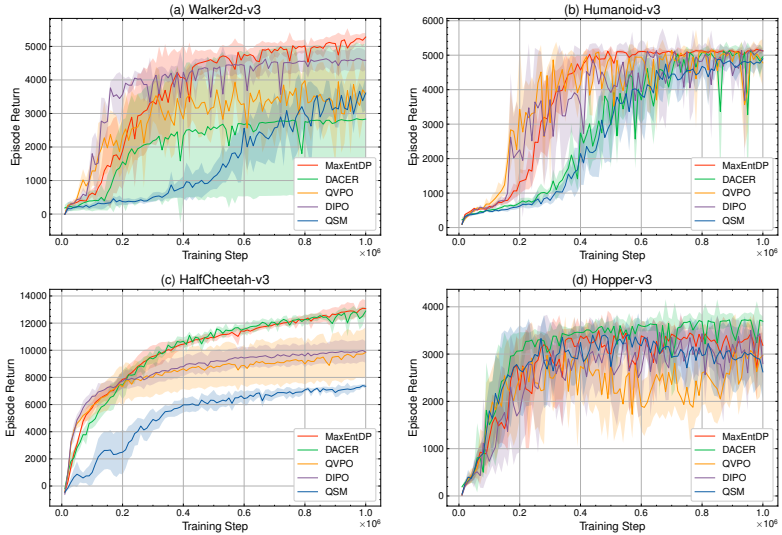

Figure 4. Learning curves of diffusion-based online RL algorithms after applying action selection to all algorithms with a candidate number of 10. Compared to the original result in the paper, the performance of QSM increased, while DACER and DIPO changed little since their greedy learning objective (maximize Q) has already ensured generating actions with high Q-value, and QVPO performed worse due to the decrease of candidate actions. Overall, MaxEntDP continued to demonstrate high sample efficiency and stability across all tasks.