Appendix

Sparse Reward and Representation Collapse

In this section, we illustrate how the sparse rewards affect the representations of states.

Definition 2. Let m^π be the stationary distribution over the states, and k^π the distribution over pair of states. The states s_i and s_j are independently sampled from m^π. The measurement of reward is:

$$\rho_r^\pi = \mathbb{E}_{(s_i,s_j)\sim k^\pi}\left[|r_i^\pi - r_j^\pi|\right]. \tag{20}$$

When under a sparse reward environment, i.e., $\rho^\pi \approx 0$, this indicates that the reward under the policy π is uninformative. The metric of two states is:

$$\rho_d^\pi = \mathbb{E}_{(s_i,s_j)\sim k^\pi}\left[d_\pi(s_i,s_j)\right]. \tag{21}$$

From Kemertas and Aumentado-Armstrong (2021), the relation between ρ_r^π and ρ_d^π is concluded in the following:

$$\rho_d^\pi = \frac{c_R}{1-c_T}\rho_r^\pi, \tag{22}$$

where c_R and c_T are the weights for reward and transition function in Eq. 1. To further analyze, when ρ_r^π is approximated as zero, ρ_d^π will also approach zero which indicates that the representations of states collapse.

Alleviating Representation Collapse

Recalling and expanding our objective,

$$\begin{aligned}\ell(\phi) = \frac{1}{2}\mathbb{E}\Bigg[&\underbrace{d\big(\phi(s_i),\phi(s_j)\big)}_{\text{our behavioral metric}} - \underbrace{d\big(\psi(s_i,a_i),\psi(s_j,a_j)\big)}_{\text{cumulative effect measurement}} \\ &- \underbrace{d\left(f\left(\phi(s_i),a_i\right),f\left(\phi(s_j),a_j\right)\right)}_{\text{immediate effect measurement}}\Bigg]^2 \\ = \frac{1}{2}\mathbb{E}\Bigg[&\Big(\underbrace{d\big(\phi(s_i),\phi(s_j)\big)}_{\text{our behavioral metric}} - 1\Big) - \underbrace{d\big(\psi(s_i,a_i),\psi(s_j,a_j)\big)}_{\text{cumulative effect measurement}} \\ &- \Big(\underbrace{d\left(f\left(\phi(s_i),a_i\right),f\left(\phi(s_j),a_j\right)\right)}_{\text{immediate effect measurement}} - 1\Big)\Bigg]^2.\end{aligned} \tag{23}$$

From Proposition 1 in Castro et al. (2021), we have:

$$|V^\pi(s_i) - V^\pi(s_j)| \le d^\pi(s_i,s_j), \tag{24}$$

for $\forall s_i, s_j \in \mathcal{S}$ and for any policy π. When cosine similarity is used in our behavioral metric and immediate effect $\phi(s_i) = \phi(s_j)$ happens, the following is ensured:

$$|V(s_i) - V(s_j)| \le d_{cos}\left(\phi(s_i),\phi(s_j)\right) - 1 = 0, \tag{25}$$

where d_{cos} is the cosine distance. This implies that their value functions should also be identical, thereby suggesting that our behavioral metric can mitigate the issue of representation collapse to a certain extent.

Hyperparameters

Table 2 lists the hyperparameters used in our BCD.

Hyperparameter	Value
Number of parallel environments	128
Number of time steps of each rollout	128
PPO clip range	0.2
Number of epochs	4
Coefficient of value loss	0.5
Coefficient of entropy bonus	0.01
Learning rate	5×10^{-4}

Table 2: Hyerparameters.

Method	ObstructedMaze-2Q	KeyCorridor-S6R3
w/o inverse	0.77 ± 0.245	0.94 ± 0.039
w/o backward	0.74 ± 0.288	0.93 ± 0.021
w/o inverse and backward	0.74 ± 0.354	0.92 ± 0.047
l2 distance on $\psi(s,a)$	0.72 ± 0.323	0.90 ± 0.049
l2 distance on z^k	0.73 ± 0.275	0.90 ± 0.030
BCD	0.78 ± 0.182	0.94 ± 0.012

Table 3: Ablation studies on cyclic dynamics and effect of distance measurement in our behavioral metric.

Ablation Studies and Effect of Distance Measurement

In Table 3, we report an additional ablation study, i.e., without backward and inverse dynamics learning. The results demonstrate the necessity of our cyclic dynamics framework, highlighting the significant roles played by inverse and backward dynamics. Additionally, we substitute the Euclidean distance for the cosine similarity in the computation of distances involving the successor feature $\psi(s,a)$ and discrete variable z^k within our forward dynamics learning. The observed performance deterioration demonstrates the effectiveness of the cosine distance employed in the proposed BCD framework. The merit of employing cosine similarity lies in its ability to normalize embeddings to unit length, enhancing robustness against varying feature scales.

More Results with Baselines

As shown in Figure 6, we plot more results of different environments (ObstructedMaze, KeyCorridor, and MultiRoom), comparing with other methods. Following Wang et al. (2023), the visit-based and count-based intrinsic rewards are defined as follows:

$$\begin{aligned} r_t^{\text{visit}} &= \mathbb{1}(N_{ep}(s_{t+1}) = 1), \\ r_t^{\text{count}} &= \frac{1}{\sqrt{N_{ep}(s_{t+1})}}, \end{aligned} \tag{26}$$

where $\mathbb{1}$ is the indicator function, and N_{ep} is the number of visit in the current episode. From the experimental results, we can see that BCD demonstrates both effectiveness and efficiency. Specifically, in the ObstructedMaze-2Q and ObstructedMaze-2dlhb environments, BCD outperforms other baseline methods. Moreover, in the Obstructedmaze-2dlh, KeyCorridorS6R3-count, and MultiRoomN7S8-count environments, BCD exhibits faster convergence speed.

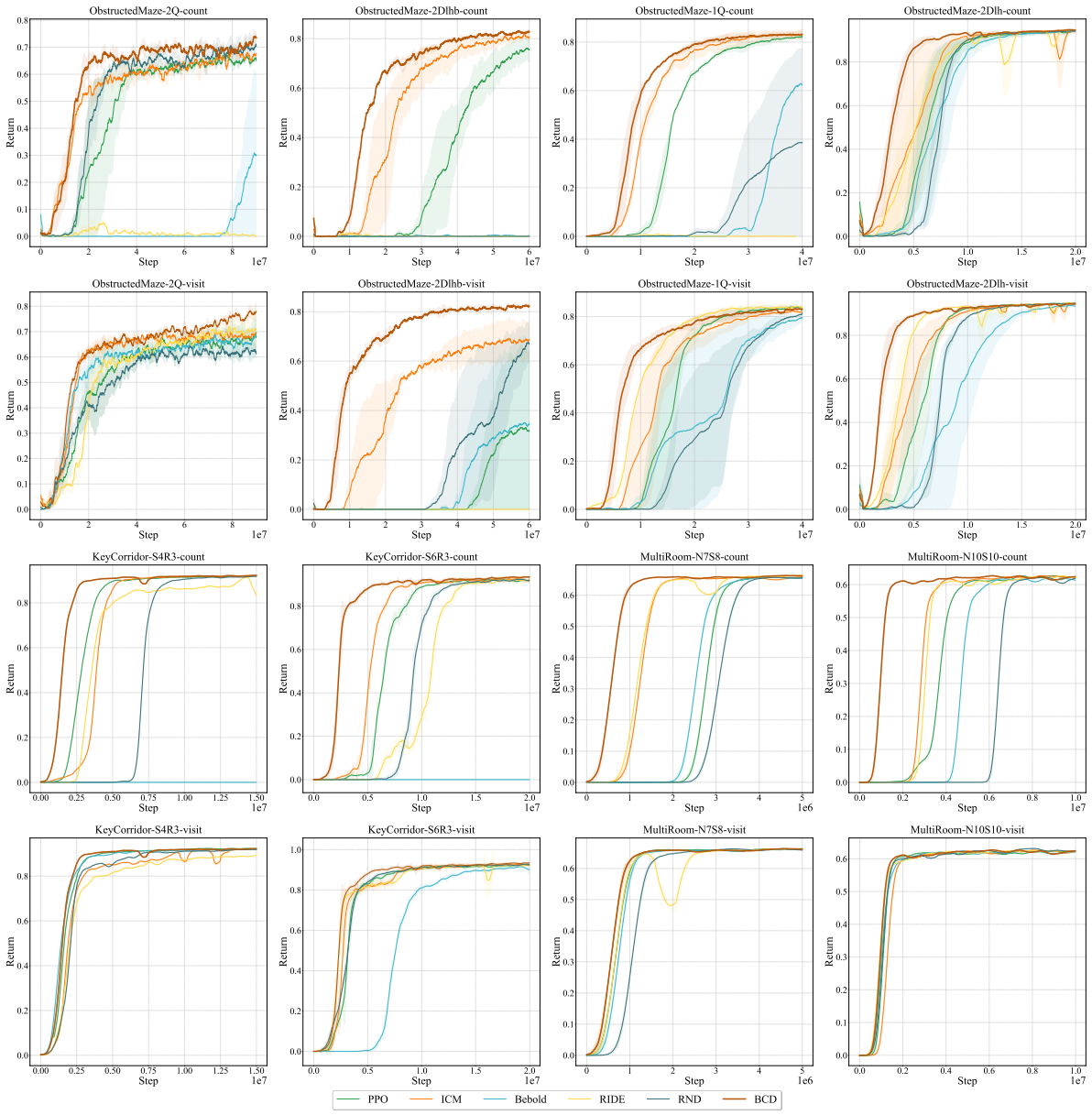

Figure 6: Comparisons with state-of-the-art methods in more environments.