Offline Imitation Learning with Model-based Reverse Augmentation

Jie-Jing Shao, Hao-Sen Shi, Lan-Zhe Guo, Yu-Feng Li
National Key Laboratory for Novel Software Technology
Nanjing University, China
{shaojj, shihs, guolz, liyf}@lamda.nju.edu.cn

KDD2024 BARCELONA, SPAIN

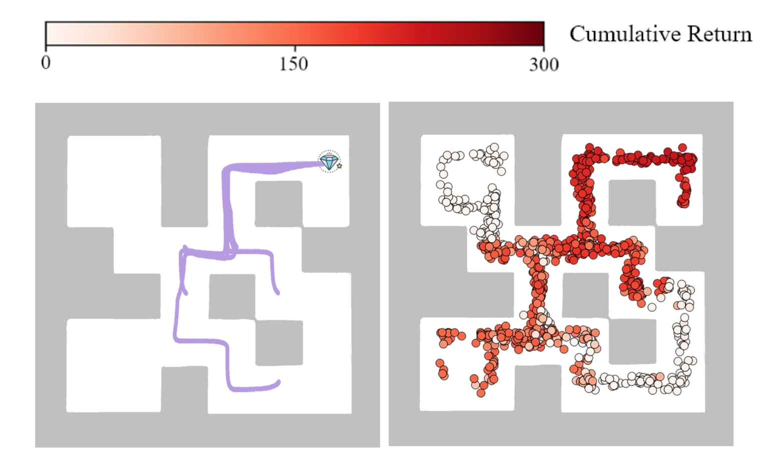
Webpage Code WeChat

Offline Imitation Learning

Imitation Learning: recover a policy from expert dataset.
In offline IL, expert data is limited, supplementary offline data is rich but low-quality.

Without reward supervision, it is difficult to determine what action an agent should take when outside the state distribution of the expert demonstrations.

Previous methods keep agents conservative, confining them to the expert-observed area.
MILO [NeurIPS'21], CLARE [ICLR'23], ML-IRL[NeurIPS'23]…

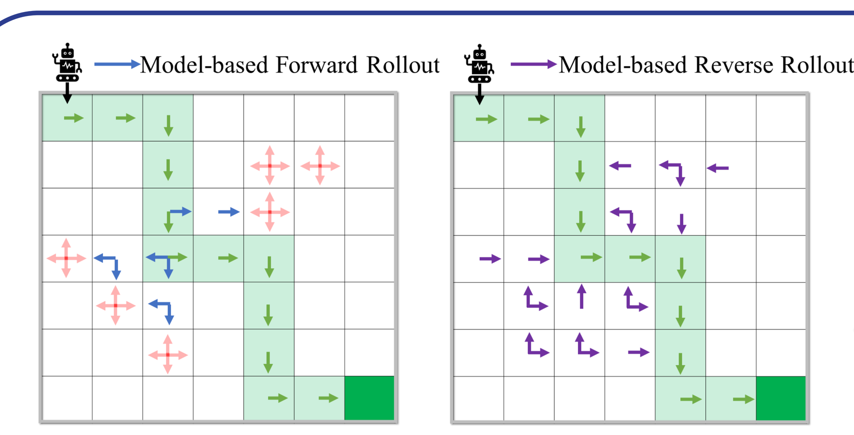


Evaluation of MILO

The Proposed Method: Self-Paced Reverse Augmentation

Main Idea

We prefer the actions which lead the agent from expert-unobserved states to expert-observed states.



Generate reverse trajectories

Agents could follow reverse trajectories to reach the expert-observed area, improving long-term returns.

Reverse Models

Reverse dynamic model $\hat{T}_r$, approximating $T(s_t|s_{t+1}, a_t)$:

$$\max_{\hat{T}_r} \sum_{(s_t, a_t, s_{t+1})} \log \hat{T}_r\,(s_t|s_{t+1}, a_t)$$

Reverse behavior policy π_r, a VAE-based actor, approximating $p(a_t|s_{t+1})$:

$$\log \pi_r\,(a|s)$$
$$\geq \mathbb{E}_{z\sim\pi_r^e(\cdot|s,a)} \log \pi_r^d\,(a|z,s) - KL(\pi_r^e(z|s,a)||p(z|s))$$

Self-Paced Augmentation

First step of reverse augmentation with h steps:

$$\{s_{-h'}, a_{-h'}, s_{-h'+1}, a_{-h'+1}, \ldots, s_{-2}, a_{-2}, s_{-1}, a_{-1}, s_0\}$$
$$s_0 \sim D^E, a_i \sim \pi_r(s_{i+1}), s_i \sim \hat{T}_r(s_{i+1}, a_i)$$

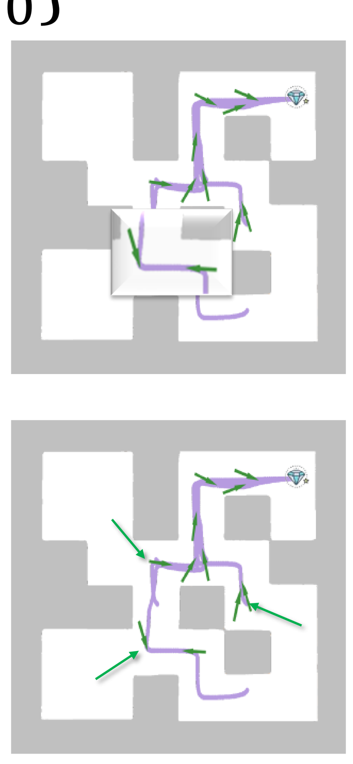
Beyond expert-observed area:

1) expand the target states G

$$Conf_\pi(s) = \pi(\mathbb{E}[\pi(a|s)]|s)$$
$$s_0 \sim G = \{s|Conf_\pi(s) \geq \mathbb{E}_{s'\sim D^E} Conf_\pi(s')\}$$
$$\cup \{s'|s' \in D^E\}$$

2) re-sample the augmented instances

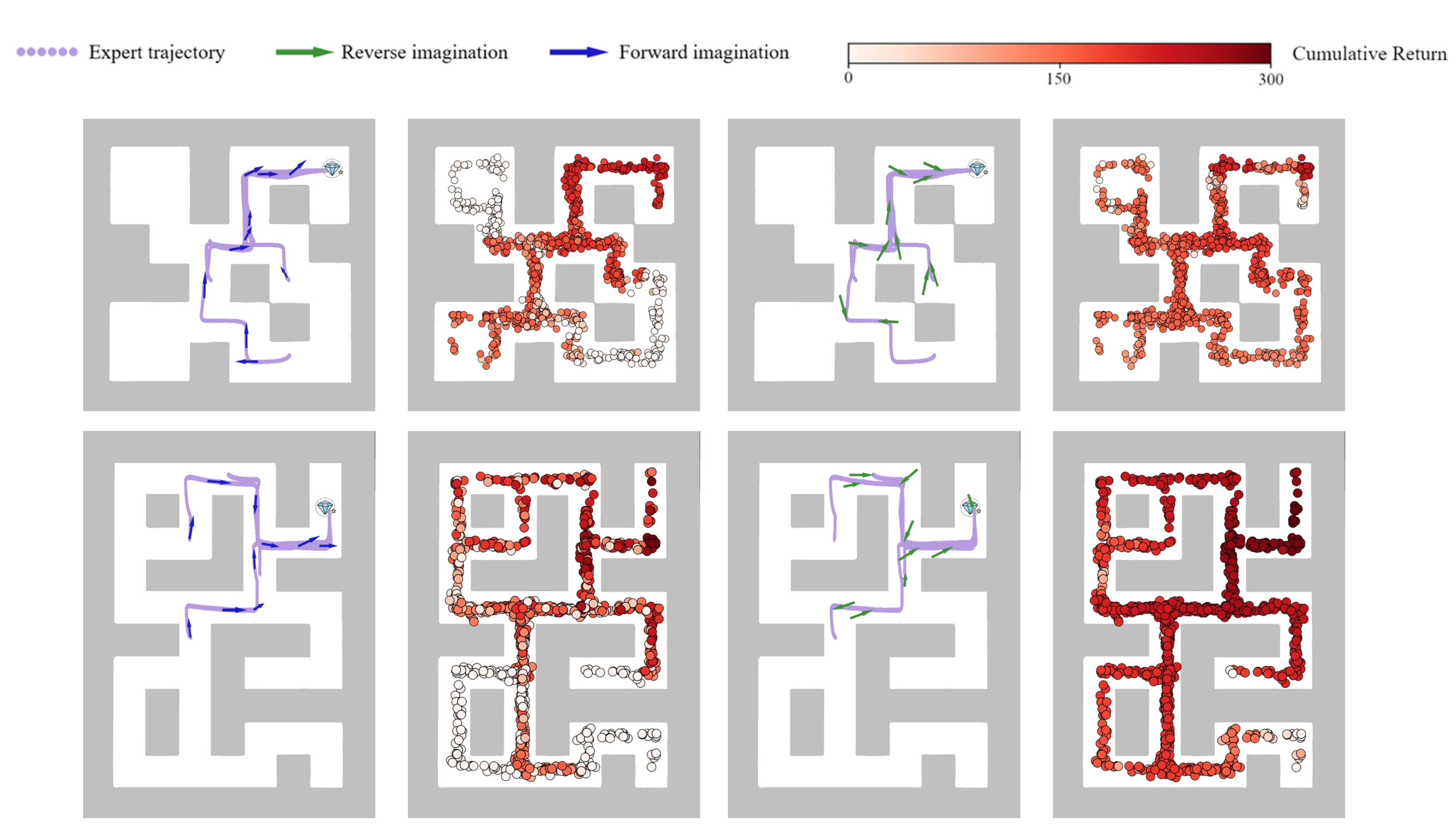
$$p(s) = 1/Conf_\pi(s)$$



Performance on Benchmarks

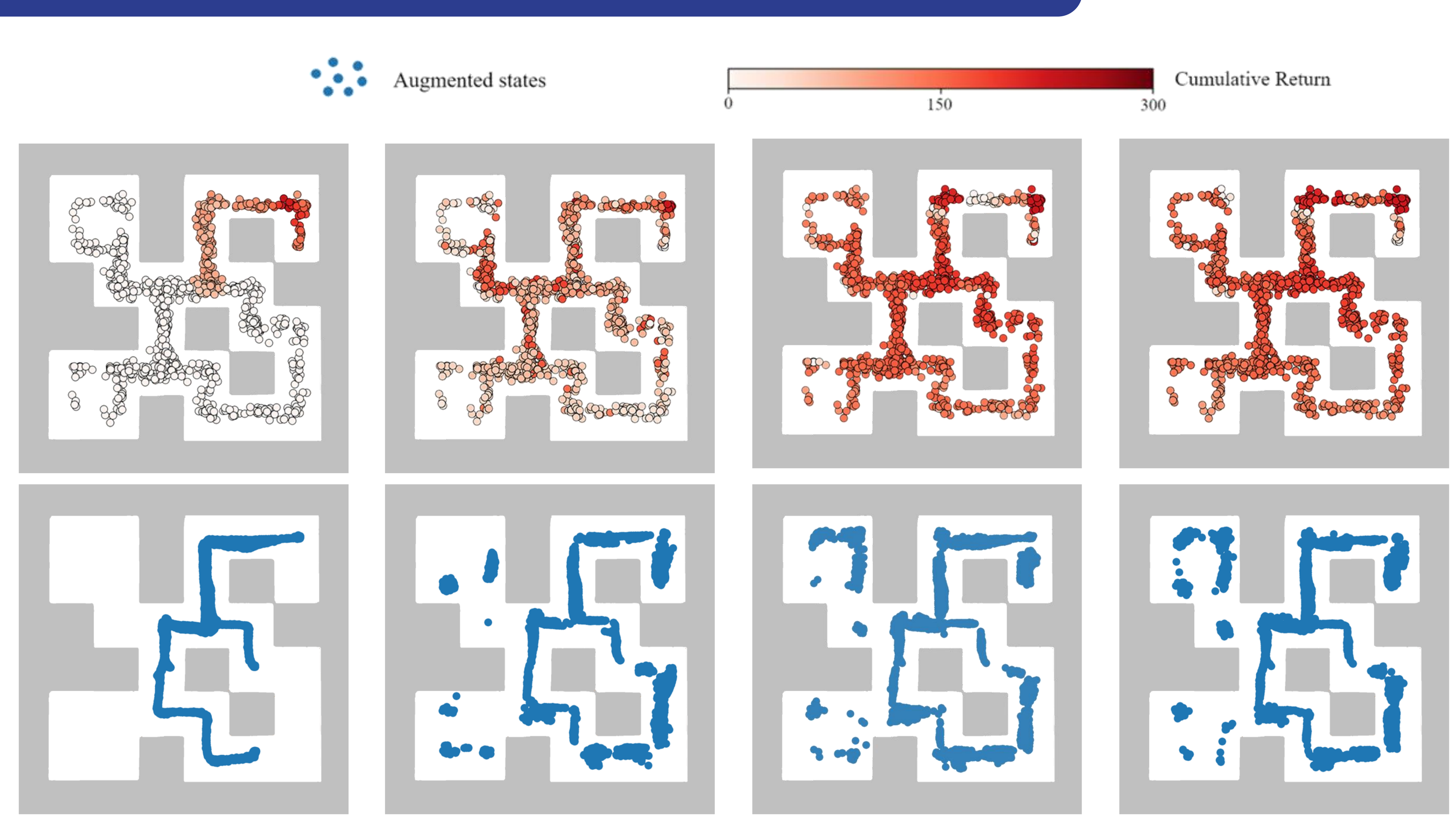
DataSet	BC-exp	DemoDICE	DWBC	OTIL	CLARE	MILO	ROMI	UDS	SRA
maze2d-umaze-sparse-v1	100.±11.6	88.7±10.4	25.8±5.65	128.±8.22	-3.08±6.02	75.0±11.2	154.±5.96	64.9±8.51	**155.±6.20**
maze2d-medium-sparse-v1	44.6±11.1	15.4±7.83	22.7±4.77	98.2±11.0	33.5±7.82	47.9±13.5	123.±10.5	83.0±8.84	**147.±5.67**
maze2d-large-sparse-v1	15.5±7.98	8.68±3.55	35.1±4.18	129.±14.4	18.6±9.12	51.2±17.1	101.±20.2	108.±16.7	**150.±14.9**
maze2d-umaze-dense-v1	70.6±9.55	69.1±9.21	39.2±4.77	100.±6.98	5.84±6.61	54.9±6.96	111.±6.23	62.3±6.99	**113.±5.80**
maze2d-medium-dense-v1	45.0±10.2	34.3±7.08	39.1±3.34	95.7±8.66	46.3±7.81	44.4±11.0	112.±9.10	87.3±7.80	**138.±5.29**
maze2d-large-dense-v1	18.2±8.57	21.7±6.30	56.1±5.56	120.±11.0	26.5±8.78	40.7±14.0	101.±16.6	109.±14.4	**140.±11.4**
hopper-medium	72.9±5.50	54.1±1.67	88.1±4.71	26.2±2.28	82.2±6.56	75.0±7.46	67.3±4.82	59.5±4.51	**90.2±4.93**
halfcheetah-medium	13.3±2.74	41.1±1.00	22.5±3.94	38.7±0.75	32.2±3.14	41.9±0.92	43.6±1.53	43.6±5.15	**43.7±1.72**
walker2d-medium	99.1±3.66	73.0±2.09	84.8±5.65	86.9±3.63	49.9±5.37	67.9±3.13	96.6±3.76	97.6±2.85	**101.±3.60**
ant-medium	51.3±6.87	91.2±3.79	37.5±5.95	72.4±5.68	68.5±7.35	92.0±3.55	**92.7±6.46**	87.3±5.10	88.9±7.18
hopper-medium-expert	72.9±5.50	98.6±4.32	99.4±4.43	42.5±3.70	93.9±5.81	90.9±5.42	100.±3.40	97.4±3.35	**104.±3.37**
halfcheetah-medium-expert	13.3±2.74	48.9±5.46	**82.3±3.79**	43.7±2.76	31.4±5.15	44.5±1.57	58.8±3.29	67.1±2.63	63.4±3.52
walker2d-medium-expert	99.1±3.66	93.1±5.49	**106.±1.57**	82.5±2.76	39.9±7.66	95.4±3.87	103.±2.12	103.±2.32	104.±4.88
ant-medium-expert	51.3±6.87	69.8±7.97	58.2±8.81	79.2±7.40	3.61±2.86	**115.±4.63**	105.±6.90	92.2±8.14	94.1±7.86

Forward Augmentation v.s. Reverse Augmentation



Forward Rollout MILO Reverse Rollout SRA

Self-paced Learning Process



State-wise cumulative return and Re-sampled augmented datasets after 0, 50000, 100000, and 150000 iterations

Scalability for Different RL Methods

DataSet	IQL	SRA+IQL	TD3BC	SRA+TD3BC	AWAC	SRA+AWAC	SAC-N	SRA+SAC-N
maze2d-umaze-sparse-v1	64.9±8.51	155.±6.20 ↑	38.1±12.9	145.±7.27 ↑	68.6±14.5	135.±10.1 ↑	151.±6.44	150.±6.47
maze2d-medium-sparse-v1	83.0±8.84	147.±5.67 ↑	22.4±9.64	140.±8.55 ↑	100.±13.5	87.8±16.4	147.±10.3	153.±7.36 ↑
maze2d-large-sparse-v1	108.±16.7	150.±14.9 ↑	57.9±13.2	143.±17.7 ↑	74.8±16.8	89.9±24.1 ↑	128.±20.7	158.±18.0 ↑
hopper-medium	59.5±4.51	90.2±4.93 ↑	57.2±1.90	96.4±5.74 ↑	38.3±3.88	85.6±6.04 ↑	3.33±0.49	107.±2.31 ↑
halfcheetah-medium	43.6±5.15	43.7±1.72 ↑	43.2±0.82	1.30±1.25	42.0±1.77	44.9±1.95 ↑	-.15±0.08	7.49±2.71 ↑
walker2d-medium	97.6±2.85	101.±3.60 ↑	89.6±3.40	103. ±3.58 ↑	90.8±5.91	94.3±4.46 ↑	4.19±0.42	86.5±7.72 ↑
ant-medium	87.3±5.10	88.9±7.18 ↑	90.4±5.42	42.0±8.30	57.0±8.39	82.2±9.29 ↑	-27.±3.40	47.2±7.43 ↑
Win/Tie/Loss	7/0/0		5/0/2		6/0/1		6/0/1	