Offline Reinforcement Learning with Reverse Model-based Imagination

Jianhao Wang*[1] Wenzhe Li*[1] Haozhe Jiang[1] Guangxiang Zhu[2] Siyuan Li[1] Chongjie Zhang[1]
[1]IIIS, Tsinghua University [2]Baidu Inc.

Motivation

Distributional shift is one of the main challenges in offline RL.
To address this problem, recent offline RL methods attempt to introduce **conservatism bias**.

Examples:
Model-free methods: BCQ, BEAR, BRAC, CQL, …

- Encode the bias into policy or value functions by using conservative regularizations or specially designed network structures.
- Constrained policy search can limit the generalization beyond the offline dataset.

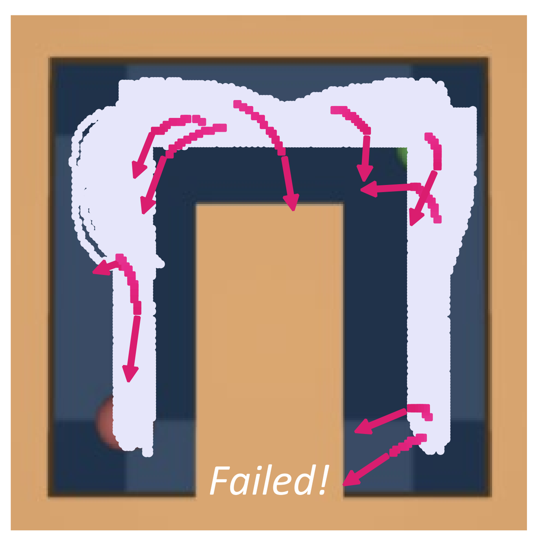
Dataset trajectory

BCQ's execution path

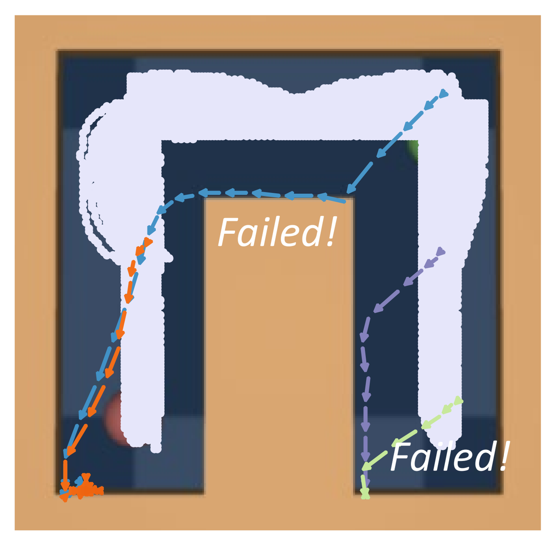
Model-based methods: MOPO, MOReL, Repb-SDE, …

- First learn a forward dynamics model from the offline dataset with conservatism quantifications, and then generate imaginary trajectories on high confidence regions to extend the offline dataset.
- Conservatism quantifications often suffer from overgeneralization and mislead forward model-based imaginations to undesired areas.

MOPO's model-uncertainty

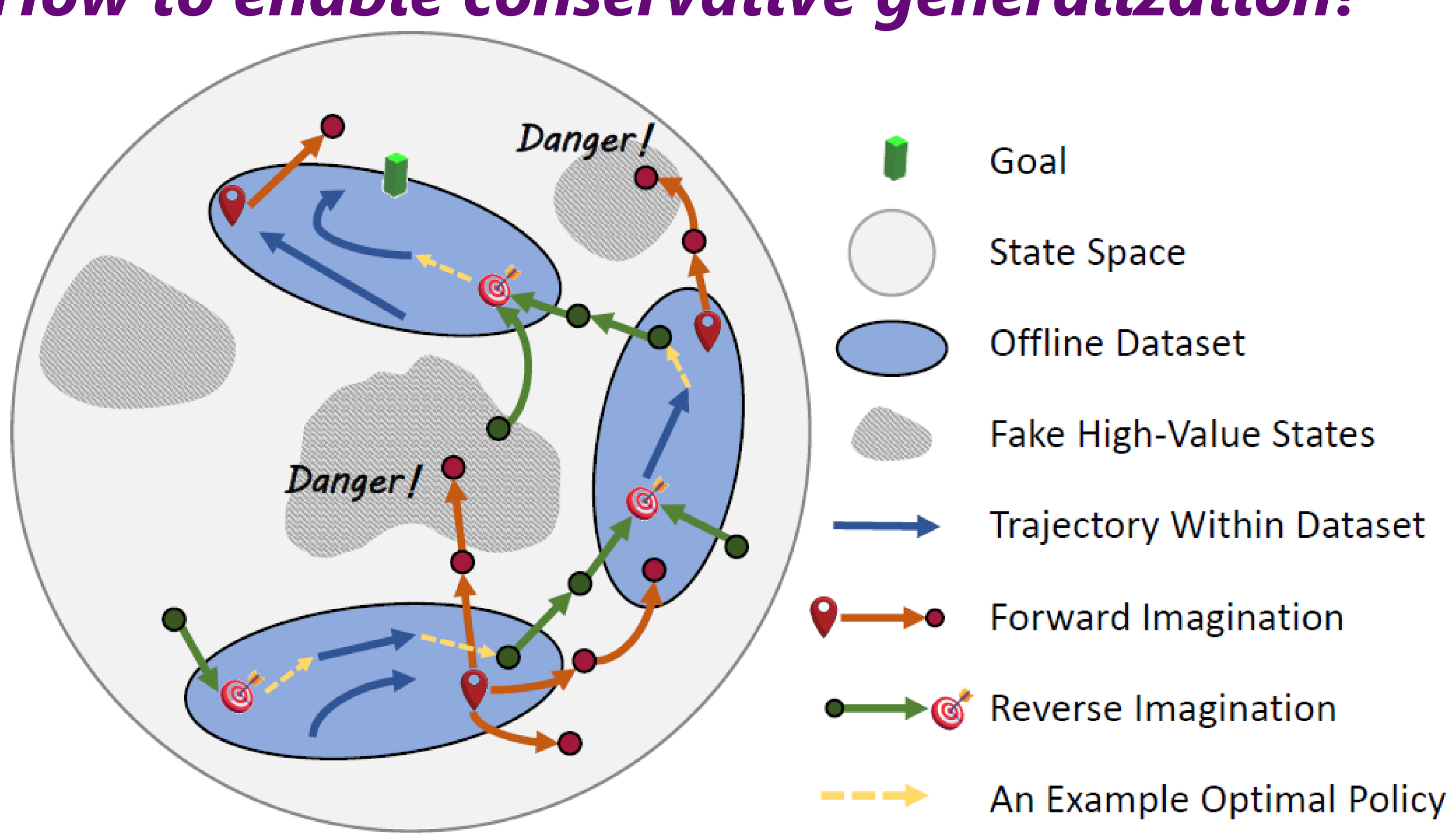


MOPO's imagination



MOPO's execution path

How to enable conservative generalization?



Reverse Offline Model-based Imagination

- The optimal policy requires a composition of multiple trajectories in the offline dataset.
- Forward imaginations potentially discover a better policy outside the offline dataset, but may also lead to undesirable regions consisting of fake high-value states due to overgeneralization errors.
- Reverse imaginations generate possible traces leading to target goal states ( ) inside the offline dataset, which provides a conservative way of augmenting the offline dataset.

Reverse model:

$$p(s,r|s',a) = p(s|s',a)p(r|s',a,s) = T_r(s|s',a)p(r|s,a)$$

$$\mathcal{L}_M(\phi) = \mathop{\mathbb{E}}_{(s,a,r,s')\sim\mathcal{D}_{\text{env}}}\left[-\log \widehat{p}_\phi(s,r|s',a)\right]$$

Reverse policy:

- Generative models: conditional VAE.
- Uniform policy.

Combination with model-free algorithms:

- ROMI provides informed data augmentation to extend the offline dataset.
- Since the reverse rollout policy is agnostic to policy learning, ROMI can be combined with any model-free offline RL algorithm

Experiments

Table 1: Performance of ROMI and best performance of prior methods on the *maze* and *antmaze* domains, on the normalized return metric proposed by D4RL benchmark [18]. Scores roughly range from 0 to 100, where 0 corresponds to a random policy performance and 100 corresponds to an expert policy performance. *med* is short for *medium*.

Environment	BC	ROMI-BCQ	MF	MB
sparse-maze2d-umaze	-3.2	**139.5** ± 3.6	65.7 ± 6.9$^{\text{BEAR}}$	76.4 ± 19.2$^{\text{COMBO}}$
sparse-maze2d-med	-0.5	**82.4** ± 15.2	70.6 ± 34.3$^{\text{BRAC-v}}$	68.5 ± 83.6$^{\text{COMBO}}$
sparse-maze2d-large	-1.7	**83.1** ± 22.1	**81.0** ± 65.3$^{\text{BEAR}}$	14.1 ± 10.7$^{\text{COMBO}}$
dense-maze2d-umaze	-6.9	**98.3** ± 2.5	51.5 ± 8.2$^{\text{BRAC-p}}$	94.3 ± 13.6$^{\text{Repb-SDE}}$
dense-maze2d-med	2.7	**102.6** ± 32.4	41.7 ± 2.0$^{\text{BAIL}}$	84.2 ± 9.5$^{\text{COMBO}}$
dense-maze2d-large	-0.3	**124** ± 1.3	**133.0** ± 25.5$^{\text{BEAR}}$	36.8 ± 12.4$^{\text{MOPO}}$
fixed-antmaze-umaze	82.0	68.7 ± 2.7	75.3 ± 13.7$^{\text{BCQ}}$	**80.3** ± 18.5$^{\text{COMBO}}$
play-antmaze-med	0.0	**35.3** ± 1.3	1.7 ± 1.0$^{\text{BAIL}}$	0.0
play-antmaze-large	0.0	**20.2** ± 14.8	2.2 ± 1.3$^{\text{BAIL}}$	0.0
diverse-antmaze-umaze	47.0	**61.2** ± 3.3	54.0 ± 15.0$^{\text{BAIL}}$	57.3 ± 33.6$^{\text{COMBO}}$
diverse-antmaze-med	0.0	27.3 ± 3.9	**61.5** ± 10.0$^{\text{CQL}}$	0.0
diverse-antmaze-large	0.0	**41.2** ± 4.2	1.0 ± 0.9$^{\text{BAIL}}$	0.0

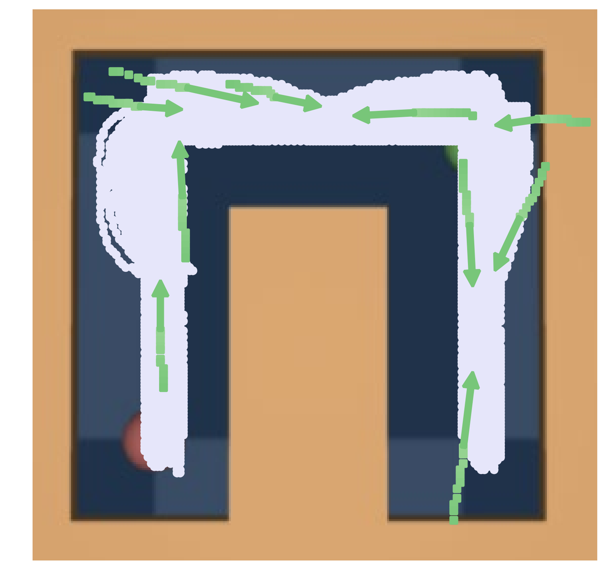
Table 2: Performance of ROMI and best performance of prior methods on Gym-MuJoCo tasks.

Environment	BC	ROMI-CQL	MF	MB
random-walker2d	0.0	7.5 ± 20.0	**11.1** ± 8.8$^{\text{BEAR}}$	7.0$^{\text{COMBO}}$
random-hopper	0.9	**30.2** ± 4.4	**31.4** ± 0.1$^{\text{CQL}}$	**31.7** ± 0.1$^{\text{Repb-SDE}}$
random-halfcheetah	-0.1	24.5 ± 0.7	19.6 ± 1.2$^{\text{CQL}}$	**38.8**$^{\text{COMBO}}$
medium-walker2d	41.7	**84.3** ± 1.1	**83.8** ± 0.2$^{\text{CQL}}$	**85.3** ± 2.2$^{\text{Repb-SDE}}$
medium-hopper	40.0	72.3 ± 17.5	66.6 ± 4.1$^{\text{CQL}}$	**95.4**$^{\text{MOReL}}$
medium-halfcheetah	39.2	49.1 ± 0.8	49.0 ± 0.4$^{\text{CQL}}$	**69.5** ± 0.0$^{\text{MOPO}}$
medium-replay-walker2d	2.2	**109.7** ± 9.8	88.4 ± 1.1$^{\text{CQL}}$	83.8 ± 7.6$^{\text{Repb-SDE}}$
medium-replay-hopper	8.1	**98.1** ± 2.6	**97.0** ± 0.8$^{\text{CQL}}$	93.6$^{\text{MOReL}}$
medium-replay-halfcheetah	25.6	47.0 ± 0.7	46.4 ± 0.3$^{\text{CQL}}$	**68.2** ± 3.2$^{\text{MOPO}}$
medium-expert-walker2d	73.4	**109.7** ± 5.3	**109.5** ± 0.1$^{\text{CQL}}$	**111.2** ± 0.2$^{\text{Repb-SDE}}$
medium-expert-hopper	36.0	**111.4** ± 5.6	**106.8** ± 2.9$^{\text{CQL}}$	**111.1**$^{\text{COMBO}}$
medium-expert-halfcheetah	39.7	86.8 ± 19.7	90.8 ± 5.6$^{\text{CQL}}$	**95.6**$^{\text{MOReL}}$

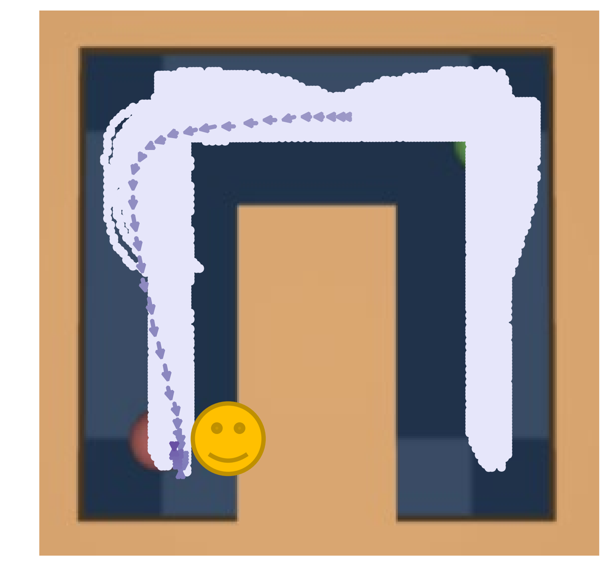
Reverse Imagination VS Forward Imagination

Table 3: Ablation study about ROMI with model-based imagination. Delta equals the improvement of ROMI-BCQ over BCQ on the normalized return metric.

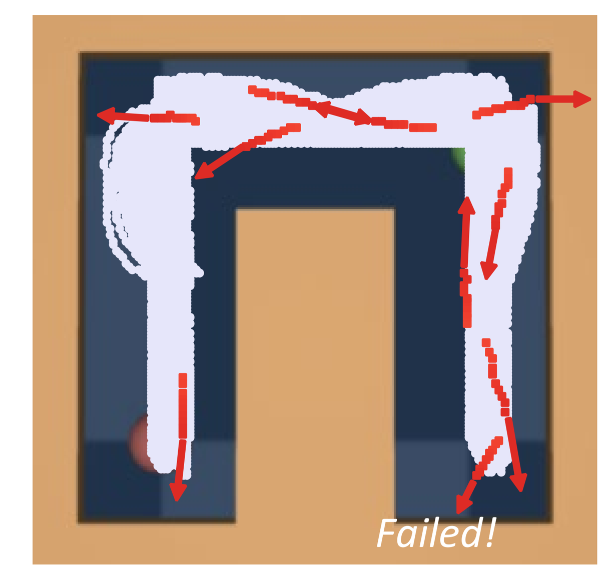
Dataset type	Environment	ROMI-BCQ (ours)	FOMI-BCQ	BCQ (base)	Delta
sparse	maze2d-umaze	**139.5** ± 3.6	8.1 ± 15.5	41.1 ± 7.6	98.4
sparse	maze2d-medium	82.4 ± 15.2	**93.6** ± 41.3	9.7 ± 14.2	72.7
sparse	maze2d-large	**83.1** ± 22.1	-2.5 ± 0.0	38.3 ± 10.4	44.8
dense	maze2d-umaze	**98.3** ± 2.5	30.7 ± 0.9	37.0 ± 5.3	61.3
dense	maze2d-medium	**102.6** ± 32.4	64.7 ± 37.0	37.9 ± 4.5	64.7
dense	maze2d-large	**124.0** ± 1.3	-0.7 ± 7.1	79.8 ± 12.2	44.2
fixed	antmaze-umaze	68.7 ± 2.7	**79.5** ± 2.5	75.3 ± 13.7	-6.6
play	antmaze-medium	**35.3** ± 1.3	26.2 ± 5.5	0.0	35.3
play	antmaze-large	**20.2** ± 14.8	12.0 ± 3.3	0.0	20.2
diverse	antmaze-umaze	61.2 ± 3.3	**66.8** ± 3.5	49.3 ± 9.9	11.9
diverse	antmaze-medium	**27.3** ± 3.9	12.3 ± 2.1	0.0	27.3
diverse	antmaze-large	**41.2** ± 4.2	17.8 ± 2.1	0.0	41.2



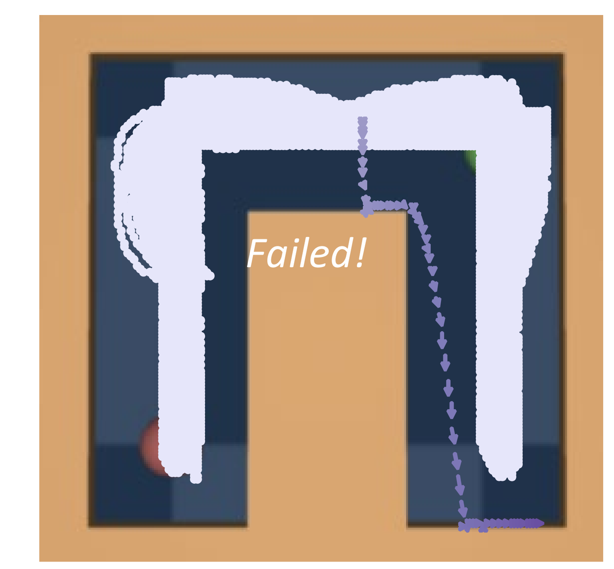
ROMI-BCQ's imagination



ROMI-BCQ's execution path



FOMI-BCQ's imagination



FOMI-BCQ's execution path

Conclusion

- We show that reverse imaginations could enable conservative generalization.
- ROMI provides a novel bidirectional learning paradigm for offline RL.
- We show that ROMI could achieve better or comparable performance to state-of-the-art baselines.

NeurIPS 2021 – December 2021, Online