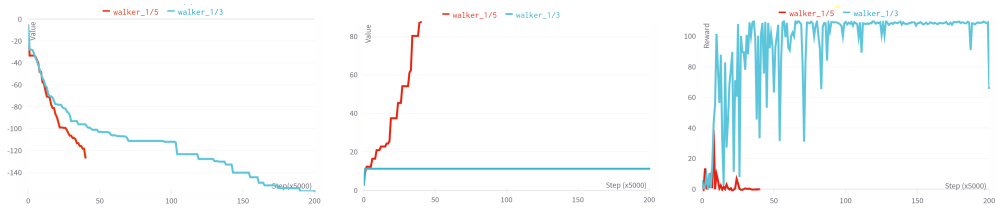
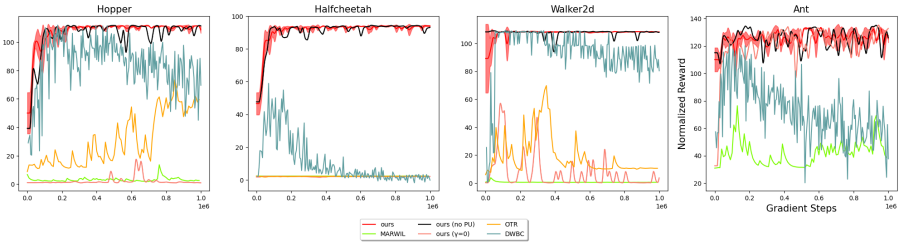
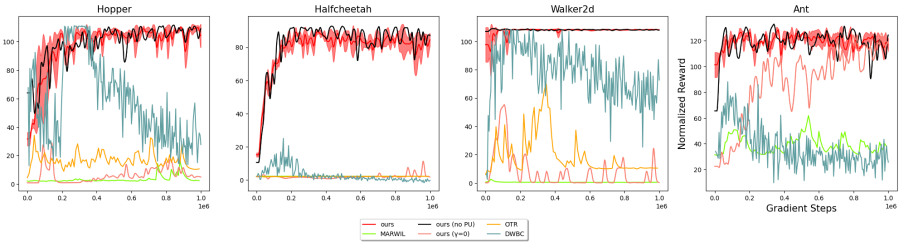


(a) Minimum $V(s)$ (b) Maximum $V(s)$ (c) Normalized Reward

Figure 25: Visualization of collapsing patterns of SMODICE-KL in Sec. 4.1 (walker_1/5 is collpasing; walker_1/3 is not). Walker_1/5 stops early due to NaN values in training, which we denote as 0 reward in our paper. It is clearly shown that the reward decrease is closely related to $V(s)$ divergence as analyzed in Sec. C. **(For reviewer iGFb Q6; reviewer Re2g Q5)**



(a) SMODICE (Sec. F.5) setting



(b) Sec. 4.3 setting

Figure 26: Comparison to DWBC (with expert action), OTR, MARWIL (single-iteration AWR), ours with no positive-unlabeled learning (no PU) and $\gamma = 0$. We found that both DWBC and OTR are not good at handling task-agnostic dataset with small portion of expert data. MARWIL performs similarly to plain BC, which is consistent with the findings from prior work. Our method with no PU generally works well, but $\gamma = 0$ fails on many environments. **(For reviewer iGFb Q3, Q4; reviewer 6nYj Q3, Q5; reviewer Re2g Q2, Q6)**

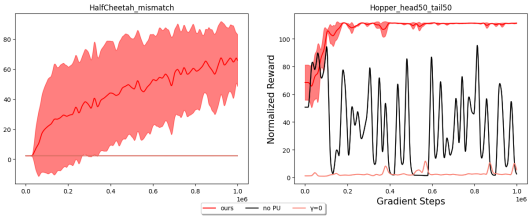


Figure 27: Examples of failing cases of our method with no positive-unlabeled learning and $\gamma = 0$. **(For reviewer iGFb Q4; reviewer 6nYj Q1; reviewer Re2g Q6)**