Supplementary Material

A. GEM Algorithm in Tabular Case

In this section, we present the formal description of the GEM algorithm in tabular case, as shown in Algorithm 3.

Algorithm 3 Generalizable Episodic Memory in Tabular Case

```
Initialize table $Q^{(1)}(s,a), Q^{(2)}(s,a)$ arbitrarily,
Initial learning step size $\alpha_t$, small $\epsilon > 0$ and episode length $l = 0$
Set $\pi$ to be the $\epsilon$-greedy policy with respect to $Q^{(1)}(s,a)$ or $Q^{(2)}(s,a)$
for $t = 1, \cdots,$ do
  Initialize and store $s_0$
  Select action $a_0 \sim \pi(\cdot|s_0)$
  Observe reward $r$ and new state $s'$
  Store transition tuple $(s, a, r, s')$
  $l \leftarrow l + 1$
  if an episode is ended then
    for $\tau = t - l, \cdots, t$ do
      Compute $R_\tau^{(1)}, R_\tau^{(2)}$ according to Equation (7)
      Uniformly choose $i \in \{1, 2\}$
      Update $Q^{(i)}(s_\tau, a_\tau) \leftarrow Q^{(i)}(s_\tau, a_\tau) + \alpha_\tau(R_\tau^{(i)} - Q^{(i)}(s_\tau, a_\tau))$
    end for
    Set $\pi$ to be the $\epsilon$-greedy policy with respect to $Q^{(1)}(s,a)$ or $Q^{(2)}(s,a)$
    $l \leftarrow 0$
  end if
end for
```

B. Proofs of Theoremss

Theorem 1. Given unbiased and independent estimators $Q_\theta^{(1,2)}(s_{t+h}, a_{t+h}) = Q^\pi(s_{t+h}, a_{t+h}) + \epsilon_h^{(1,2)}$, Equation (8) will not overestimate the true objective, i.e.

$$\mathbb{E}_{\tau,\epsilon}\left[R_t^{(1,2)}(s_t)\right] \le \mathbb{E}_\tau\left[\max_{0\le h\le T-t-1} Q_{t,h}^\pi(s_t)\right], \tag{12}$$

where

$$Q_{t,h}^\pi(s,a) = \begin{cases} \sum_{i=0}^{h} \gamma^i r_{t+i} + \gamma^{h+1} Q^\pi(s_{t+h+1}, a_{t+h+1}) & \text{if } h < T - t, \\ \sum_{i=0}^{h} \gamma^i r_{t+i} & \text{if } h = T - t. \end{cases} \tag{13}$$

and $\tau = \{(s_t, a_t, r_t, s_{t+1})_{t=1,\cdots,T}\}$ is a trajectory.

Proof. By unrolling and rewritten Equation (7), we have

$$R_t^{(1,2)} = V_{t,h^*} = \sum_{i=0}^{h^*} \gamma^i r_{t+i} + \gamma^{h^*+1} Q_\theta^{(2,1)}(s_{t+h^*+1}, a_{t+h^*+1}),$$

Where h^* is the abbrevation for $h^*_{(1,2)}$ for simplicity. Then we have

$$\begin{aligned}\mathbb{E}_\epsilon\left[R_t^{(1,2)} - Q^\pi_{t,h^*_{(1,2)}}(s_t)\right] &= \mathbb{E}\left[V_{t,h^*_{(1,2)}} - Q^\pi_{t,h^*_{(1,2)}}(s_t)\right] \\ &= \mathbb{E}\left[\gamma^{h^*+1}\left(Q_\theta^{(2,1)}(s_{t+h^*+1}, a_{t+h^*+1}) - Q^\pi(s_{t+h^*+1}, a_{t+h^*+1})\right)\right] \\ &= 0.\end{aligned}$$

Then naturally

$$\mathbb{E}_{\tau,\epsilon}[R_t^{(1,2)}] = \mathbb{E}_\tau[Q^\pi_{t,h^*_{(1,2)}}(s_t)] \le \mathbb{E}_\tau\left[\max_{0\le h\le T-t} Q^\pi_{t,h}(s_t)\right].$$

□

To prepare for the theorem below, we need the following lemma:

Lemma 1. Consider a stochastic process $(\zeta_t, \Delta_t, F_t), t \ge 0$, where $\zeta, \Delta_t, F_t : X \to \mathbb{R}$ satisfy the equations

$$\Delta_{t+1}(x) = (1 - \zeta_t(x))\Delta_t(x) + \zeta_t(x)F_t(x) \tag{14}$$

Let $\{P_t\}$ be a filter such that ζ_t and Δ_t are P_t-measurable, F_t is P_{t+1}-measurable, $t \ge 0$. Assume that the following hold:

- X is finite: $|X| < +\infty$.
- $\zeta_t(x) \in [0, 1]$, $\sum_t \zeta_t(x) = +\infty$, $\sum_t \zeta_t^2(x) < +\infty$ a.s. for all $x \in X$.
- $\|\mathbb{E}(F_t|P_t)\|_\infty \le \kappa\|\Delta_t\|_\infty + c_t$, where $\kappa \in [0, 1)$ and $c_t \overset{\text{a.s}}{\to} 0$.
- Var$(F_t|P_t) \le K(1 + \|\Delta_t\|_\infty)^2$, where K is some constant.

Then Δ_t converge to zero w.p.1.

This lemma is also used in Double-Q learning (Van Hasselt, 2010) and we omit the proof for simplicity.
In the following sections, we use $\|\cdot\|$ to represent the infinity norm $\|\cdot\|_\infty$.

Theorem 2. Algorithm 3 converge to Q^* w.p.1 with the following conditions:

1. The MDP is finite, i.e. $|\mathcal{S} \times \mathcal{A}| \le \infty$
2. $\gamma \in [0, 1)$
3. The Q-values are stored in a lookup table
4. $\alpha_t(s, a) \in [0, 1], \sum_t \alpha_t(s, a) = \infty, \sum_t \alpha_t^2(s, a) \le \infty$
5. The environment is fully deterministic, i.e. $P(s'|s, a) = \delta(s' = f(s, a))$ for some deterministic transition function f

Proof. This is a sketch of proof and some technical details are omitted.

We just need to show that without double-q version, the update will be a γ-contraction and will converge. Then we need to show that $\|Q^1 - Q^2\|$ converge to zero, which is similar with double-q learning.

We only prove convergence of $Q^{(1)}$, and by symmetry we have the conclusion.

Let $\Delta_t = Q_t^{(1)} - Q^*$, and $F_t(s_t, a_t) = R_t^{(1)} - Q^*(s_t, a_t)$,

Then the update rule can be written exactly as Equation (14):

$$\Delta_{t+1} = (1 - \alpha_t)\Delta_t + \alpha_t F_t.$$

We define

$$G_t = \tilde{R}_t^{(1)} - Q^*(s_t, a_t) = F_t + (\tilde{R}_t^{(1)} - R_t^{(1)}),$$

where $\tilde{R}^{(1)} = R_{t,h^*_{(1)}}^{(1)}$,and the notation is kept the same as in Equation (7)(8).

To use Lemma 1, we only need to prove that G_t is a γ-contraction and $c_t = \tilde{R}_t^{(1)} - R_t^{(1)}$ converge to zero.

On the one hand,

$$\begin{aligned}
\tilde{R}_t^{(1)} - Q^*(s_t, a_t) &\geq r_t + \gamma \tilde{Q}^{(1)}(s_{t+1}, \tilde{a}^*) - Q^*(s_t, a_t) \\
&= r_t + \gamma \tilde{Q}^{(1)}(s_{t+1}, \tilde{a}^*) - r_t + \gamma Q^*(s_{t+1}, a^*) \\
&= \gamma(\tilde{Q}^{(1)}(s_{t+1}, \tilde{a}^*) - Q^*(s_{t+1}, a^*)) \\
&\geq -\gamma \left\| \Delta_t \right\|.
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\tilde{R}_t^{(1)} - Q^*(s_t, a_t) &= \sum_{i=0}^{h^*_{(1)}} \gamma^i r_{t+i} + \gamma^{h^*_{(1)}+1} \tilde{Q}^{(1)}(s_{t+h^*_{(1)}+1}, \tilde{a}^*) - Q^*(s_t, a_t) \\
&\leq \sum_{i=0}^{h^*_{(1)}} \gamma^i r_{t+i} + \gamma^{h^*_{(1)}+1} \tilde{Q}_\pi^{(1)}(s_{t+h^*_{(1)}+1}) \\
&- (\sum_{i=0}^{h^*_{(1)}} \gamma^i r_{t+i} + \gamma^{h^*_{(1)}+1} Q^*(s_{t+h^*_{(1)}+1}, a^*)) \\
&= \gamma^{h^*_{(1)}+1}(\tilde{Q}^{(1)}(s_{t+h^*_{(1)}+1}, \tilde{a}^*) - Q^*(s_{t+h^*_{(1)}+1}, a^*)) \\
&\leq \gamma(\tilde{Q}^{(1)}(s_{t+h^*_{(1)}+1}, \tilde{a}^*) - Q^*(s_{t+h^*_{(1)}+1}, a^*)) \\
&\leq \gamma \left\| \Delta_t \right\|.
\end{aligned}$$

Thus G_t is a γ-contraction w.r.t Δ_t.

Finally we show $c_t = \tilde{R}^{(1)} - R_t^{(1)}$ converges to zero.

Note that $c_t = \gamma^{h^*_{(1)}}(\tilde{Q}^{(1)} - \tilde{Q}^{(2)})$, it suffices to show that $\Delta^{1,2} = \tilde{Q}^{(1)} - \tilde{Q}^{(2)}$ converge to zero.

Depending on whether $\tilde{Q}^{(1)}$ or $\tilde{Q}^{(2)}$ is updated, the update rule can be written as

$$\Delta_{t+1}^{1,2} = \Delta_t^{1,2} + \alpha_t F_t^{(2)}(s_t, a_t),$$

or

$$\Delta_{t+1}^{1,2} = \Delta_t^{1,2} - \alpha_t F_t^{(1)}(s_t, a_t),$$

where $F_t^{(1)} = R_t^{(1)} - \tilde{Q}_t^{(2)}$ and $F_t^{(2)} = R_t^{(2)} - \tilde{Q}_t^{(1)}$.

Now let $\zeta_t = \frac{1}{2}\alpha_t$, we have

$$\begin{aligned}\mathbb{E}[\Delta^{1,2}_{t+1}|P_t] &= \frac{1}{2}(\Delta^{1,2} + \alpha_t \,\mathbb{E}[F^{(2)}_t]) + \frac{1}{2}(\Delta^{1,2} - \alpha_t \,\mathbb{E}[F^{(1)}_t]) \\ &= (1-\zeta_t)\Delta^{1,2}_t + \zeta_t \,\mathbb{E}[R^{(2)}_t - R^{(1)}_t]\end{aligned}$$

when $\mathbb{E}[R^{(2)}_t] \geq \mathbb{E}[R^{(1)}_t]$, by definition we have $\mathbb{E}[R^{(2)}_t] \leq \mathbb{E}[\tilde{R}^{(2)}_t]$.

Then

$$\begin{aligned}|\,\mathbb{E}[R^{(2)}_t - R^{(1)}_t]| &\leq \mathbb{E}[\tilde{R}^{(2)}_t - R^{(1)}_t] \\ &\leq \gamma^{h^*_{(2)}+1}(Q^{(1)}(s_{t+h^*_{(2)}+1}, a^*_{(1)}) - Q^{(2)}(s_{t+h^*_{(2)}+1}, a^*_{(1)})) \\ &\leq \gamma \left\|\Delta^{1,2}_t\right\|.\end{aligned}$$

Similarly, $\mathbb{E}[R^{(2)}_t] < \mathbb{E}[R^{(1)}_t]$,, we have

$$\begin{aligned}|\,\mathbb{E}[R^{(2)}_t - R^{(1)}_t]| &\leq \mathbb{E}[\tilde{R}^{(1)}_t - R^{(2)}_t] \\ &\leq \gamma^{h^*_{(1)}+1}(Q^{(2)}(s_{t+h^*_{(1)}+1}, a^*_{(2)}) - Q^{(1)}(s_{t+h^*_{(1)}+1}, a^*_{(2)})) \\ &\leq \gamma \left\|\Delta^{1,2}_t\right\|.\end{aligned}$$

Now in both scenairos we have $|E\{F^{(1,2)}_t|P_t\}| \leq \gamma \left\|\Delta^{1,2}_t\right\|$ holds. Applying Lemma 1 again we have the desired results. □

The theorem apply only to deterministic scenairos. Nevertheless, we can still bound the performance when the environment is stochastic but nearly deterministic.

Theorem 3. $\tilde{Q}(s,a)$ learned by Algorithm 3 satisfy the following inequality:

$$\forall s \in \mathcal{S}, a \in \mathcal{A}, Q^*(s,a) \leq \tilde{Q}(s,a) \leq Q_{\text{max}}(s,a), \tag{15}$$

w.p.1 with condition 1-4 in Theorem 2.

Proof. We just need to prove that $(Q^* - Q^{(1,2)})_+$ and $(Q^{(1,2)} - Q_{\text{max}})_+$ converge to 0 w.p.1, where $(\cdot)_+ = \max(0,\cdot)$.

On the one hand, similar from the proof of Theorem 2 and let $\Delta_t = (Q^*(s_t,a_t) - Q^{(1,2)}(s_t,a_t))_+$.

$$\begin{aligned}Q^*(s_t,a_t) - \tilde{R}^{(1,2)}_t &\leq Q^*(s_t,a_t) - (r_t + \gamma\tilde{Q}^{(1,2)}(s_{t+1},\tilde{a}^*)) \\ &= r_t + \gamma Q^*(s_{t+1},a^*) - r_t - \gamma\tilde{Q}^{(1,2)}(s_{t+1},\tilde{a}^*) \\ &= \gamma(\tilde{Q}(s_{t+1},\tilde{a}^*) - Q^*(s_{t+1},a^*)) \\ &\leq \gamma \left\|\Delta_t\right\|.\end{aligned}$$

The rest is the same as the proof of Theorem 2, and we have $(Q^* - Q^{(1,2)})_+$ converge to zero w.p.1.

On the other hand, let $\Delta_t = (Q_{\text{max}}(s_t,a_t) - Q^{(1,2)}(s_t,a_t))_+$,

We have

$$\begin{aligned}
F_{t+1} &= \tilde{R}_t^{(1,2)} - Q_t^{max} \\
&\leq \sum_{i=0}^{h^*_{(2,1)}} \gamma^i r_{t+i} + \gamma^{h^*+1} \tilde{Q}^{(1,2)}_{t+h^*_{(2,1)+1}} - (\sum_{i=0}^{h^*_{(2,1)}} \gamma^i r_{t+i} + \gamma^{h^*+1} Q^{max}_{t+h^*_{(2,1)+1}}) \\
&\leq \gamma^{h^*+1} (\tilde{Q}^{(1,2)}_{t+h^*_{(2,1)+1}} - Q^{max}_{t+h^*_{(2,1)+1}}) \\
&\leq \gamma \left\| \Delta_t \right\| .
\end{aligned}$$

The rest is the same as the proof of Theorem 2, and we have $(Q^{(1,2)} - Q_{\text{max}})_+$ converge to zero w.p.1.

□

When the enironment is nearly-deterministic, we can bound the performance of Q despite its non-convergence:

Theorem 4. For a nearly-deterministic environment with factor μ, in limit, GEM's performance can be bounded by

$$V^{\tilde{\pi}}(s) \geq V^*(s) - \frac{2\mu}{1-\gamma}, \forall s \in \mathcal{S}. \tag{16}$$

Proof. since we have $\left\| \tilde{Q} - Q^* \right\| \leq \mu$, It is easy to show that

$$\begin{aligned}
& V^*(s) - V_{\tilde{\pi}}(s) \\
&= Q^*(s, a^*) - Q_{\tilde{\pi}}(s, \tilde{a}) \\
&= Q^*(s, a^*) - \tilde{Q}(s, a^*) + \tilde{Q}(s, a^*) - Q_{\tilde{\pi}}(s, \tilde{a}) \\
&\leq \epsilon + \tilde{Q}(s, \tilde{a}) - Q_{\tilde{\pi}}(s, \tilde{a}) \\
&= \epsilon + (\tilde{Q}(s, \tilde{a}) - Q^*(s, \tilde{a})) + (Q^*(s, \tilde{a}) - Q_{\tilde{\pi}}(s, \tilde{a})) \\
&\leq 2\epsilon + \gamma(V^*(s) - V_{\tilde{\pi}}(s)).
\end{aligned}$$

So we have the conclusion. □

C. Hyperparameters

Here we listed the hyperparameters we used for the evaluation of our algorithm.

Task	HalfCheetah	Ant	Swimmer	Humanoid	Walker	Hopper
Maximum Length d	1000	1000	1000	5	5	5

Table 1. Maximum length of rollouts used in GEM across different tasks

Hyper-parameter	GEM
Critic Learning Rate	1e-3
Actor Learning Rate	1e-3
Optimizer	Adam
Target Update Rate(τ)	0.6
Memory Update Period(u)	100
Memory Size	100000
Policy Delay(p)	2
Batch Size	100
Discount Factor	0.99
Exploration Policy	$\mathcal{N}(0, 0.1)$
Gradient Steps per Update	200

Table 2. List of Hyperparameters used in GEM across different tasks

The hyper-parameters for Atari games are kept the same as in the continuous domain, and other hyper-parameters are kept the same as Rainbow (Hessel et al., 2018).

D. Additional Ablation Results

Here we include more ablation results of GEM. To verify the effectiveness of our proposed implicit planning, we compare our method with simple n-step Q learning combined with TD3. For a fair comparison, we include all different rollout lengths used in GEM's result. The result is shown in Figure 6. We can see that GEM significantly outperform simple n-step learning.

To understand the effects of rollout lengths, we also compare the result of different rollout lengths on Atari games. The result is shown below in Figure 7. We can see that using short rollout length greatly hinders the performance of GEM.

To verify the effectiveness of GEM on the stochastic domain, we conduct experiments on Atari games with sticky actions, as suggested in (Machado et al., 2018). As illustrated in Figure 5, GEM is still competitive on stochastic domains.

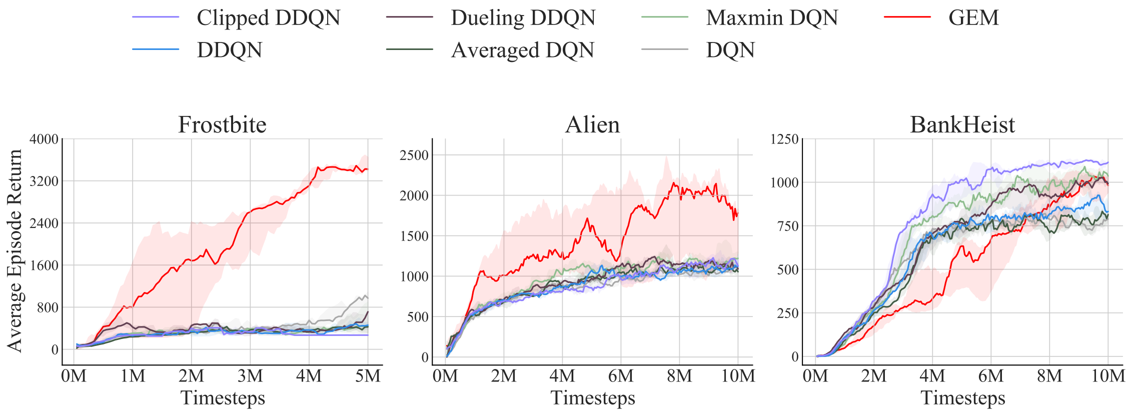

Figure 5. Comparison on 3 Atari games, with sticky actions to make the environment stochastic.

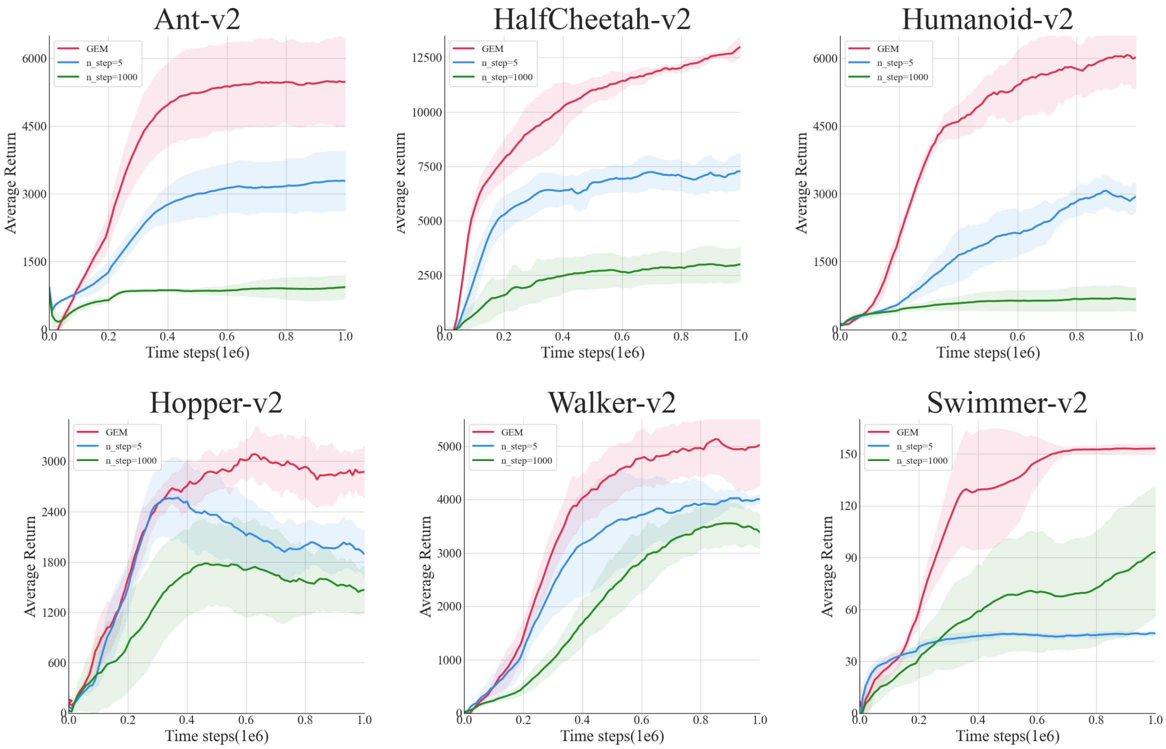

Figure 6. Comparison with simple n-step learning. The shaded region represents half a standard deviation of the average evaluation. Curves are smoothed uniformly for visual clarity.

References

Hessel, M., Modayil, J., Van Hasselt, H., Schaul, T., Ostrovski, G., Dabney, W., Horgan, D., Piot, B., Azar, M., and Silver, D. Rainbow: Combining improvements in deep reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Machado, M. C., Bellemare, M. G., Talvitie, E., Veness, J., Hausknecht, M., and Bowling, M. Revisiting the arcade learning environment: Evaluation protocols and open problems for general agents. *Journal of Artificial Intelligence Research*, 61: 523–562, 2018.

Van Hasselt, H. Double q-learning. In *Advances in neural information processing systems*, pp. 2613–2621, 2010.

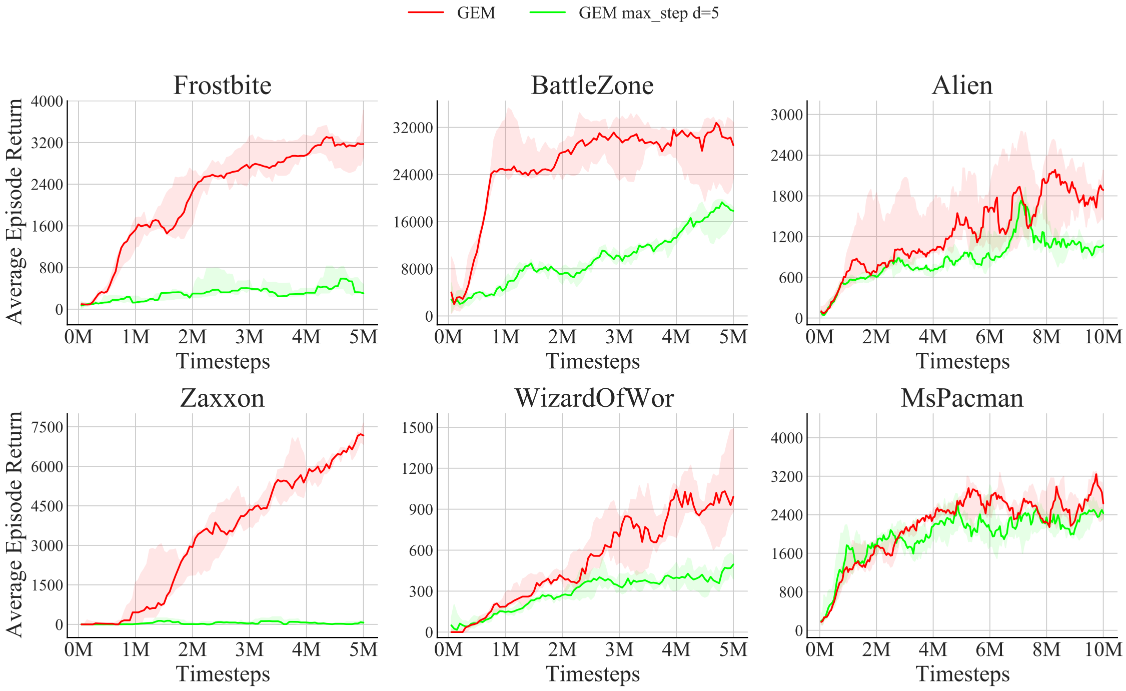

Figure 7. Ablation study on 6 Atari games. Limiting rollout lengths greatly affects the performance of GEM, which proves that GEM can use long rollout trajectories effectively.