Cross-Domain Offline Policy Adaptation with Optimal Transport and Dataset Constraint

Jiafei Lyu[1], Mengbei Yan[1], Zhongjian Qiao[1], Runze Liu[1], Xiaoteng Ma[1], Deheng Ye[2], Jing-Wen Yang[2], Zongqing Lu[3,4], Xiu Li[1]

[1]Tsinghua University [2]Tencent [3]BAAI [4]Peking University

2025/3/11

Introduction

- We study the offline policy adaptation problem where there exists dynamics shift between the source domain dataset and the target domain dataset
- We consider the setting that we have limited target domain data, about 5000 samples, since if one has a large amount of target domain data, one can directly use off-the-shelf offline RL methods like ReBRAC to achieve quite strong performance
- We consider the offline policy adaptation problem given very limited target domain data, with which the single-domain offline RL methods often struggle.

Framework

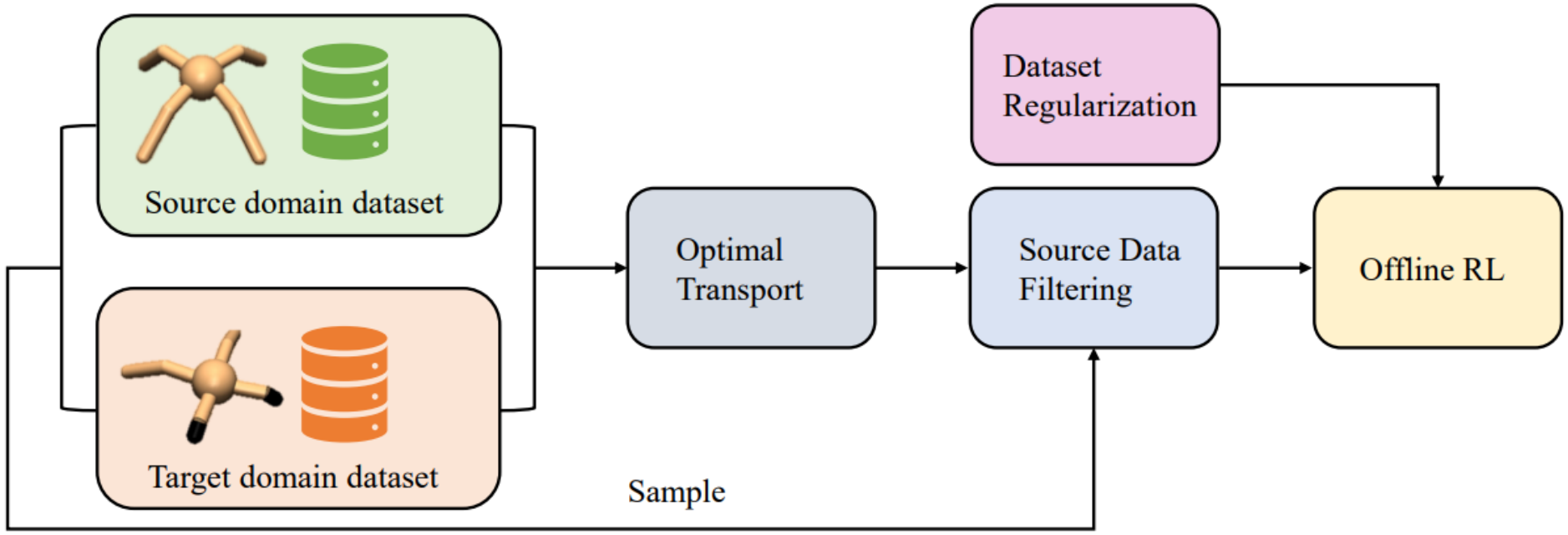


Figure 1: **An overview of our proposed framework.** We first align source domain data and target domain data via the Wasserstein distance. Then we adopt the solved optimal coupling for selectively sharing source domain transitions with the downstream offline RL algorithms. We further introduce a regularization term to encourage the learned policy to lie in the support region of the target domain.

Theoretical analysis

Theorem 3.1. *Denote the empirical policy distribution in the offline dataset D_{src} from source domain $\mathcal{M}_{\text{src}}$ and the offline dataset D_{tar} from target domain $\mathcal{M}_{\text{tar}}$ as $\pi_{D_{\text{src}}} := \frac{\sum_{D_{\text{src}}} \mathbb{1}(s,a)}{\sum_{D_{\text{src}}} \mathbb{1}(s)}$ and $\pi_{D_{\text{tar}}} := \frac{\sum_{D_{\text{tar}}} \mathbb{1}(s,a)}{\sum_{D_{\text{tar}}} \mathbb{1}(s)}$, respectively. Denote $C_1 = \frac{2r_{\max}}{(1-\gamma)^2}$, then the return difference of any policy π between the empirical source domain $\widehat{\mathcal{M}}_{\text{src}}$ and the true target domain $\mathcal{M}_{\text{tar}}$ is bounded:*

$$J_{\mathcal{M}_{\text{tar}}}(\pi) - J_{\widehat{\mathcal{M}}_{\text{src}}}(\pi) \geq -C_1 \underbrace{\mathbb{E}_{\rho^{\pi_{D_{\text{src}}}}_{\widehat{\mathcal{M}}_{\text{src}}}, P_{\widehat{\mathcal{M}}_{\text{src}}}} \left[D_{\text{TV}}(\pi_{D_{\text{src}}} \| \pi)\right]}_{(a)\text{: source policy deviation}} - C_1 \underbrace{\mathbb{E}_{\rho^{\pi}_{\mathcal{M}_{\text{tar}}}, P_{\mathcal{M}_{\text{tar}}}} \left[D_{\text{TV}}(\pi_{D_{\text{tar}}} \| \pi)\right]}_{(b)\text{: target policy deviation}}$$

$$- C_1 \underbrace{\mathbb{E}_{\rho^{\pi_D}_{\mathcal{M}_{\text{src}}}, \pi_{D_{\text{src}}}} \left[D_{\text{TV}}(P_{\mathcal{M}_{\text{tar}}} \| P_{\widehat{\mathcal{M}}_{\text{src}}})\right]}_{(c)\text{: dynamics mismatch}} - \text{constant}.$$

- The performance bound is related to three components

A novel objective

$$\mathcal{L}_Q = \mathbb{E}_{D_{\text{tar}}}\left[(Q_\theta - \mathcal{T}Q_\theta)^2\right] + \mathbb{E}_{(s,a,s')\sim D_{\text{src}}}\left[\mathbb{1}(\hat{p} > \epsilon)(Q_\theta - \mathcal{T}Q_\theta)^2\right] + \mathcal{R}_{D_{\text{src}}}(Q_\theta, \mathcal{T}Q_\theta),$$

$$\mathcal{L}_Q = \mathbb{E}_{D_{\text{tar}}}\left[(Q_\theta - \mathcal{T}Q_\theta)^2\right] + \mathbb{E}_{(s,a,s')\sim D_{\text{src}}}\left[\omega(s,a,s')\mathbb{1}(\hat{p} > \hat{p}_{\xi\%})(Q_\theta - \mathcal{T}Q_\theta)^2\right],$$

We define $u = s_{\text{src}} \oplus a_{\text{src}} \oplus s'_{\text{src}}$ and $u' = s_{\text{tar}} \oplus a_{\text{tar}} \oplus s'_{\text{tar}}$, where $\oplus$ is the the vector concatenation operator, $(s_{\text{src}}, a_{\text{src}}, s'_{\text{src}}) \sim D_{\text{src}}, (s_{\text{tar}}, a_{\text{tar}}, s'_{\text{tar}}) \sim D_{\text{tar}}$. Let $p_s = \frac{1}{|D_{\text{src}}|}\sum_{t=1}^{|D_{\text{src}}|}\delta_{u_t}$ and $p_t = \frac{1}{|D_{\text{tar}}|}\sum_{t=1}^{|D_{\text{tar}}|}\delta_{u'_t}$ denote the state-action-next-state joint distribution of the source domain dataset and the target domain dataset, respectively. Given a cost function C, the Wasserstein distance

$$\mathcal{W}(u, u') = \min_{\mu \in M} \sum_{t=1}^{|D_{\text{src}}|} \sum_{t'=1}^{|D_{\text{tar}}|} C(u_t, u'_{t'})\mu_{t,t'}$$

A novel objective

- Suppose the optimal coupling by solving the optimization problem above gives $\mu*$, we determine the deviation between a source domain data and the target domain dataset via:

$$d(u_t) = -\sum_{t'=1}^{|D_{\text{tar}}|} C(u_t, u'_{t'})\mu^*_{t,t'}, \quad u_t = (s^t_{\text{src}}, a^t_{\text{src}}, (s'_{\text{src}})^t) \sim D_{\text{src}}.$$

- Intuitively, d is larger if the source domain data aligns the distribution of the target domain dataset (since the cost is smaller by then) and smaller otherwise. It can hence work as a good proxy for $\hat{p}$

Theorem 3.2. *Assume that the cost is bounded, i.e.,* $C(u, u') \leq C_{\max} < \infty, \forall\, u, u'$*, then we have*

$$0 \geq d(u_t) \geq -C_{\max} D_{\text{TV}}(p_s \| p_t).$$

A novel objective

- Another benefit of computing d is that we can calculate the weight $\omega = \exp(\alpha \times d(u))$

$$\mathcal{L}_Q = \mathbb{E}_{D_{\text{tar}}}\left[(Q_\theta - \mathcal{T}Q_\theta)^2\right] + \mathbb{E}_{(s,a,s')\sim D_{\text{src}}}\left[\exp(\alpha \times d)\mathbb{1}(d > d_{\xi\%})(Q_\theta - \mathcal{T}Q_\theta)^2\right],$$

- However, it is insufficient since the performance bound in Theorem 3.1 is also connected with policy deviation terms. If only limited target domain data is available (e.g., 5000 transitions), the learned policy can get biased towards the behavior policy of the source domain dataset, incurring unsatisfying policy adaptation performance.

$$\widehat{\mathcal{L}}_\pi = \mathcal{L}_\pi - \beta \times \mathbb{E}_{s\sim D_{\text{src}}\cup D_{\text{tar}}} \log \pi^b_{\text{tar}}(\pi(\cdot|s)|s),$$

Practical implementation

- We normalize the deviation to omit the introduced hyperparameter

$$\widehat{d}_i = \frac{d_i - \max_i d_i}{\max_i d_i - \min_i d_i}, \quad i \in \{1, 2, \ldots, N\},$$

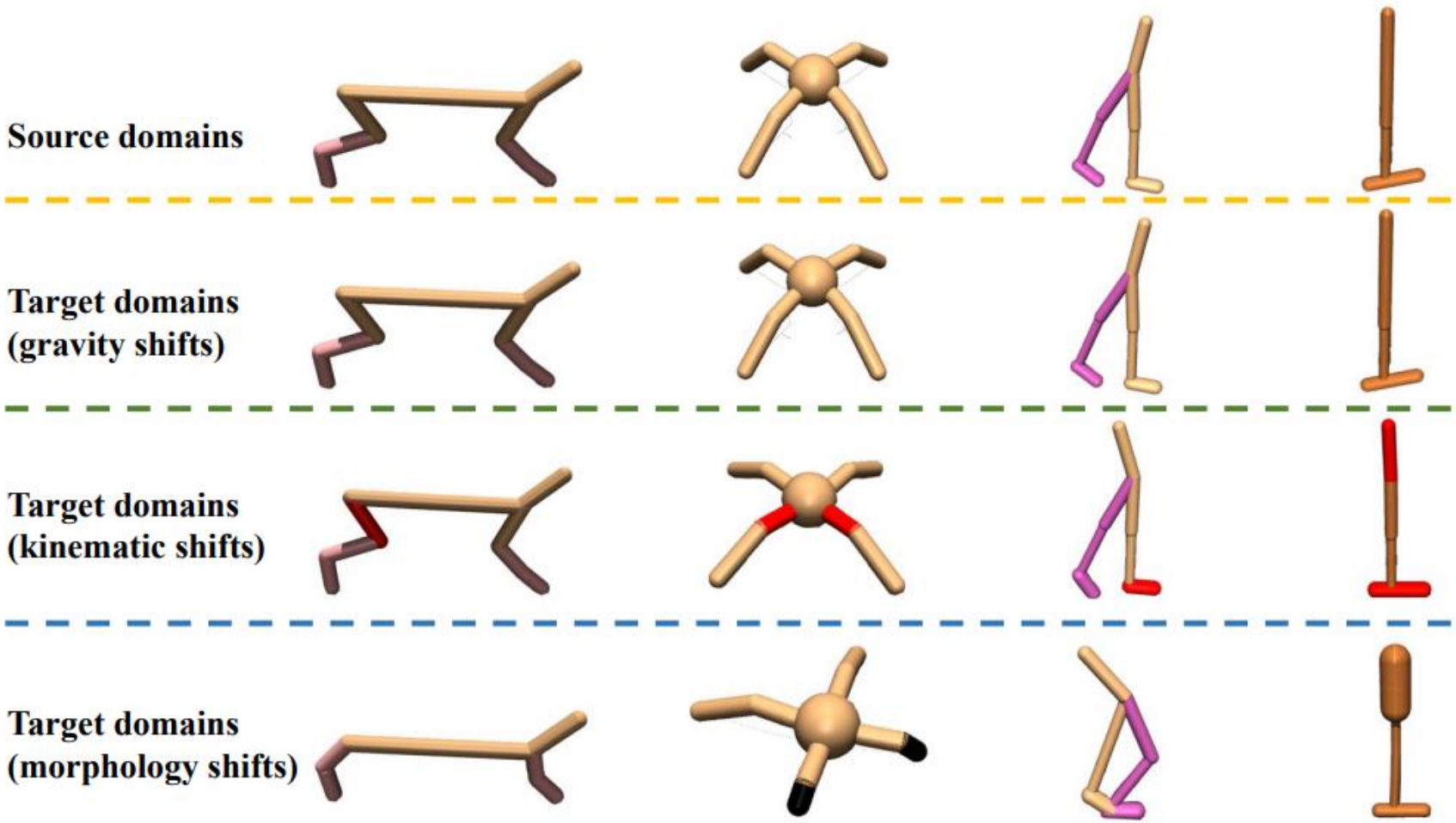
$$\mathcal{L}_Q = \mathbb{E}_{D_{\text{tar}}}\left[(Q_\theta - \mathcal{T}Q_\theta)^2\right] + \mathbb{E}_{(s,a,s')\sim D_{\text{src}}}\left[\exp(\widehat{d})\mathbb{1}(d > d_{\xi\%})(Q_\theta - \mathcal{T}Q_\theta)^2\right].$$

- Furthermore, we model the behavior policy in the target domain with CVAE

$$\mathcal{L}_{\text{CVAE}} = \mathbb{E}_{(s,a)\sim D_{\text{tar}}, z\sim E_\nu(s,a)}\left[(a - D_\varsigma(s,z))^2 + D_{\text{KL}}\left(E_\nu(s,a)\|\mathcal{N}(0,\mathbf{I})\right)\right],$$

$$\widehat{\mathcal{L}}_\pi = \mathcal{L}_\pi - \beta \times \mathop{\mathbb{E}}_{s\sim D_{\text{src}}\cup D_{\text{tar}}} \log\left[\sum_{i=1}^{M} \exp(\log \widehat{\pi}^i_{\text{tar}}(\pi(\cdot|s)|s))\right],$$

Experiments

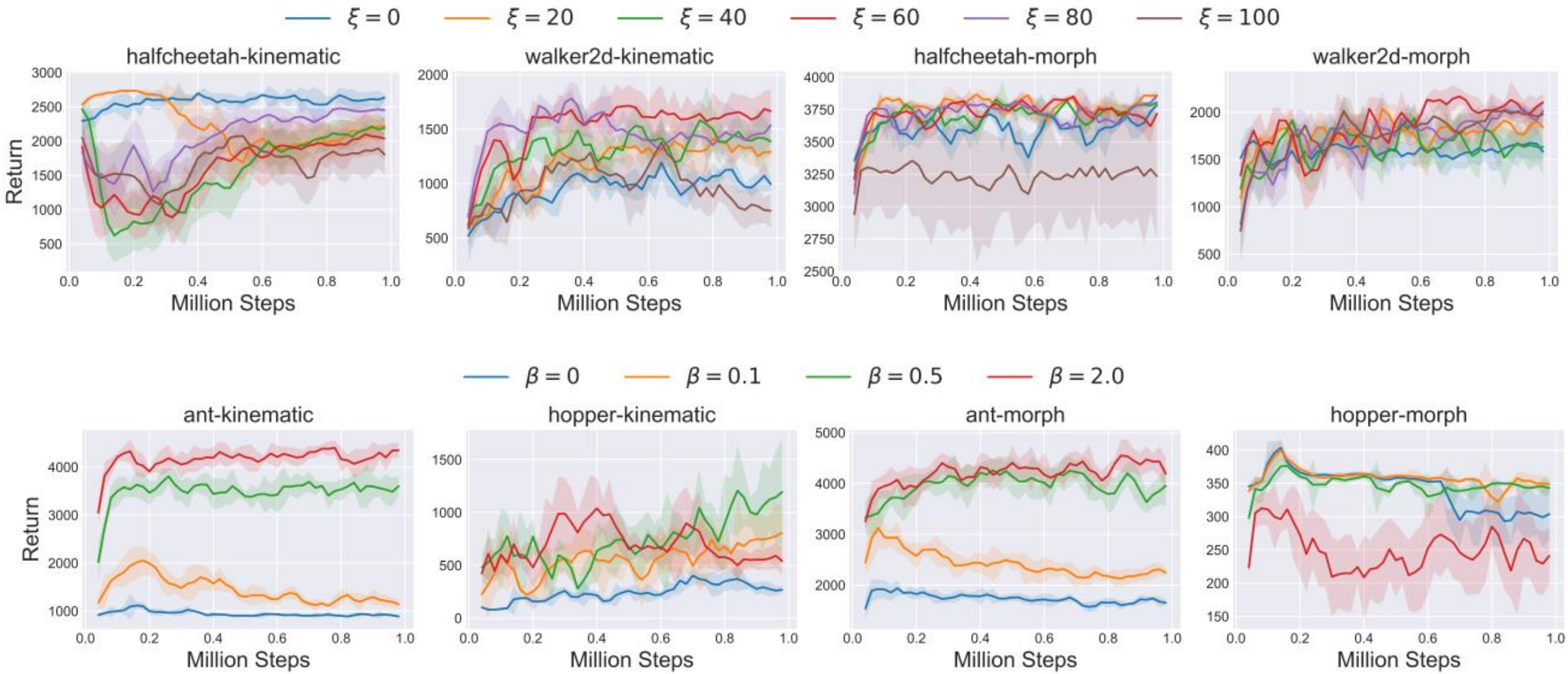


Experiments

Source	Target	IQL*	DARA	BOSA	SRPO	IGDF	OTDF (ours)
half-m	medium	30.0±1.6	26.6±3.3	19.3±3.5	41.3±0.4	**41.6**±0.5	39.1±2.3
half-m	medium-expert	31.8±1.1	32.0±0.7	33.6±1.1	30.7±0.8	29.6±2.2	**35.6**±0.7
half-m	expert	8.5±1.0	9.3±1.6	7.9±0.8	8.6±0.9	10.0±0.8	**10.7**±1.2
half-m-r	medium	30.8±4.4	35.6±0.7	35.0±4.6	32.0±1.4	28.0±2.0	**40.0**±1.2
half-m-r	medium-expert	12.9±2.2	16.9±4.1	19.9±5.5	12.4±1.6	12.0±3.7	**34.4**±0.7
half-m-r	expert	5.9±1.7	3.7±2.7	2.4±1.9	6.2±1.4	5.3±2.3	**8.2**±2.7
half-m-e	medium	**41.5**±0.1	40.3±1.2	41.3±0.3	41.3±0.4	40.9±0.4	41.4±0.3
half-m-e	medium-expert	25.8±2.0	30.6±2.8	32.1±0.8	27.2±0.8	26.2±1.8	**35.1**±0.6
half-m-e	expert	7.8±1.3	8.3±1.3	9.1±0.8	7.8±0.9	7.5±0.9	**9.8**±1.0
hopp-m	medium	**13.5**±0.2	**13.5**±0.4	13.2±0.3	13.4±0.1	13.4±0.2	11.0±0.9
hopp-m	medium-expert	13.4±0.1	**13.6**±0.2	11.2±4.6	13.3±0.2	13.3±0.4	12.6±0.8
hopp-m	expert	13.5±0.2	13.6±0.3	13.3±0.4	13.6±0.2	**13.9**±0.1	10.7±4.7
hopp-m-r	medium	10.8±1.1	10.2±1.0	1.2±0.0	10.7±1.6	**12.0**±4.4	8.7±2.8
hopp-m-r	medium-expert	**11.6**±1.6	10.4±0.9	1.3±0.2	10.4±1.2	8.2±2.8	9.7±2.7
hopp-m-r	expert	9.8±0.5	9.0±0.3	1.3±0.1	10.4±1.4	**11.4**±1.5	10.7±2.4
hopp-m-e	medium	12.6±1.4	13.0±0.5	**15.7**±7.2	14.0±2.3	12.7±0.8	7.9±3.2
hopp-m-e	medium-expert	**14.1**±1.3	13.8±0.6	12.0±1.4	13.5±0.3	13.3±1.2	9.6±3.5
hopp-m-e	expert	13.8±0.5	12.3±1.8	10.5±5.0	**14.7**±2.3	12.8±0.9	5.9±4.0
walk-m	medium	23.0±4.7	23.3±3.3	6.2±2.9	24.7±1.7	27.5±9.5	**50.5**±5.8
walk-m	medium-expert	21.5±8.6	22.2±7.6	7.2±2.9	18.7±7.3	20.7±5.9	**44.3**±23.8
walk-m	expert	20.3±2.8	17.3±3.4	15.8±8.7	21.1±7.2	15.8±4.5	**55.3**±8.3
walk-m-r	medium	11.3±3.0	10.9±4.6	5.4±4.0	10.4±4.8	13.4±7.2	**37.4**±5.1
walk-m-r	medium-expert	7.0±1.5	4.5±1.1	4.0±2.2	4.9±1.7	6.9±2.2	**33.8**±6.9
walk-m-r	expert	6.3±0.9	4.5±1.1	3.8±3.4	5.5±0.9	5.5±2.2	**41.5**±6.8
walk-m-e	medium	24.1±7.4	31.7±6.6	18.7±6.5	29.9±4.7	27.5±2.3	**49.9**±4.6
walk-m-e	medium-expert	27.0±5.5	23.3±5.5	11.1±0.9	22.9±3.8	25.3±6.4	**40.5**±11.0
walk-m-e	expert	22.4±3.3	25.2±5.7	9.9±3.9	18.7±5.7	24.7±2.4	**45.7**±6.9
ant-m	medium	38.7±3.8	**41.3**±1.8	18.2±1.9	40.6±2.1	40.9±1.7	39.4±1.7
ant-m	medium-expert	47.0±5.1	43.3±2.0	45.3±7.0	47.2±4.3	44.4±1.7	**58.3**±8.9
ant-m	expert	36.2±3.5	48.5±4.2	72.2±10.5	42.2±9.9	41.4±4.2	**85.4**±4.4
ant-m-r	medium	38.2±2.9	38.9±2.7	20.2±3.7	38.3±1.9	39.7±1.2	**41.2**±0.9
ant-m-r	medium-expert	38.1±3.5	33.4±5.5	15.2±1.6	35.0±5.7	37.3±2.4	**50.8**±4.5
ant-m-r	expert	24.1±1.9	24.5±2.6	16.0±1.7	22.7±3.0	23.6±1.4	**67.2**±7.5
ant-m-e	medium	32.9±5.1	**40.2**±1.5	28.1±5.6	35.9±2.5	36.1±4.4	39.9±2.9
ant-m-e	medium-expert	35.7±3.9	36.5±8.7	14.8±15.9	24.5±15.7	30.7±10.8	**65.7**±4.5
ant-m-e	expert	36.1±8.5	34.6±5.8	53.9±5.0	38.4±9.4	35.2±6.6	**86.4**±2.2
Total Score		798.0	816.8	646.3	803.1	808.7	**1274.3**

Source	Target	IQL*	DARA	BOSA	SRPO	IGDF	OTDF (ours)
half-m	medium	39.6±3.3	**41.2**±3.9	38.9±4.0	36.9±4.5	36.6±5.5	40.7±7.7
half-m	medium-expert	39.6±3.7	**40.7**±2.8	40.4±3.0	**40.7**±2.3	38.7±6.2	28.6±3.2
half-m	expert	**42.4**±3.8	39.8±4.4	40.5±3.9	39.4±1.6	39.6±4.6	36.1±5.3
half-m-r	medium	20.1±5.0	17.6±6.2	20.0±4.9	17.5±5.2	14.4±2.2	**21.5**±6.5
half-m-r	medium-expert	17.2±1.6	**20.2**±5.2	16.7±4.2	16.3±1.7	10.0±2.5	14.7±4.1
half-m-r	expert	20.7±5.5	22.4±1.7	15.4±4.2	**23.1**±4.0	15.3±3.7	11.4±1.9
half-m-e	medium	38.6±6.0	37.8±3.3	41.8±5.1	**42.5**±2.3	37.7±7.3	39.5±3.5
half-m-e	medium-expert	39.6±3.0	39.4±4.4	38.7±3.7	**43.3**±2.7	40.7±3.2	32.4±5.5
half-m-e	expert	43.4±0.9	**45.3**±1.3	39.9±2.7	43.3±3.0	41.1±4.1	26.5±9.1
hopp-m	medium	11.2±1.1	17.3±3.8	15.2±3.3	12.4±1.0	15.3±3.5	**32.4**±8.0
hopp-m	medium-expert	14.7±3.6	15.4±2.5	21.1±9.3	14.2±1.8	15.1±3.6	**24.2**±3.6
hopp-m	expert	12.5±1.6	19.3±10.5	12.7±1.7	11.8±0.9	14.8±4.0	**33.7**±7.8
hopp-m-r	medium	13.9±2.9	10.7±4.3	3.3±1.9	14.0±2.6	15.3±4.4	**31.1**±13.4
hopp-m-r	medium-expert	13.3±6.3	12.5±5.6	4.6±1.7	14.4±4.2	15.4±5.5	**24.2**±6.1
hopp-m-r	expert	11.0±2.6	14.3±6.0	3.2±0.8	16.4±5.0	16.1±4.0	**31.0**±9.8
hopp-m-e	medium	19.1±6.6	18.5±12.3	15.9±5.9	19.7±8.5	22.3±5.4	**26.4**±10.1
hopp-m-e	medium-expert	16.8±2.7	16.0±6.1	17.3±2.5	15.8±3.3	16.6±7.7	**28.3**±6.7
hopp-m-e	expert	20.9±4.1	23.9±14.8	23.2±7.9	21.4±1.9	26.0±9.2	**44.9**±10.6
walk-m	medium	28.1±12.9	28.4±13.7	**38.0**±11.2	21.4±7.0	22.1±8.4	36.6±2.3
walk-m	medium-expert	35.7±4.7	30.7±9.7	40.9±7.2	34.0±9.9	35.4±9.1	**44.8**±7.5
walk-m	expert	37.3±8.0	36.0±7.0	41.3±8.6	39.5±3.8	36.2±13.6	**44.0**±4.0
walk-m-r	medium	14.6±2.5	14.1±6.1	7.6±5.8	17.9±3.8	11.6±4.6	**32.7**±7.0
walk-m-r	medium-expert	15.3±1.9	15.9±5.8	4.8±5.8	15.3±4.5	13.9±6.5	**31.6**±6.1
walk-m-r	expert	15.8±7.2	15.7±4.5	7.1±4.6	13.7±8.1	15.2±5.3	**31.3**±5.3
walk-m-e	medium	39.9±13.1	41.6±13.0	32.3±7.2	**46.4**±3.5	33.8±3.1	30.2±9.8
walk-m-e	medium-expert	49.1±6.9	45.8±9.4	40.1±4.5	36.4±3.4	44.7±2.9	**53.3**±7.1
walk-m-e	expert	40.4±11.9	56.4±3.5	43.7±4.4	45.8±8.0	45.3±10.4	**61.1**±3.4
ant-m	medium	10.2±1.8	9.4±0.9	12.4±2.0	11.7±1.0	11.3±1.3	**45.1**±12.4
ant-m	medium-expert	9.4±1.2	10.0±0.9	11.6±1.3	10.2±1.2	9.4±1.4	**33.9**±5.4
ant-m	expert	10.2±0.3	9.8±0.6	11.8±0.4	9.5±0.6	9.7±1.6	**33.2**±9.0
ant-m-r	medium	18.9±2.6	21.7±2.1	13.9±1.5	18.7±1.7	19.6±1.0	**29.6**±10.7
ant-m-r	medium-expert	19.1±3.0	18.3±2.1	15.9±2.7	18.7±1.8	20.3±1.6	**25.4**±2.1
ant-m-r	expert	18.5±0.9	20.0±1.3	14.5±1.7	19.9±2.1	18.8±2.1	**24.5**±2.8
ant-m-e	medium	9.8±2.4	8.1±1.8	8.1±3.0	8.4±2.1	8.9±1.5	**18.6**±11.9
ant-m-e	medium-expert	9.0±0.8	6.4±1.4	6.2±1.5	6.1±3.5	7.2±2.9	**34.0**±9.4
ant-m-e	expert	9.1±2.6	10.4±2.9	4.2±3.9	8.8±1.0	9.2±1.5	**23.2**±2.9
Total Score		825.0	851.0	763.2	825.5	803.6	**1160.7**

Parameter study



Thanks!