NEURAL INFORMATION PROCESSING SYSTEMS
NEURREPS @ NEURIPS 2025

清華大学智能产业研究院 Institute for AI Industry Research,Tsinghua University

北京交通大學 BEIJING JIAOTONG UNIVERSITY

上海人工智能实验室 Shanghai Artificial Intelligence Laboratory

Sample Efficient Offline RL via T-symmetry Enforced Latent State-Stitching

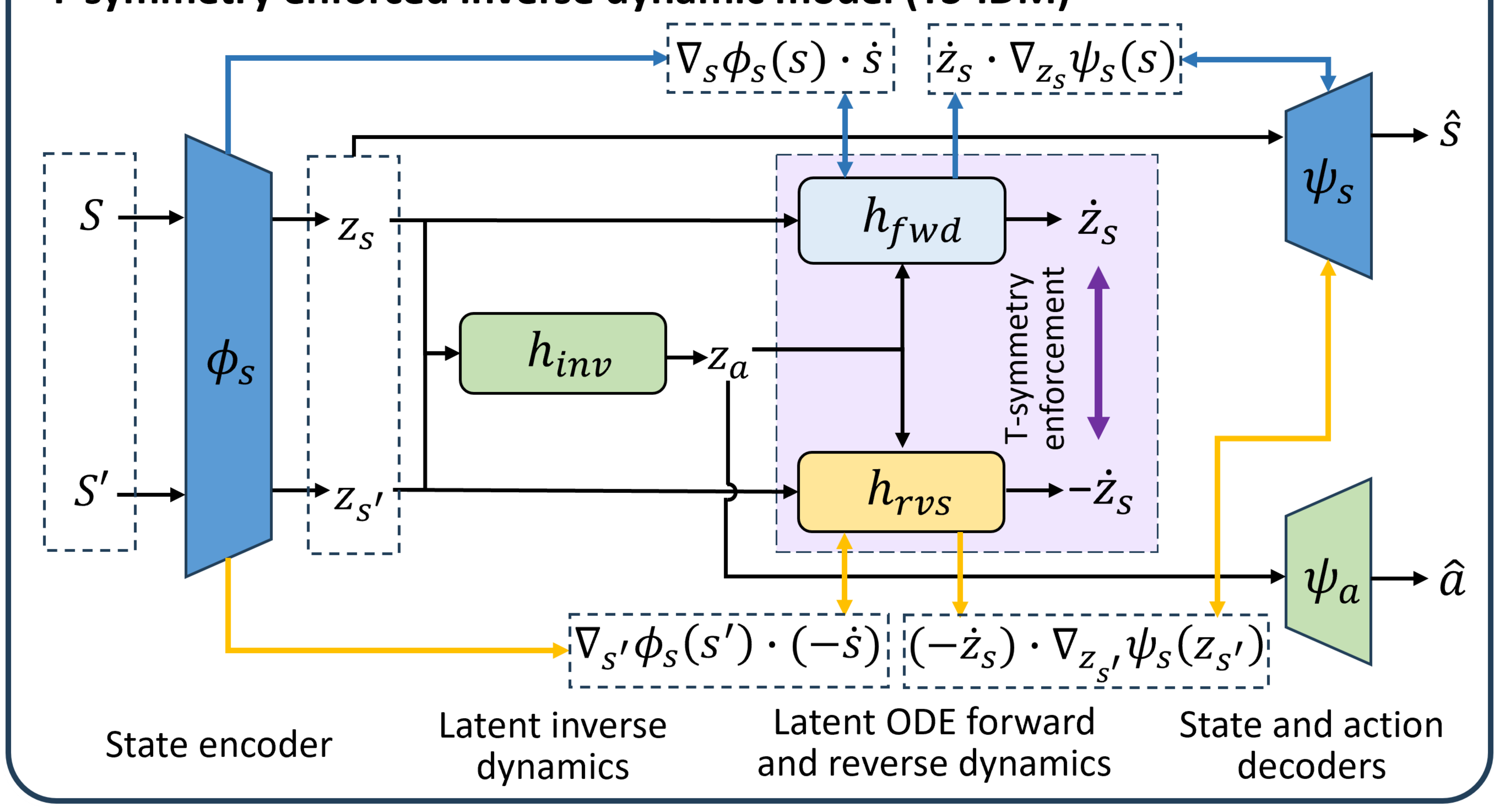
Peng Cheng, Zhihao Wu, Jianxiong Li, Ziteng He, Haoran Xu, Wei Sun, Youfang Lin, Xianyuan Zhan

Introduction

- Current offline RL methods require a large amount of training data to achieve reasonable performance and offer limited generalizability in out-of-distribution (OOD) regions due to conservative data-related regularizations.
- We propose a **highly sample-efficient offline RL algorithm (TELS)** that learns optimized policy within the latent space regulated by the fundamental T-symmetry in the dynamical systems.
- Our approach achieves amazing sample efficiency and OOD generalizability, significantly outperforming existing offline RL methods in various small-sample tasks, even **using as few as 1% of the data samples** in D4RL datasets.

T-symmetry Enforced Latent State-Stitching (TELS)

T-symmetry enforced inverse dynamic model (TS-IDM)



1 Latent state-value function learning

Learning in the latent space

$$\min_V \mathbb{E}_{(s,r,s')\sim\mathcal{D}}\left[L_2^\tau\left(r+\gamma\bar{V}(\phi_s(s'))-V(\phi_s(s))\right)\right]$$

2 T-symmetry regularized guide-policy optimization

$$\max_{\pi_g} \mathbb{E}_{(s,s')\sim\mathcal{D}}\left[\lambda_\alpha V\left(\pi_g(z_s)\right)-\eta\left\|\psi_s\left(\pi_g(z_s)\right)-s'\right\|_2^2 - \ell_{T-sym}\left(z_s, h_{inv}\left(z_s,\pi_g(z_s)\right)\right)\right]$$

Or

$$\max_{\pi_g} \mathbb{E}_{(s,s')\sim\mathcal{D}}\left[\exp\left(\alpha\cdot A(z_s,z_{s'})\right)\log\pi_g\left(z_{s'}|z_s\right) - \ell_{T-sym}\left(z_s, h_{inv}\left(z_s,\pi_g(\cdot|z_s)\right)\right)\right]$$

T-symmetry regularization

3 Action inference

Extract optimized action using latent inverse dynamics

$$a^*=\psi_a\left(h_{inv}\left(z_s,\pi_g(z_s)\right)\right)$$

- **TS-IDM:** Learns well-behaved latent representations that address OOD generalization challenges and enhance action inference efficiency.
- **Latent Space Offline Policy Optimization:** Learns a latent state-value function and T-symmetry-regularized guide-policy to generate valuable and reliable next latent states, enabling TS-IDM's inverse dynamics to infer optimal actions.

Main Results

Table 1: Average normalized scores on reduced-size D4RL datasets. The scores are taken over the final 10 evaluations with 5 seeds.

Task	Size (ratio)	BC	TD3+BC	CQL	IQL	DOGE	IDQL	POR	TSRL	TELS
Hopper-m	10k (1%)	29.7±11.7	40.1±18.6	43.1±24.6	46.7±6.5	44.2 ± 10.2	44.2±12.1	46.4 ± 1.7	62.0±3.7	**77.3 ± 10.7**
Hopper-mr	10k (2.5%)	12.1±5.3	7.3±6.1	2.3±1.9	13.4±3.1	17.9 ± 4.5	21.7±7.0	17.4 ± 6.2	21.8±8.2	**43.2 ± 3.5**
Hopper-me	10k (0.5%)	27.8±10.7	17.8±7.9	29.9±4.5	34.3±8.7	50.5 ± 25.2	43.2±4.4	37.9 ± 6.1	50.9±8.6	**100.9 ± 6.8**
Halfcheetah-m	10k (1%)	26.4±7.3	16.4±10.2	35.8±3.8	29.9±0.12	36.2 ± 3.4	36.4±1.5	33.3±3.2	38.4±3.1	**40.8 ± 0.6**
Halfcheetah-mr	10k (5%)	14.3±7.8	17.9±9.5	8.1±9.4	22.7±6.4	23.4 ± 3.6	26.7±1.0	27.5±3.6	28.1±3.5	**33.2 ± 1.0**
Halfcheetah-me	10k (0.5%)	19.1±9.4	15.4±10.7	26.5±10.8	10.5±8.8	26.7 ± 6.6	38.8±1.9	34.7±2.6	39.9±21.1	**40.7 ± 1.2**
Walker2d-m	10k (1%)	15.8±14.1	7.4±13.1	18.8±18.8	22.5±3.8	45.1 ± 10.2	31.7±14.2	22.2±3.6	49.7±10.6	**62.4 ± 5.3**
Walker2d-mr	10k (3.3%)	1.4±1.9	5.7±5.8	8.5±2.19	10.7±11.9	13.5 ± 8.4	12.2±10.5	14.8±4.2	26.0±11.3	**54.8 ± 6.0**
Walker2d-me	10k (0.5%)	21.7±8.2	7.9±9.1	19.1±14.4	26.5±8.6	35.3 ± 11.6	21.8±14.5	20.1±8.6	46.4±17.4	**87.4 ± 13.3**
Antmaze-u	10k (1%)	44.7 ± 42.1	0.7 ± 1.2	0.1 ± 0.0	65.1 ± 19.4	56.3 ± 24.4	67.5 ±12.4	6.1 ± 7.3	76.1 ± 15.6	**88.7 ± 7.7**
Antmaze-u-d	10k (1%)	24.1 ± 22.2	16.27 ± 16.4	0.5 ± 0.1	34.6 ± 18.5	41.7 ± 18.9	55.1 ± 36.8	42.1 ± 14.2	52.2 ± 22.1	**60.9 ± 16.9**
Antmaze-m-d	100k (10%)	0.0	0.0	0.0	4.8 ± 5.9	0.0	9.0 ±3.4	0.0	0.0	**47.2 ± 17.3**
Antmaze-m-p	100k (10%)	0.0	0.0	0.0	12.5 ± 5.4	0.0	9.4 ± 14.7	0.0	0.0	**62.9 ± 17.8**
Antmaze-l-d	100k (10%)	0.0	0.0	0.0	3.6 ± 4.1	0.0	16.1 ± 8.4	0.0	0.0	**39.8 ± 14.1**
Antmaze-l-p	100k (10%)	0.0	0.0	0.0	3.5 ± 4.1	0.0	9.7 ±8.5	0.0	0.0	**47.3 ± 13.1**

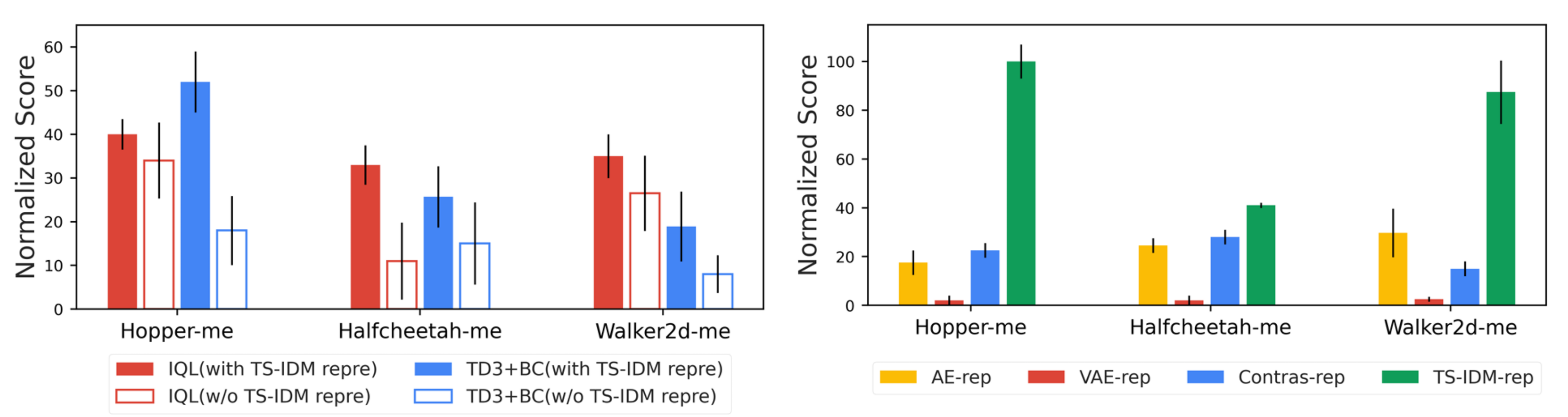


Figure 4: **Left:** The performance of IQL and TD3+BC on 10k datasets with or without using the representation from TS-IDM. **Right:** Performance of TELS with different representation models on 10k datasets, error bars indicate the normalized scores over 5 random seeds.

Table 2: Ablation results on the design components of TS-IDM.

	$\phi/\psi + h_{inv}$	$+ h_{fwd}, h_{rvs}$ ↑	$+ \ell_{\text{ode}}$ ↑	$+ \ell_{\text{T-sym}}$ ↑
Hopper-me	17.2 ± 7.0	35.5 ± 7.3	61.4 ± 23.7	**100.9 ± 6.8**
Halfcheetah-me	29.7 ± 3.6	31.3 ± 1.1	31.2 ± 1.2	**40.7 ± 1.2**
Walker2d-me	24.5 ± 10.1	33.6 ± 9.2	58.5 ± 18.1	**87.4 ± 13.3**

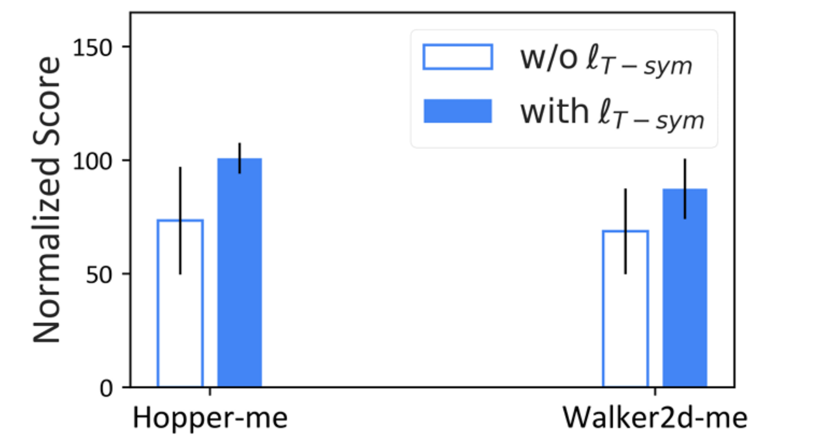


Figure 5: Impact of $\ell_{\text{T-sym}}$ on policy optimization

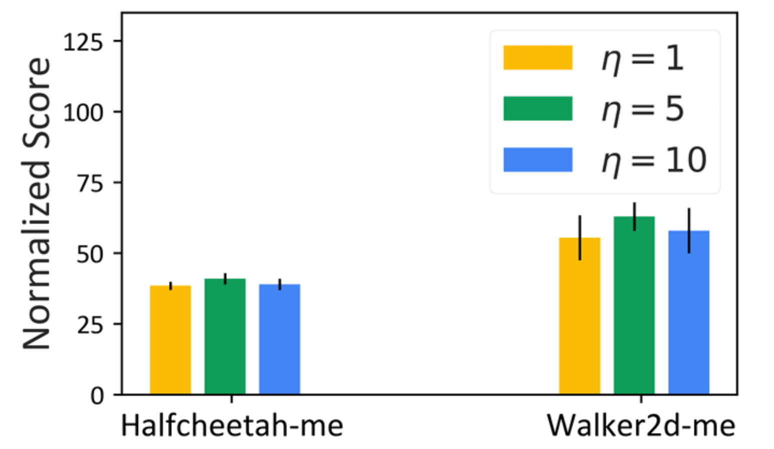


Figure 6: Performance of TELS with different η

Out-of-Distribution Generalizability of TELS

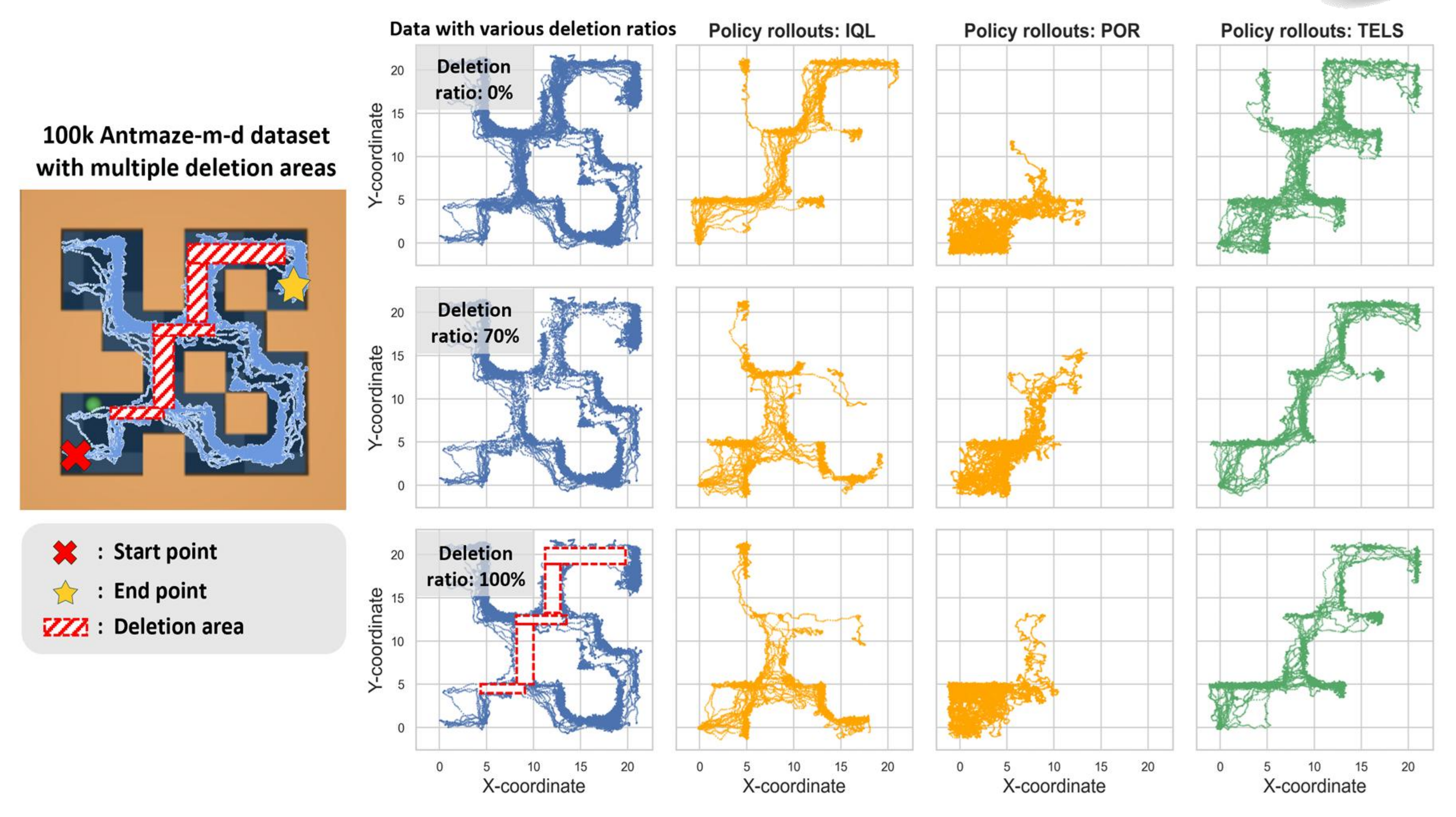
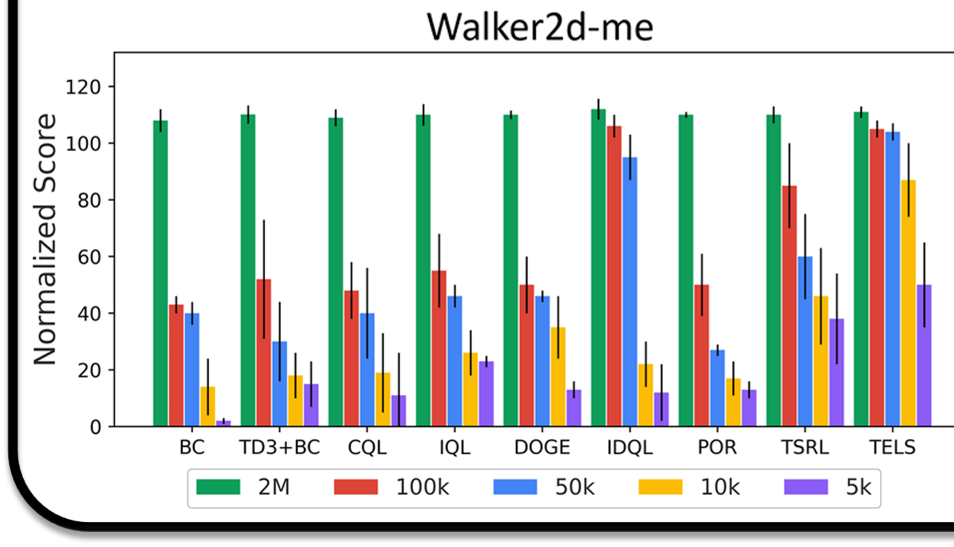


Figure 3: **Left:** Illustration of the 100k Antmaze-m-d task with multiple deletion areas, where the red cross denotes the start point, the yellow star denotes the goal locations, and the red shaded areas denote the data deletion regions. **Right:** Visualization of the training dataset and policy rollout trajectories generated by trained policies from various algorithms under varying deletion ratios.

- We randomly remove samples within 5 critical regions along the critical paths from the start to the goal locations.
- **Only TELS** consistently learns optimal policy even with **70% and 100% deletion rates**.
- These highlight the OOD generalization capability of TELS in extremely challenging low-data regimes.

Offline RL under Various Data Size



Performance of offline RL methods and TELS on the D4RL MuJoCo *Walker2d-mediu-expert-v2* datasets when reducing the number of samples from 2M (full dataset) to 5k (0.5%).