# Offline RL with Reverse Model-based Imagination


Presenter: Wenzhe Li

Joint work with: Jianhao Wang, Haozhe Jiang, Guangxiang Zhu, Siyuan Li, Chongjie Zhang



NeurIPS 2021



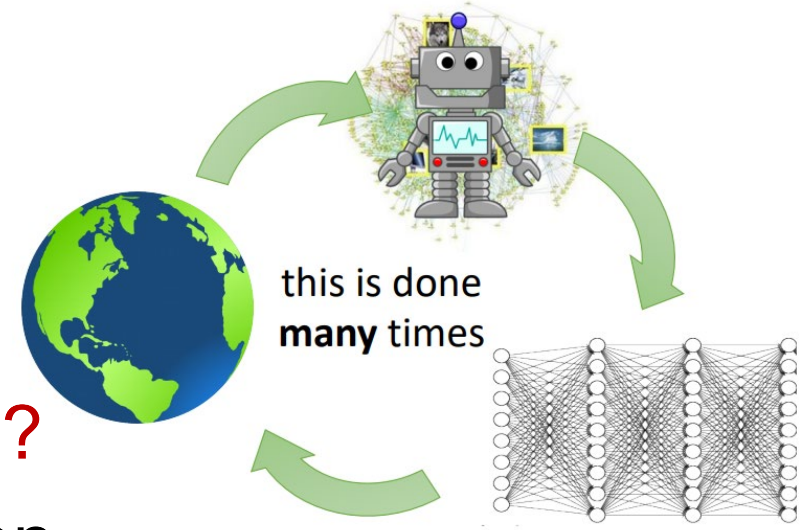
Machine Intelligence Group

清華大學 交叉信息研究院

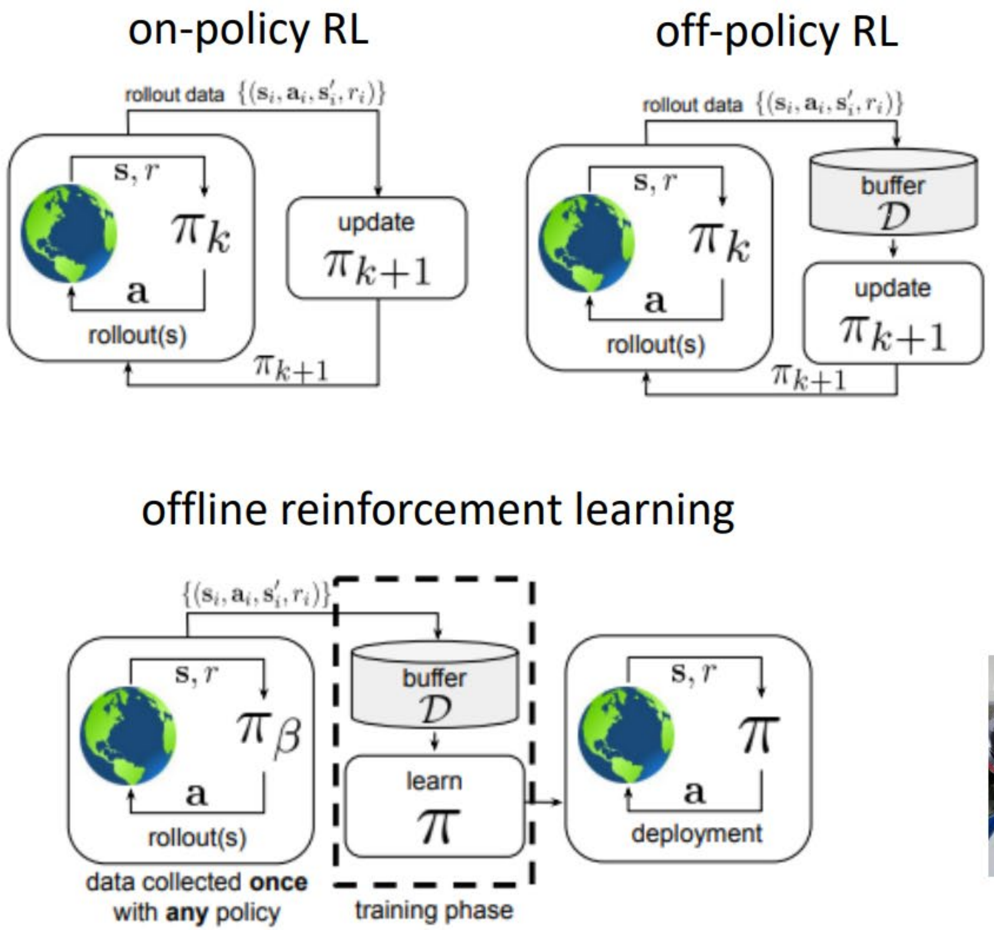
Tsinghua University Institute for Interdisciplinary Information Sciences

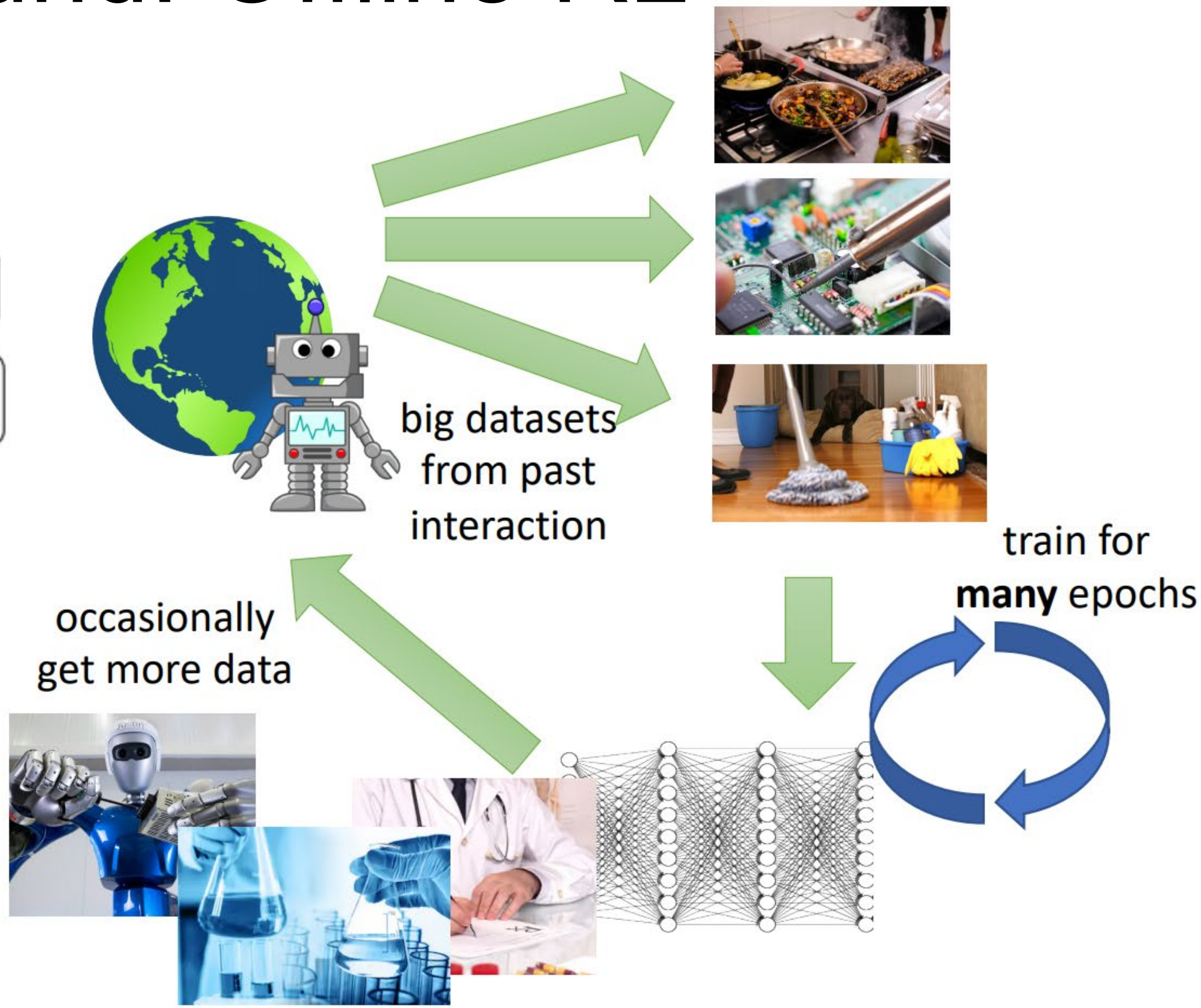
# Background: Offline RL

- The success of modern machine learning
  - Scalable data-driven learning methods
- Reinforcement learning
  - Online learning paradigm
  - Interaction is expensive & dangerous
  - Can we develop data-driven offline RL?
  - Healthcare, Robotics, Recommendation…



# Background: Offline RL

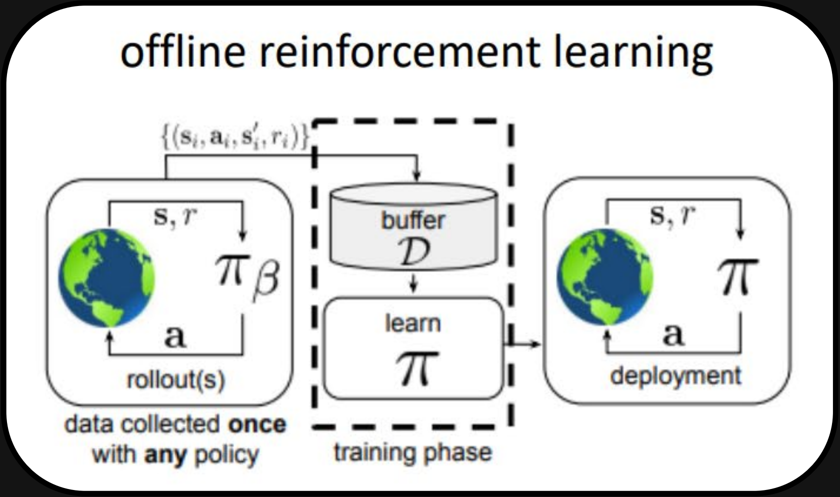


# Background: Offline RL

**Offline RL**

- the policy $\pi_k$ is updated with a static dataset $\mathcal{D}$, which is collected by unknown behavior policy $\pi_\beta$
- Interactions are not allowed

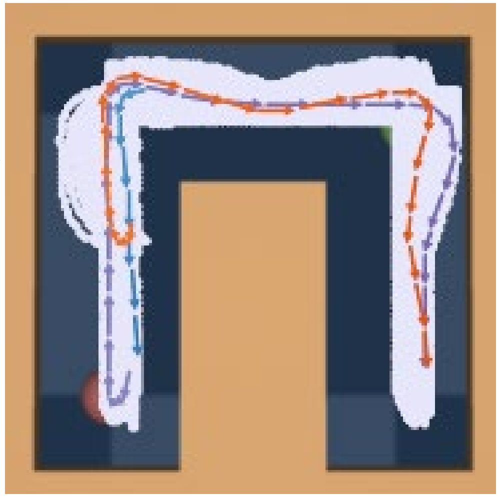



- $\mathcal{D} = \{(s_i, a_i, s'_i, r_i)\}$
- $s \sim d^{\pi_\beta}(s)$
- $a \sim \pi_\beta(a \mid s)$
- $s' \sim p(s' \mid s, a)$
- $r \leftarrow r(s, a)$
- Objective:

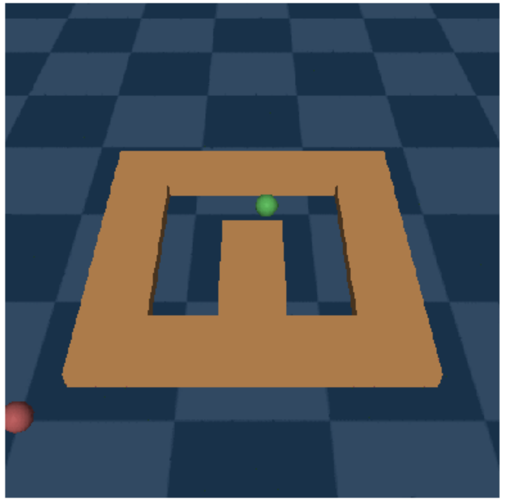
$$\max_{\pi} \sum_{t=0}^{T} E_{s_t \sim d^{\pi}(s), a_t \sim \pi(a|s)}[\gamma^t r(s_t, a_t)]$$

# Offline RL: Challenges

- Distributional shift
  - Learning with the dataset $\mathcal{D}$ only guarantees accurate predictions on the data distribution
- Common idea: conservatism
  - Model-free: stay inside the support
  - BCQ, BEAR, BRAC, CQL, …
  - Cons: overly conservative



Dataset support



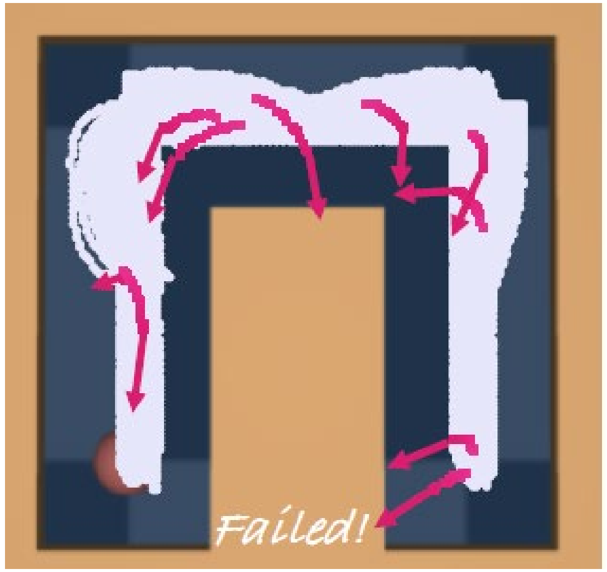
BCQ behavior

# Offline RL: Conservatism

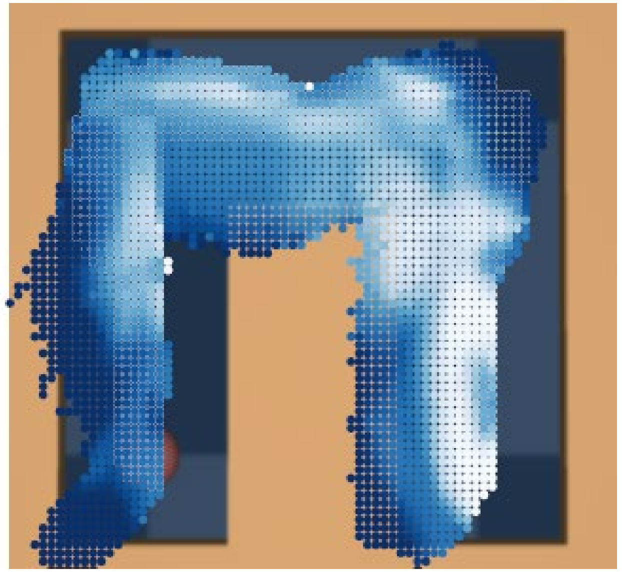
- Common idea: conservatism
  - Model-free: stay inside the support of the dataset distribution
  - Cons: overly conservative
  - Model-based: generalize beyond the dataset
  - MOPO, MOReL, Repb-SDE
  - Cons: uncertainty quantification

# Offline RL: Conservatism

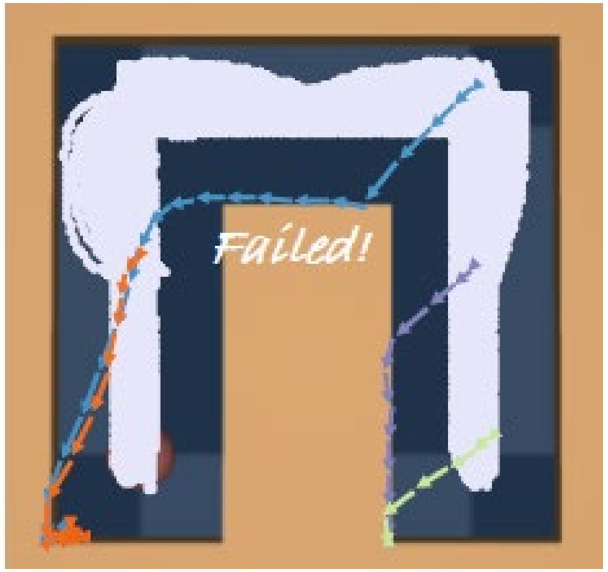
- Common idea: conservatism
  - Model-based: generalize beyond the dataset
  - Cons: uncertainty quantification




MOPO imagination

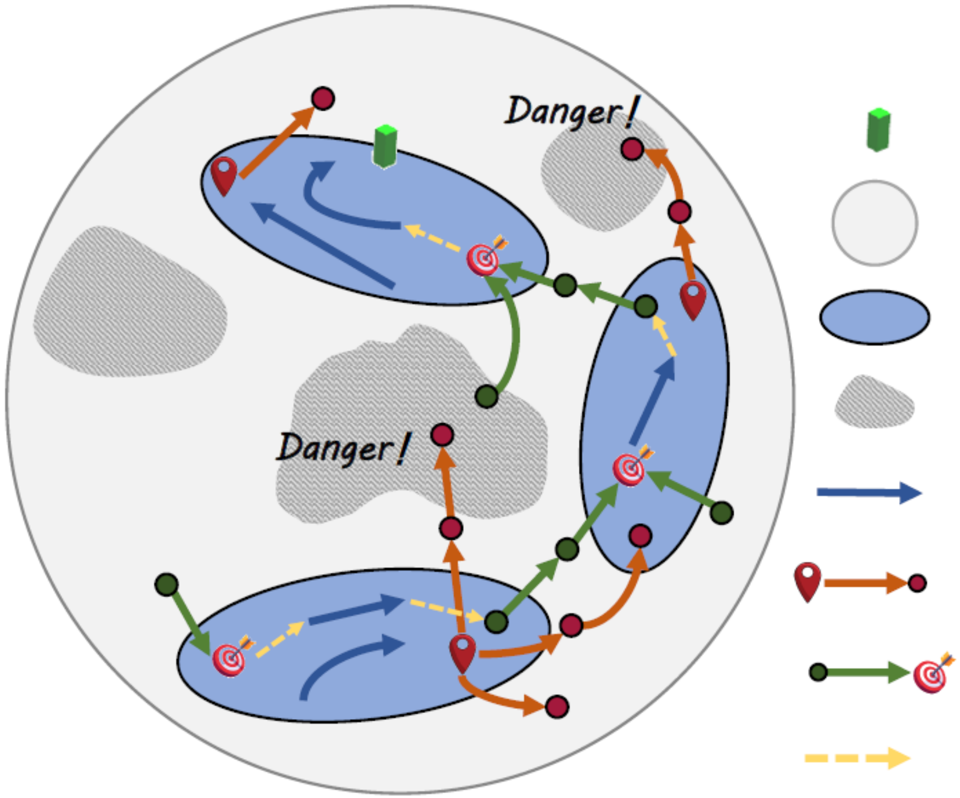


MOPO penalty




MOPO behavior

# Reverse Offline Model-based Imagination




**Algorithm 1** ROMI: Reverse Offline Model-based Imagination

**Require:** Offline dataset $\mathcal{D}_{\text{env}}$, rollout horizon $h$, the number of iterations $C_\phi, C_\theta, T$, learning rates $\alpha_\phi, \alpha_\theta$, model-free offline RL algorithm (i.e., BCQ or CQL)
Randomly initialize reverse model parameters $\phi$
**for** $i = 0 \ldots C_\phi - 1$ **do** $\triangleright$ Learning a reverse dynamics model $\widehat{p}_\phi$
  Compute $\mathcal{L}_M$ using the dataset $\mathcal{D}_{\text{env}}$
  Update $\phi \leftarrow \phi - \alpha_\phi \nabla_\phi \mathcal{L}_M$
Randomly initialize rollout policy parameters $\theta$
**for** $i = 0 \ldots C_\theta - 1$ **do** $\triangleright$ Learning a diverse rollout policy $\widehat{G}_\theta$
  Compute $\mathcal{L}_p$ using the dataset $\mathcal{D}_{\text{env}}$
  Update $\theta \leftarrow \theta - \alpha_\theta \nabla_\theta \mathcal{L}_p$
Initialize the replay buffer $\mathcal{D}_{\text{model}} \leftarrow \varnothing$
**for** $i = 0 \ldots T - 1$ **do** $\triangleright$ Collecting the replay buffer $\mathcal{D}_{\text{model}}$
  Sample target state $s_{t+1}$ from the dataset $\mathcal{D}_{\text{env}}$
  Generate reverse model rollout $\widehat{\tau} = \{(s_{t-i}, a_{t-i}, r_{t-i}, s_{t+1-i})\}_{i=0}^{h-1}$ from $s_{t+1}$ by drawing samples from the dynamics model $\widehat{p}_\phi$ and rollout policy $\widehat{G}_\theta$
  Add model rollouts to replay buffer, $\mathcal{D}_{\text{model}} \leftarrow \mathcal{D}_{\text{model}} \cup \{(s_{t-i}, a_{t-i}, r_{t-i}, s_{t+1-i})\}_{i=0}^{h-1}$
Compose the final dataset $\mathcal{D}_{\text{total}} \leftarrow \mathcal{D}_{\text{env}} \cup \mathcal{D}_{\text{model}}$
Combine model-free offline RL algorithms to derive the final policy $\pi_{\text{out}}$ using the dataset $\mathcal{D}_{\text{total}}$
**Return:** $\pi_{\text{out}}$

# ROMI: Components

- Reverse model

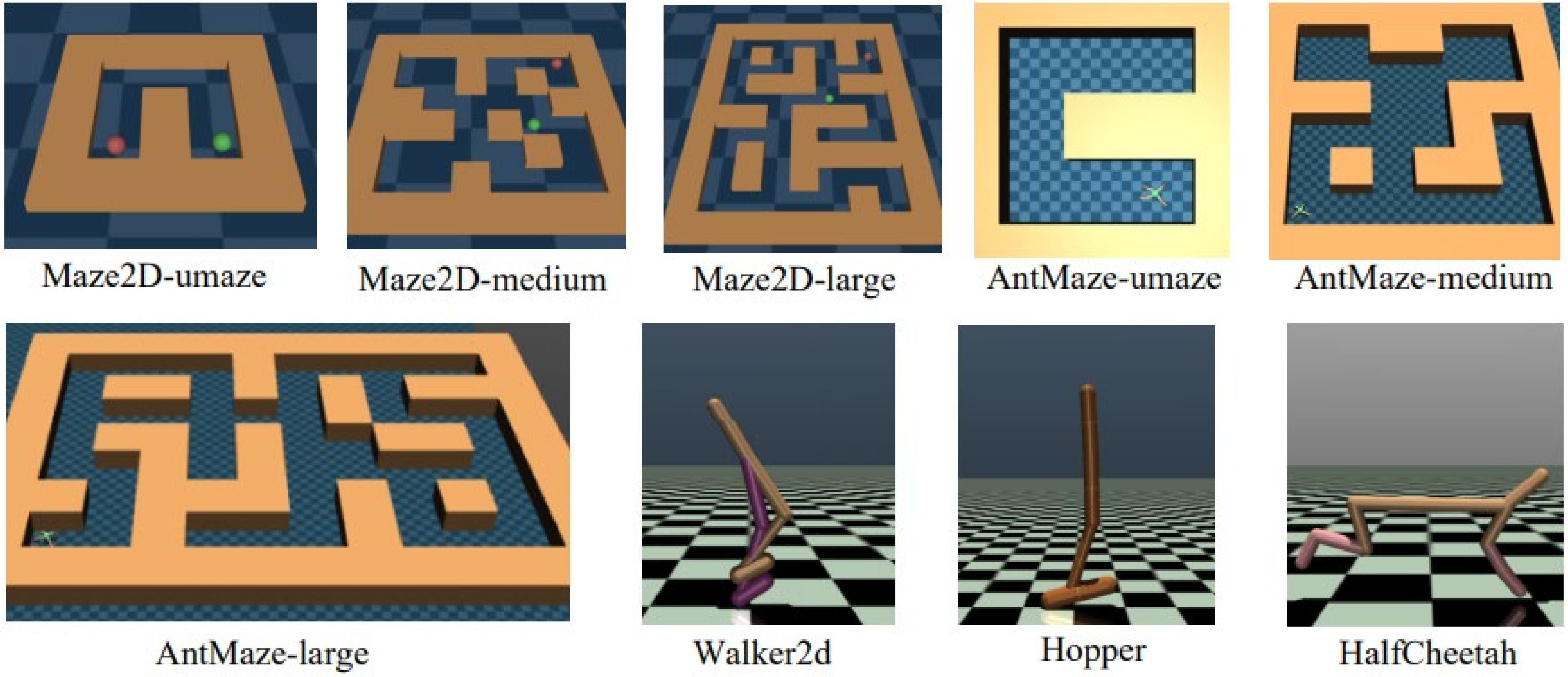
$$p(s,r|s',a) = p(s|s',a)p(r|s',a,s) = T_r(s|s',a)p(r|s,a)$$

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

- Reverse policy
  - Generative models: conditional VAE
  - Uniform policy

# Experiments: D4RL

- D4RL benchmark

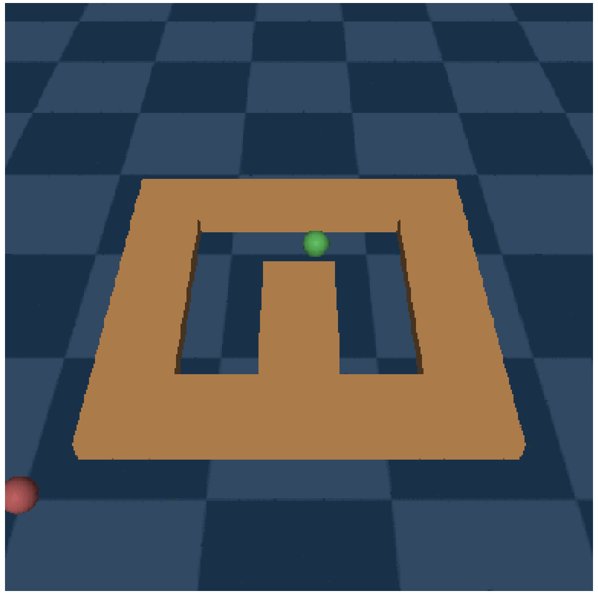


# Experiments: D4RL

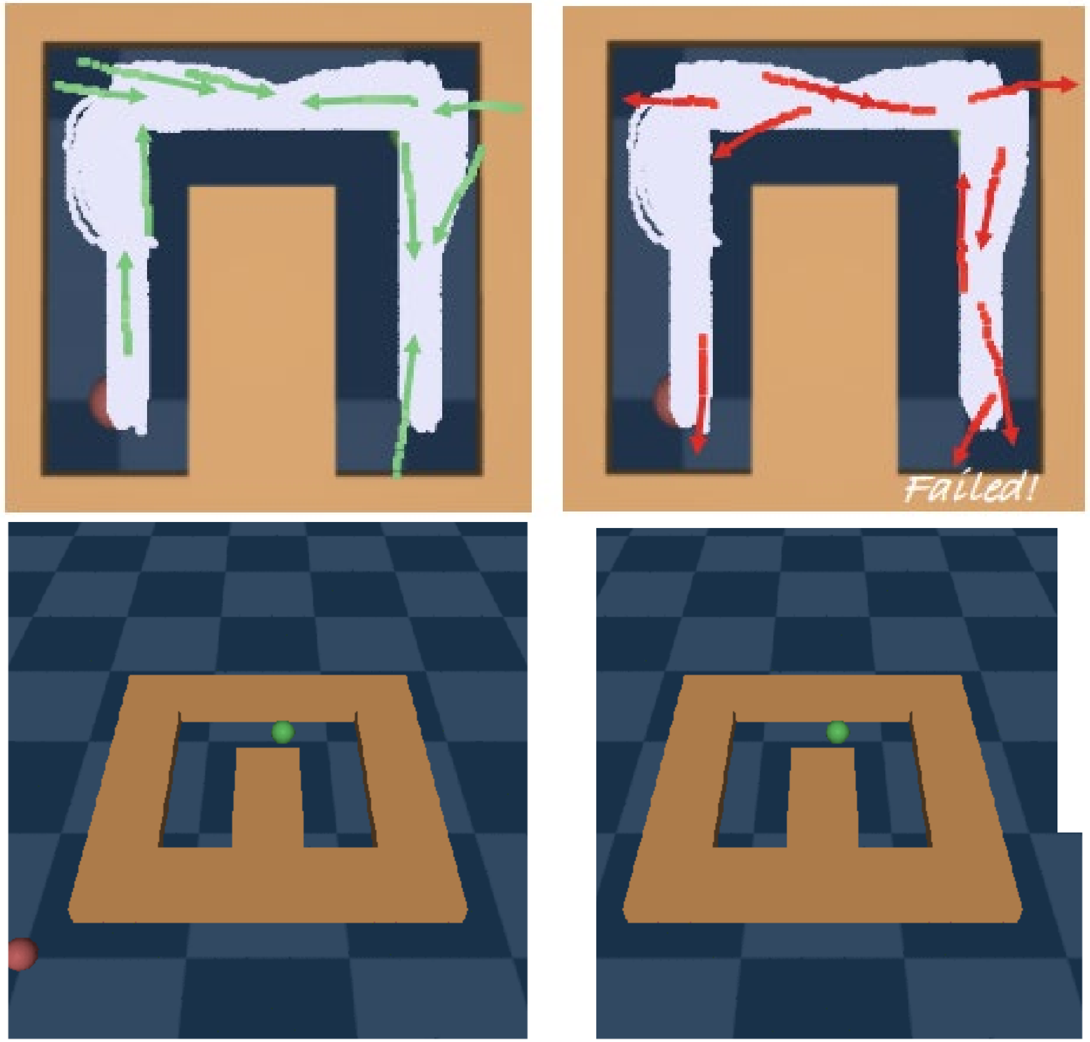
Table 1: Performance of ROMI and best performance of prior methods on the *maze* and *antmaze* domains, on the normalized return metric proposed by D4RL benchmark [18]. Scores roughly range from 0 to 100, where 0 corresponds to a random policy performance and 100 corresponds to an expert policy performance.

| Dataset type | Environment | BC | ROMI-BCQ (ours) | MF | MB |
|---|---|---|---|---|---|
| sparse | maze2d-umaze | -3.2 | **139.5** ± 3.6 | 65.7 [BEAR] | -12.6 [Repb-SDE] |
| sparse | maze2d-medium | -0.5 | **82.4** ± 15.2 | 70.6 [BRAC-v] | 21.7 [MOPO] |
| sparse | maze2d-large | -1.7 | **83.1** ± 22.1 | **81.0** [BEAR] | -1.5 [MOPO] |
| dense | maze2d-umaze | -6.9 | **98.3** ± 2.5 | 51.5 [BRAC-p] | 86.2 [Repb-SDE] |
| dense | maze2d-medium | 2.7 | **102.6** ± 32.4 | 39.5 [BAIL] | 69.9 [MOPO] |
| dense | maze2d-large | -0.3 | **124** ± 1.3 | **133.0** [BEAR] | 33.1 [MOPO] |
| fixed | antmaze-umaze | 82.0 | 68.7 ± 2.7 | **90.0** [BCQ] | 0.0 |
| play | antmaze-medium | 0.0 | **35.3** ± 1.3 | 7.5 [BAIL] | 0.0 |
| play | antmaze-large | 0.0 | **20.2** ± 14.8 | 2.0 [BAIL] | 0.0 |
| diverse | antmaze-umaze | 47.0 | **61.2** ± 3.3 | 52.0 [BAIL] | 0.0 |
| diverse | antmaze-medium | 0.0 | 27.3 ± 3.9 | **61.5** [CQL] | 0.0 |
| diverse | antmaze-large | 0.0 | **41.2** ± 4.2 | 2.0 [BAIL] | 0.0 |


ROMI imagination

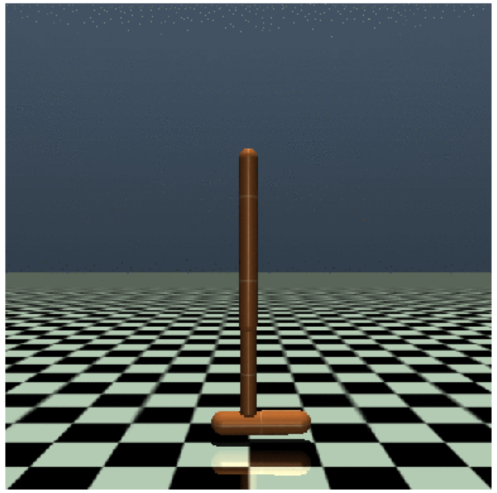

ROMI behavior

# Experiments: Ablation

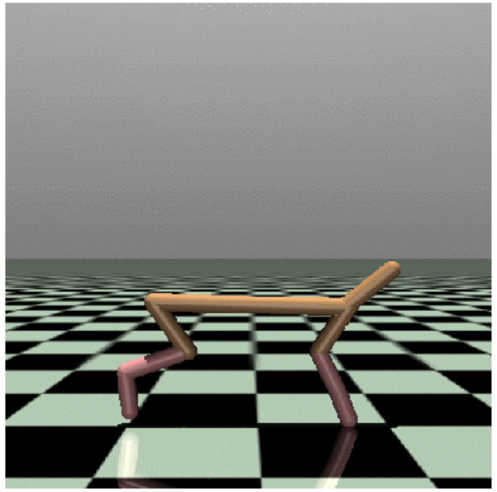



| Dataset type | Environment | ROMI-BCQ (ours) | FOMI-BCQ |
|---|---|---|---|
| sparse | maze2d-umaze | **139.5** ± 3.6 | 8.1 ± 15.5 |
| sparse | maze2d-medium | 82.4 ± 15.2 | **93.6** ± 41.3 |
| sparse | maze2d-large | **83.1** ± 22.1 | -2.5 ± 0.0 |
| dense | maze2d-umaze | **98.3** ± 2.5 | 30.7 ± 0.9 |
| dense | maze2d-medium | **102.6** ± 32.4 | 64.7 ± 37.0 |
| dense | maze2d-large | **124.0** ± 1.3 | -0.7 ± 7.1 |
| fixed | antmaze-umaze | 68.7 ± 2.7 | 79.5 ± 2.5 |
| play | antmaze-medium | **35.3** ± 1.3 | 26.2 ± 5.5 |
| play | antmaze-large | **20.2** ± 14.8 | 12.0 ± 3.3 |
| diverse | antmaze-umaze | 61.2 ± 3.3 | **66.8** ± 3.5 |
| diverse | antmaze-medium | **27.3** ± 3.9 | 12.3 ± 2.1 |
| diverse | antmaze-large | **41.2** ± 4.2 | 17.8 ± 2.1 |

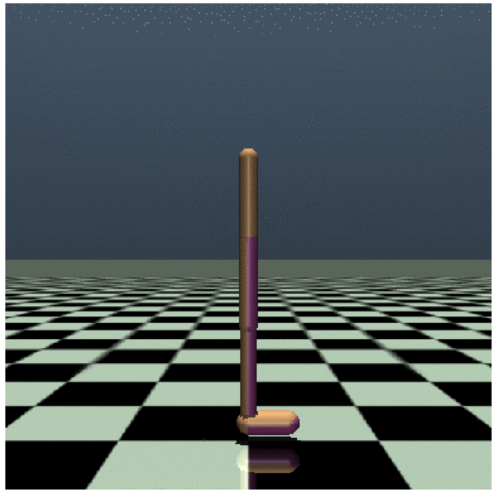
Reverse imaginations induce more conservative and effective behavior!



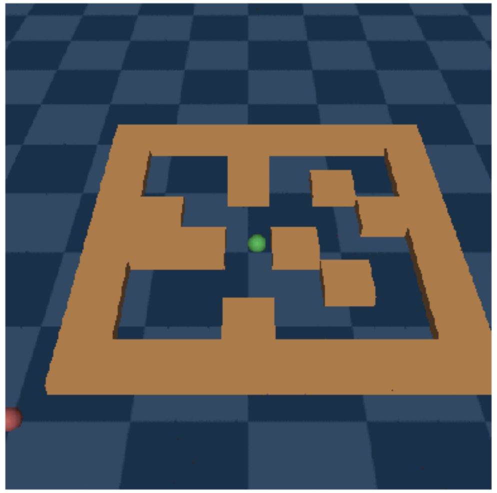
***hopper-random***



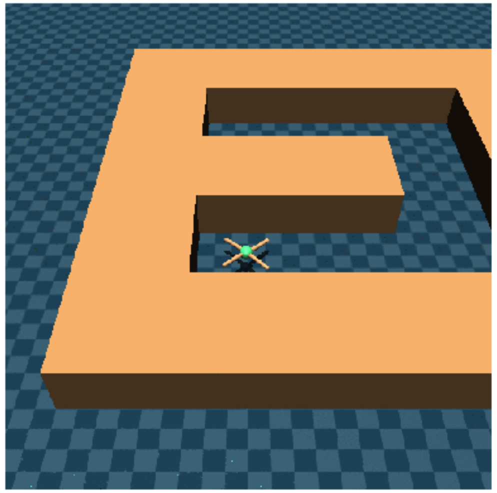
***halfcheetah-medium***



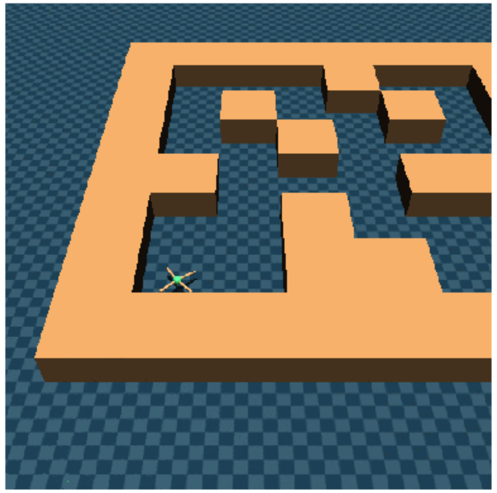
***walker2d-medium-replay***



***maze2d-medium***



***antmaze-umaze-diverse***



***antmaze-medium-diverse***

***antmaze-large-diverse***

***maze2d-large***

# Takeaway

- Reverse imaginations enable conservative generalization
- Bidirectional learning paradigm
  - Forward dataset trajectory
  - Reverse imaginary trajectory
- Better or comparable performance to state-of-the-art baselines

- More details
  - Paper & poster!

# Thanks for your listening