_2\|_\infty
\end{aligned} \tag{24}$$

Similarly, when $\tau > 1/2$, we have $\mathcal{T}_\tau^\mu V_1(s) - \mathcal{T}_\tau^\mu V_2(s) \leq (1-2\alpha(1-\tau)(1-\gamma))\|V_1 - V_2\|_\infty$. □

B.3 Proof of Lemma 2

Lemma 2. *For any $\tau, \tau' \in (0,1)$, if $\tau' \geq \tau$, we have $\mathcal{T}_{\tau'}^\mu \geq \mathcal{T}_\tau^\mu, \forall s \in S$.*

Proof. Based on Equation 22, we have

$$\begin{aligned}
&\mathcal{T}_{\tau'}^\mu V(s) - \mathcal{T}_\tau^\mu V(s) \\
&= (1-2\alpha)V(s) + 2\alpha\tau'\mathcal{T}_+^\mu V(s) + 2\alpha(1-\tau')\mathcal{T}_-^\mu V(s) \\
&\qquad - ((1-2\alpha)V(s) + 2\alpha\tau\mathcal{T}_+^\mu V(s) + 2\alpha(1-\tau)\mathcal{T}_-^\mu V(s)) \\
&= 2\alpha(\tau'-\tau)(\mathcal{T}_+^\mu V(s) - \mathcal{T}_-^\mu V(s)) \\
&= 2\alpha(\tau'-\tau)\mathbb{E}_{a\sim\mu}[[\delta(s,a)]_+ - [\delta(s,a)]_-] \geq 0.
\end{aligned} \tag{25}$$

□

B.4 Proof of Lemma 3

Lemma 3. *Let V^* denote the fixed point of Bellman optimality operator $\mathcal{T}^*$. In the deterministic MDP, we have $\lim_{\tau\to 1} V_\tau^* = V^*$.*

Proof. We first show that V^* is also a fixed point for $\mathcal{T}_+^\mu$. Based on the definition of $\mathcal{T}^*$, we have $V^*(s) = \max_a[r(s,a) + \gamma V^*(s')]$, which infers that $\delta(s,a) \leq 0, \forall s \in S, a \in A$. Thus, we have $\mathcal{T}_+^\mu V^*(s) = V^*(s) + \mathbb{E}_{a\sim\mu}[\delta(s,a)]_+ = V^*(s)$. By setting $(1-\tau) \to 0$, we eliminate the effect of $\mathcal{T}_-^\mu$. Further by the contractive property of $\mathcal{T}_\tau^\mu$, we obtain the uniqueness of V_τ^*. The proof is completed. □

B.5 PROOF OF LEMMA 4

Lemma 4. *Given $\tau \in (0,1)$ and $T \in \mathbb{N}^+$, $\mathcal{T}_{\text{vem}}$ is a γ_τ-contraction. If $\tau > \frac{1}{2}$, $\mathcal{T}_{\text{vem}}$ has the same fixed point as $\mathcal{T}_\tau^\mu$.*

Proof. We prove the contraction first. For any V_1, V_2, we have

$$\begin{aligned}\mathcal{T}_{\text{vem}}V_1(s) - \mathcal{T}_{\text{vem}}V_2(s) &= \max_{1\le n\le n_{\max}}\{(\mathcal{T}^\mu)^{n-1}\mathcal{T}_\tau^\mu V_1(s)\} - \max_{1\le n\le T}\{(\mathcal{T}^\mu)^{n-1}\mathcal{T}_\tau^\mu V_2(s)\}\\ &\le \max_{1\le n\le n_{\max}}|(\mathcal{T}^\mu)^{n-1}\mathcal{T}_\tau^\mu V_1(s) - (\mathcal{T}^\mu)^{n-1}\mathcal{T}_\tau^\mu V_2(s)|\\ &\le \max_{1\le n\le n_{\max}}\gamma^{n-1}\gamma_\tau\|V_1 - V_2\|_\infty\\ &\le \gamma_\tau\|V_1 - V_2\|_\infty.\end{aligned} \tag{26}$$

Next we show that V_τ^*, the fixed point of $\mathcal{T}_\tau^\mu$, is also the fixed point of $\mathcal{T}_{\text{vem}}$ when $\tau > \frac{1}{2}$. By definition, we have $V_\tau^* = \mathcal{T}_\tau^\mu V_\tau^*$. Following Lemma 2, we have $V_\tau^* = \mathcal{T}_\tau^\mu V_\tau^* \ge \mathcal{T}_{1/2}^\mu V_\tau^* = \mathcal{T}^\mu V_\tau^*$. Repeatedly applying $\mathcal{T}^\mu$ and using its monotonicity, we have $\mathcal{T}^\mu V_\tau^* \ge (\mathcal{T}^\mu)^{n-1}V_\tau^*, 1 \le n \le n_{\max}$. Thus, we have $\mathcal{T}_{\text{vem}}V_\tau^*(s) = \max_{1\le n\le T}\{(\mathcal{T}^\mu)^{n-1}\mathcal{T}_\tau^\mu V_\tau^*(s)\} = V_\tau^*(s)$. □

B.6 PROOF OF LEMMA 5

Lemma 5. *When the current value estimates $V(s)$ are much lower than the value of behavior policy, $\mathcal{T}_{\text{vem}}$ provides an optimistic update. Formally, we have*

$$|\mathcal{T}_{\text{vem}}V(s) - V_\tau^*(s)| \le \gamma^{n^*(s)-1}\gamma_\tau\|V - V_{n^*,\tau}^\mu\|_\infty + \|V_{n^*,\tau}^\mu - V_\tau^*\|_\infty, \forall s \in S, \tag{27}$$

where $n^*(s) = \arg\max_{1\le n\le T}\{(\mathcal{T}^\mu)^{n-1}\mathcal{T}_\tau^\mu V(s)\}$ and $V_{n^*,\tau}^\mu$ is the fixed point of $(\mathcal{T}^\mu)^{n^*(s)-1}\mathcal{T}_\tau^\mu$.

Proof. The lemma is a direct result of the triangle inequality. We have

$$\begin{aligned}\mathcal{T}_{\text{vem}}V(s) - V_\tau^*(s) &= (\mathcal{T}^\mu)^{n^*(s)-1}\mathcal{T}_\tau^\mu V(s) - V_\tau^*(s)\\ &= (\mathcal{T}^\mu)^{n^*(s)-1}\mathcal{T}_\tau^\mu V(s) - (\mathcal{T}^\mu)^{n^*(s)-1}\mathcal{T}_\tau^\mu V_{n^*,\tau}^\mu(s) + V_{n^*,\tau}^\mu(s) - V_\tau^*(s)\\ &\le \gamma^{n^*(s)-1}\gamma_\tau\|V - V_{n^*,\tau}^\mu\|_\infty + \|V_{n^*,\tau}^\mu - V_\tau^*\|.\end{aligned} \tag{28}$$

□

B.7 PROOF OF PROPOSITION 1

Proposition 1. *Let V_τ^* denote the fixed point of $\mathcal{T}_\tau^\mu$. For any $\tau, \tau' \in (0,1)$, if $\tau' \ge \tau$, we have $V_{\tau'}^*(s) \ge V_\tau^*(s), \forall s \in S$.*

Proof. With the Lemma 2, we have $\mathcal{T}_{\tau'}^\mu V_\tau^* \ge \mathcal{T}_\tau^\mu V_\tau^*$. Since V_τ^* is the fixed point of $\mathcal{T}_\tau^\mu$, we have $\mathcal{T}_\tau^\mu V_\tau^* = V_\tau^*$. Putting the results together, we obtain $V_\tau^* = \mathcal{T}_\tau^\mu V_\tau^* \le \mathcal{T}_{\tau'}^\mu V_\tau^*$. Repeatedly applying $\mathcal{T}_{\tau'}^\mu$ and using its monotonicity, we have $V_\tau^* \le \mathcal{T}_{\tau'}^\mu V_\tau^* \le (\mathcal{T}_{\tau'}^\mu)^\infty V_\tau^* = V_{\tau'}^*$. □

C DETAILED IMPLEMENTATION

C.1 GENERALIZED ADVANTAGE-WEIGHTED LEARNING

In practice, we adopt Leaky-ReLU or Softmax functions.

Leaky-ReLU:

$$\begin{aligned}&\max_\phi J_\pi(\phi) = \mathbb{E}_{(s,a)\sim\mathcal{D}}\left[\log\pi_\phi(a \mid s)\cdot f\left(\hat{A}(s,a)\right)\right],\\ &\text{where}\quad f(\hat{A}(s,a)) = \begin{cases}\hat{A}(s,a) & \text{if}\quad \hat{A}(s,a) > 0\\ \frac{\hat{A}(s,a)}{\alpha} & \text{if}\quad \hat{A}(s,a) \le 0\end{cases}\end{aligned} \tag{29}$$

Softmax:

$$\max_{\phi} J_{\pi}(\phi) = \mathbb{E}_{(s,a)\sim\mathcal{D}}\left[\log \pi_{\phi}(a \mid s) \cdot \frac{\exp(\frac{1}{\alpha}\hat{A}(s,a))}{\sum_{(s_i,a_i)\sim\text{Batch}} \exp(\frac{1}{\alpha}\hat{A}(s_i,a_i))}\right]. \tag{30}$$

C.2 BCQ-EM

The value network of BCQ-EM is trained by minimizing the following loss:

$$\min_{\theta} \mathcal{J}_Q(\theta) = \mathbb{E}_{(s_t,a_t,s_{t+1})\sim\mathcal{D}}\left[\left(R_t - Q_{\theta}(s_t,a_t)\right)^2\right] \tag{31}$$

$$R_t = \max_{0<n\leq n_{\max}} Q_{t,n}, \quad Q_{t,n} = \begin{cases} r_t + \gamma Q_{t+1,n-1}(s_{t+1},\hat{a}_{t+1}) & \text{if} \quad n > 0, \\ Q(s_t,\hat{a}_t) & \text{if} \quad n = 0, \end{cases} \tag{32}$$

where $\hat{a}_t$ corresponds to the perturbed actions, sampled from the generative model $G_w(s_t)$.

The perturbation network of BCQ-EM is trained by minimizing the following loss:

$$\min_{\phi} \mathcal{J}_{\xi}(\phi) = -\mathbb{E}_{s\sim\mathcal{D}}\left[Q_{\theta}(s, a_i + \xi_{\phi}(s,a_i,\Phi))\right], \quad \{a_i \sim G_w(s)\}_{i=1}^{n}, \tag{33}$$

where $\xi_{\phi}(s,a_i,\Phi)$ is a perturbation model, which outputs an adjustment to an action a in the range $[-\Phi,\Phi]$. We adopt conditional variational auto-encoder to represent the generative model $G_w(s)$ and it is trained to match the state-action pairs sampled from $\mathcal{D}$ by minimizing the cross-entropy loss-function.

C.3 Hyper-Parameter and Network Structure

Table 2: Hyper-parameter Sheet

Hyper-Parameter	Value
Critic Learning Rate	1e-3
Actor Learning Rate	1e-3
Optimizer	Adam
Target Update Rate (κ)	0.005
Memory Update Period	100
Batch Size	128
Discount Factor	0.99
Gradient Steps per Update	200
Maximum Length d	Episode Length T

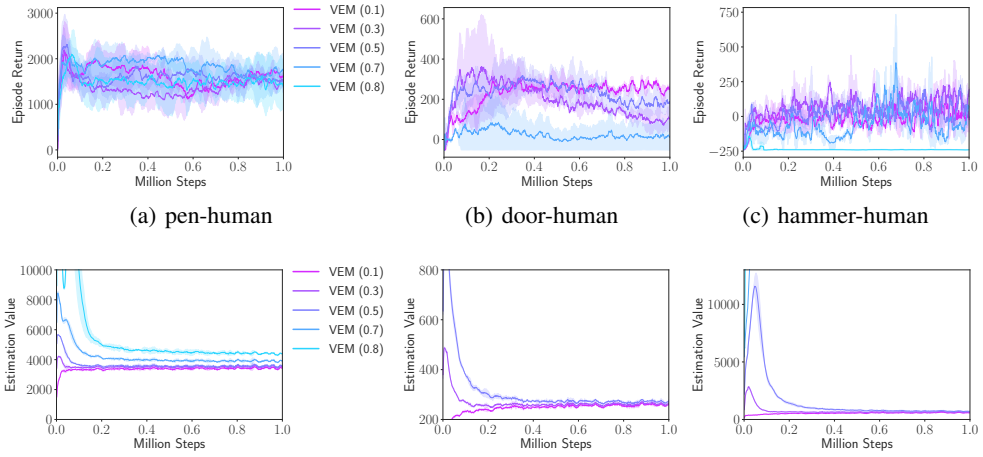
Table 3: Hyper-Parameter τ used in VEM across different tasks

AntMaze-fixed	umaze 0.4	medium 0.3	large 0.3
AntMaze-diverse	umaze 0.3	medium 0.4	large 0.1
Adroit-human	door 0.4	hammer 0.4	pen 0.4
Adroit-cloned	door 0.2	hammer 0.3	pen 0.1
Adroit-expert	door 0.3	hammer 0.3	pen 0.3
MuJoCo-medium	walker2d 0.3	halfcheetah 0.4	hopper 0.5
MuJoCo-random	walker2d 0.5	halfcheetah 0.6	hopper 0.7

We use a fully connected neural network as a function approximation with 256 hidden units and ReLU as an activation function. The structure of the actor network is $[(\text{state dim}, 256), (256, 256), (256, \text{action dim})]$. The structure of the value network is $[(\text{state dim}, 256), (256, 256), (256, 1)]$.

D ADDITIONAL EXPERIMENTS ON D4RL

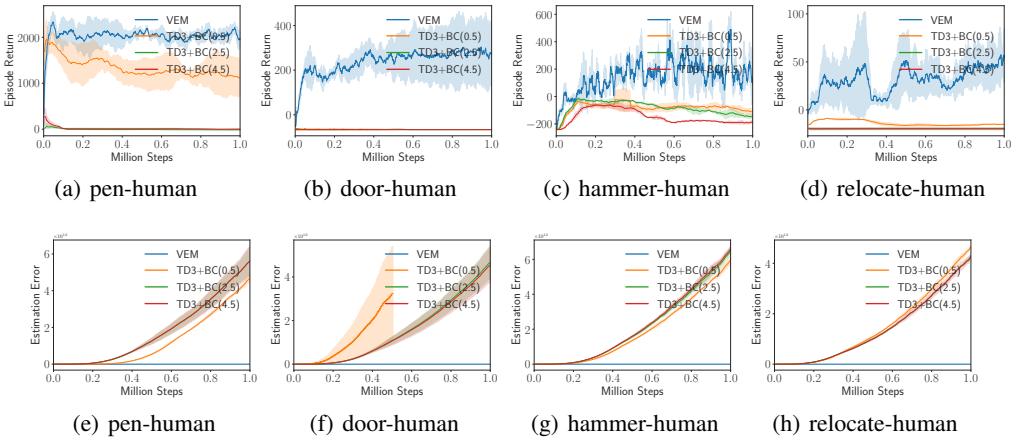
D.1 ABLATION STUDY



(a) pen-human (b) door-human (c) hammer-human

(d) pen-human (e) door-human (f) hammer-human

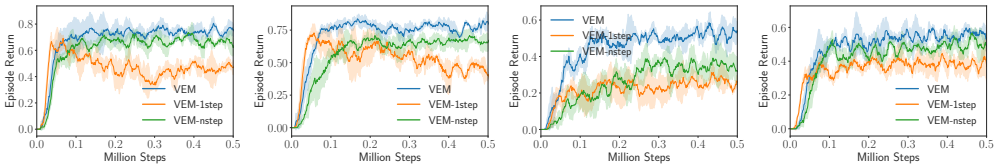
Figure 5: The results of VEM (τ) with various τ in Adroit tasks. The results in the upper row are the performance. The results in the bottom row are the estimation value.



(a) pen-human (b) door-human (c) hammer-human (d) relocate-human

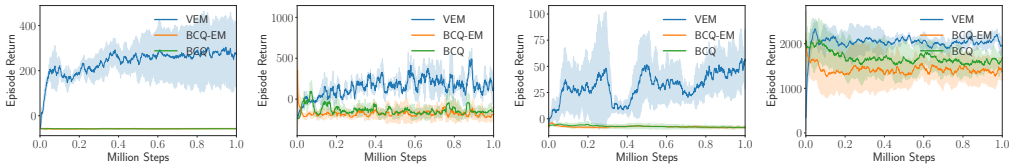
(e) pen-human (f) door-human (g) hammer-human (h) relocate-human

Figure 6: Comparison results between VEM with TD3+BC. We adopt different hyper-parameters $\alpha \in \{0.5, 2.5, 4.5\}$ in TD3+BC to test its performance. The upper row are the performance. The results in the bottom row are the estimation error (the unit is 10^{12}).

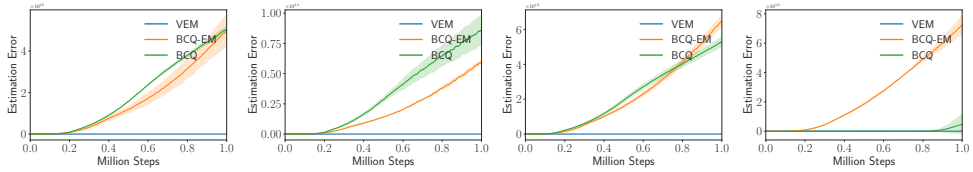


(a) medium-play (b) medium-diverse (c) large-play (d) large-diverse

Figure 7: The comparison between episodic memory and n-step value estimation on AntMaze tasks.

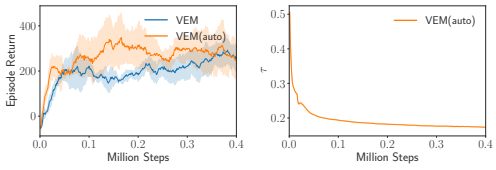


(a) door-human (b) hammer-human (c) relocate-human (d) pen-human



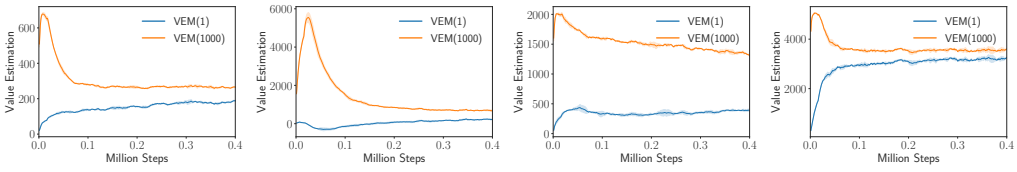
(e) door-human (f) hammer-human (g) relocate-human (h) pen-human

Figure 8: The comparison between VEM, BCQ-EM and BCQ on Adroit-human tasks. The results in the upper row are the performance. The results in the bottom row are the estimation error, where the unit is 10^{13}.



(a) Episode return (b) τ value

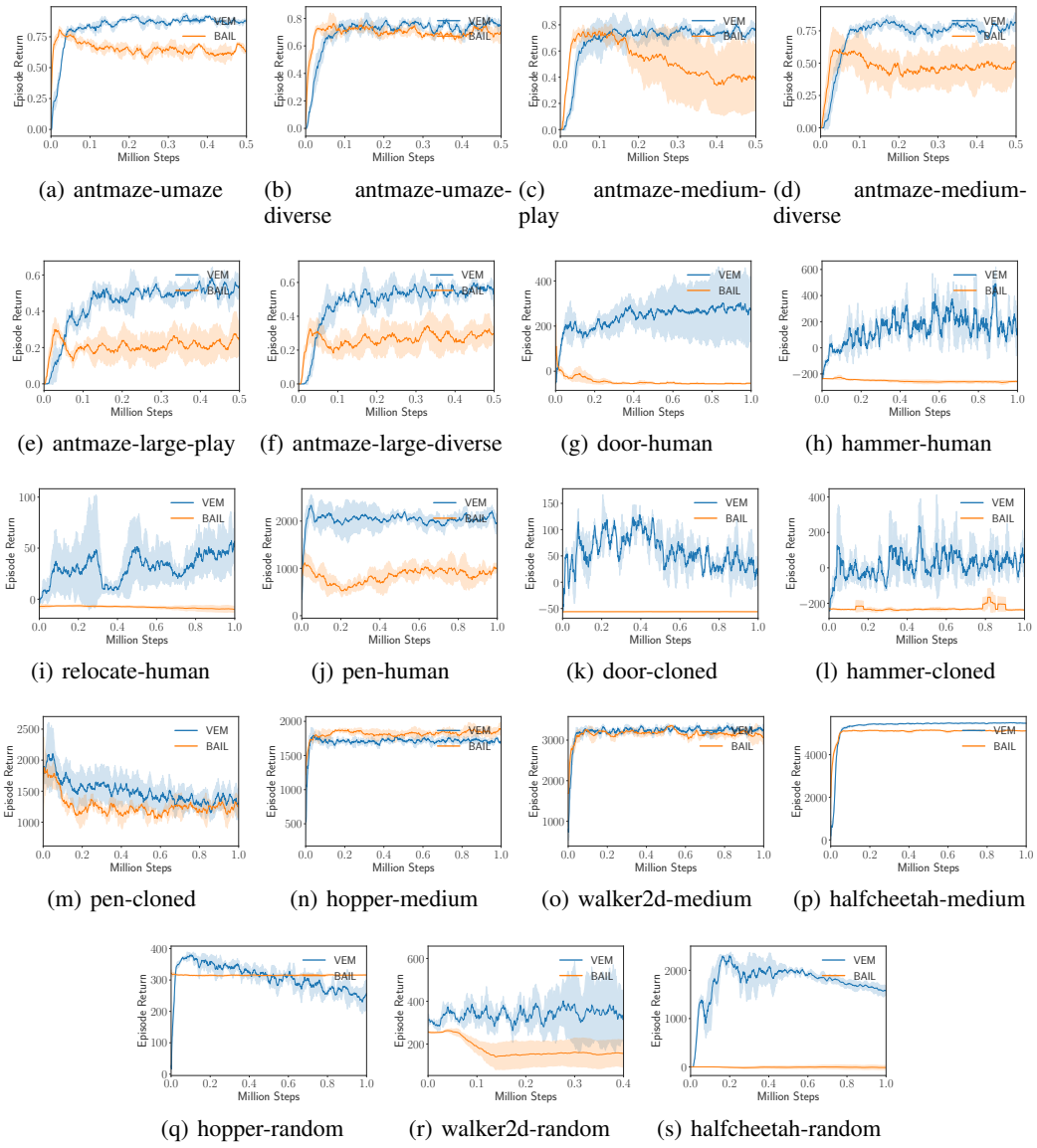
Figure 9: Comparison between fixed τ (VEM) and auto-tuning τ (VEM(auto)) in the door-human task.



(a) door-human (b) hammer-human (c) relocate-human (d) pen-human

Figure 10: Value estimation of VEM ($n_{\max}$) in adroit-human tasks, where $n_{\max}$ is the maximal rollout step for memory control (see Equation 11). We set $\tau = 0.5$ in all tasks.

D.2 COMPLETE TRAINING CURVES AND VALUE ESTIMATION ERROR



(a) antmaze-umaze (b) antmaze-umaze-diverse (c) antmaze-medium-play (d) antmaze-medium-diverse

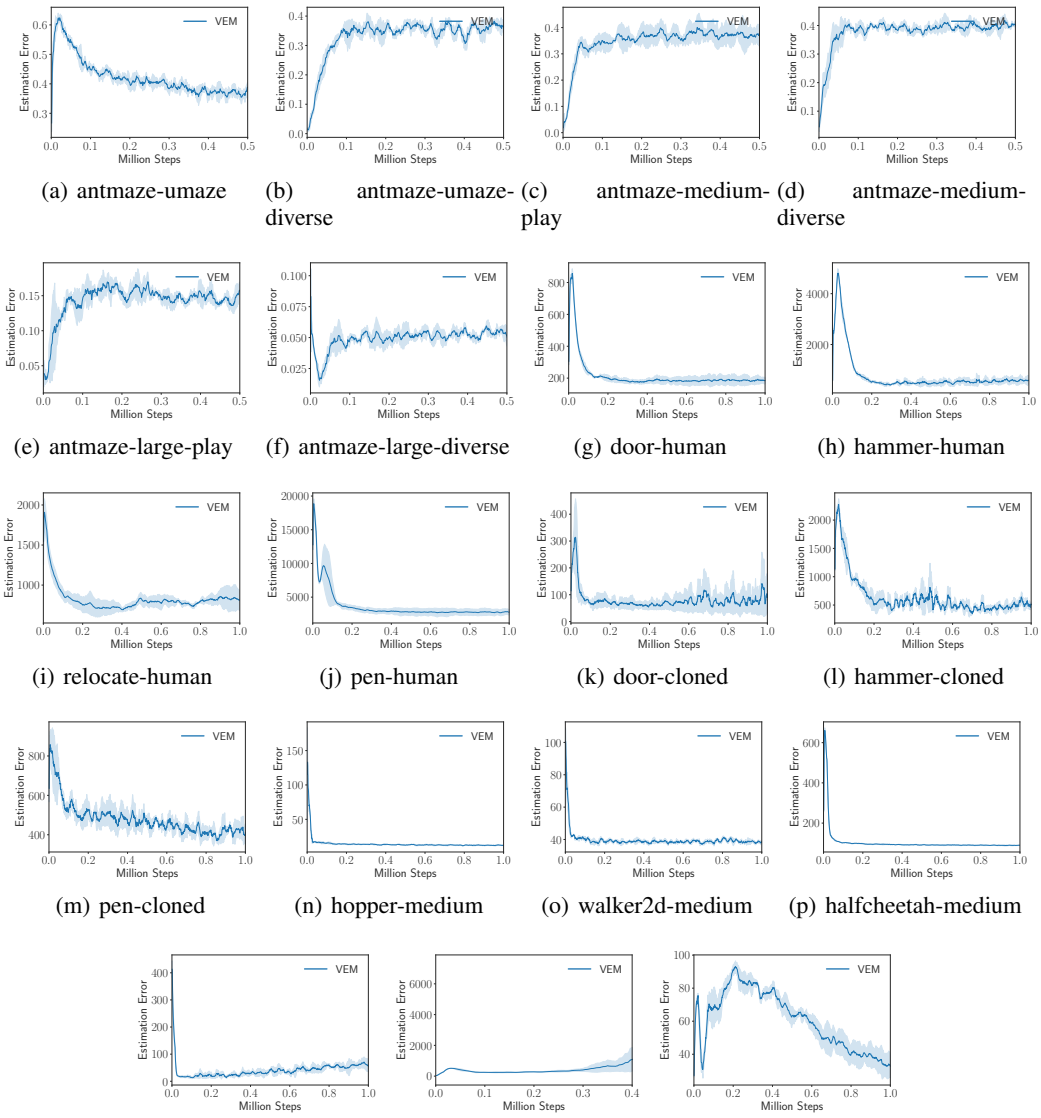
(e) antmaze-large-play (f) antmaze-large-diverse (g) door-human (h) hammer-human

(i) relocate-human (j) pen-human (k) door-cloned (l) hammer-cloned

(m) pen-cloned (n) hopper-medium (o) walker2d-medium (p) halfcheetah-medium

(q) hopper-random (r) walker2d-random (s) halfcheetah-random

Figure 11: The training curves of VEM and BAIL on D4RL tasks.



(a) antmaze-umaze (b) antmaze-umaze-diverse (c) antmaze-medium-play (d) antmaze-medium-diverse

(e) antmaze-large-play (f) antmaze-large-diverse (g) door-human (h) hammer-human

(i) relocate-human (j) pen-human (k) door-cloned (l) hammer-cloned

(m) pen-cloned (n) hopper-medium (o) walker2d-medium (p) halfcheetah-medium

(q) hopper-random (r) walker2d-random (s) halfcheetah-random

Figure 12: The value estimation error of VEM on D4RL tasks. The estimation error refers to the average estimated state values minus the average returns.