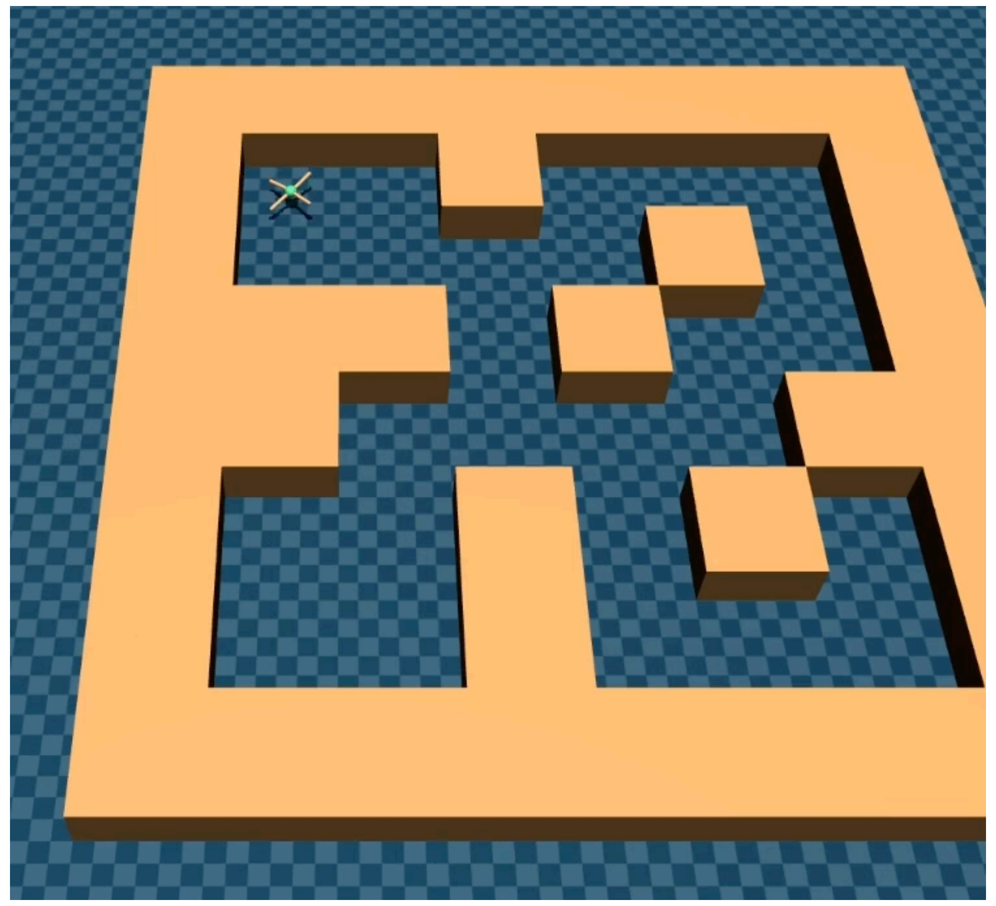
Entropy-regularized Diffusion Policy with Q-Ensembles for Offline RL

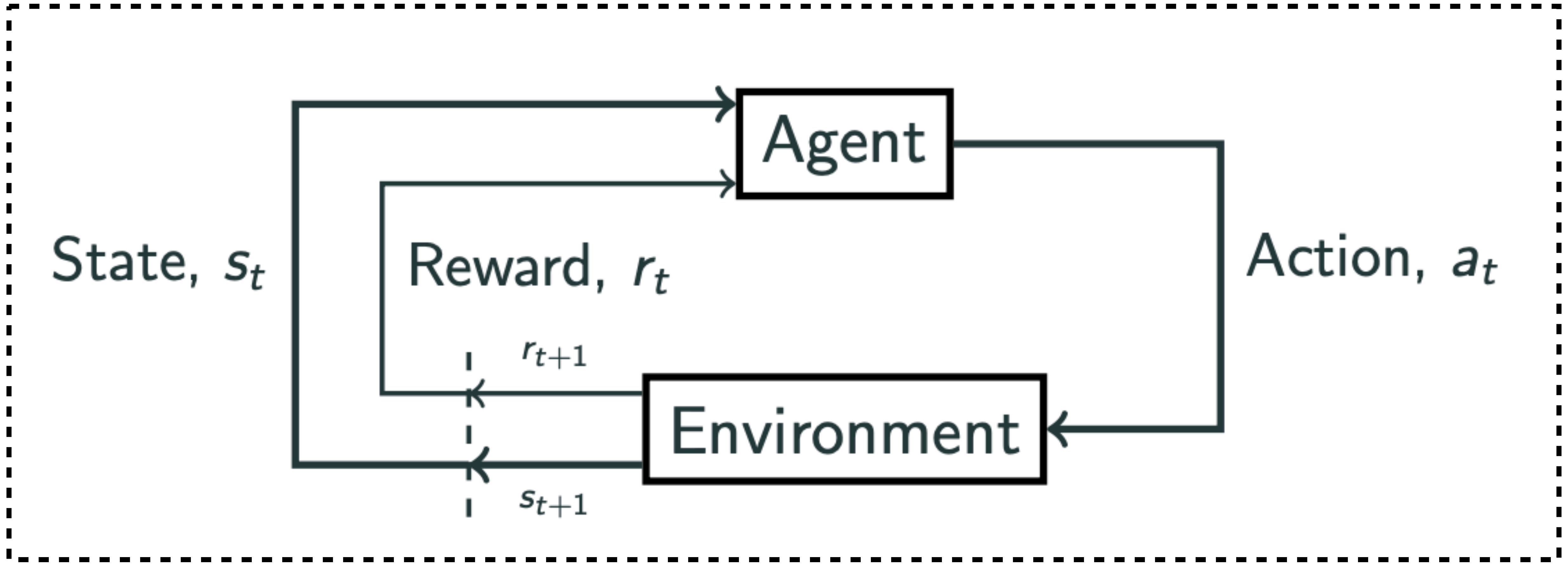
Ruoqi Zhang, Ziwei Luo, Jens Sjölund, Thomas Schön, Per Mattsson

NEURAL INFORMATION PROCESSING SYSTEMS



Offline Reinforcement Learning

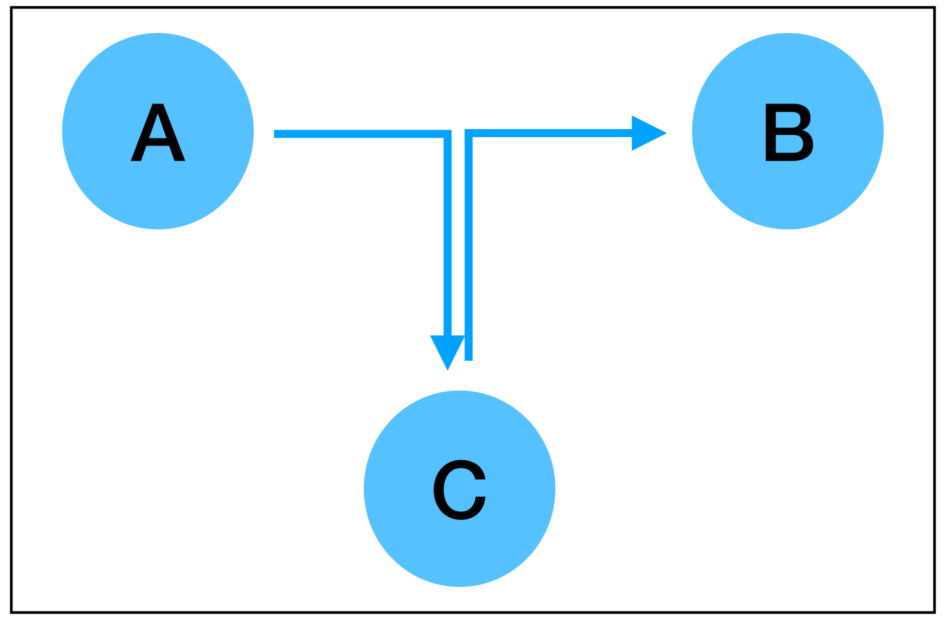
Learning from the dataset only

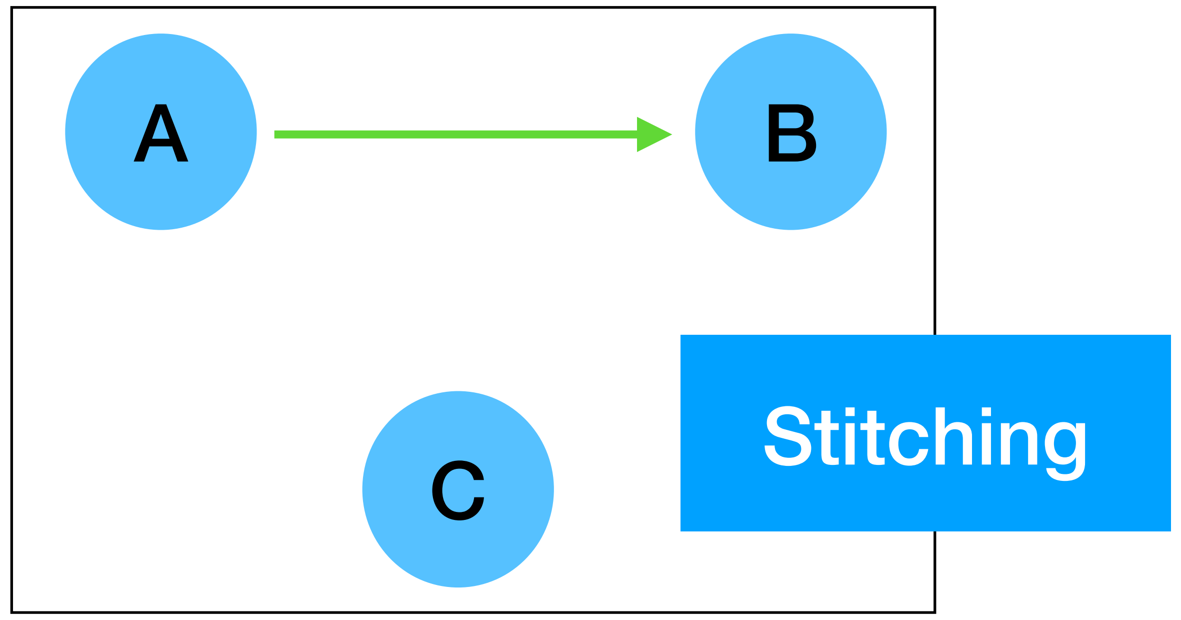


Dataset $\mathscr{D}$

Do better than the dataset!

Policy: $a_t = \pi_\theta(s_t) = \arg\max_a \mathbb{E}[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1}]$





UPPSALA
UNIVERSITET

Offline Reinforcement Learning

Multi-modality of the Dataset: t-SNE visualization

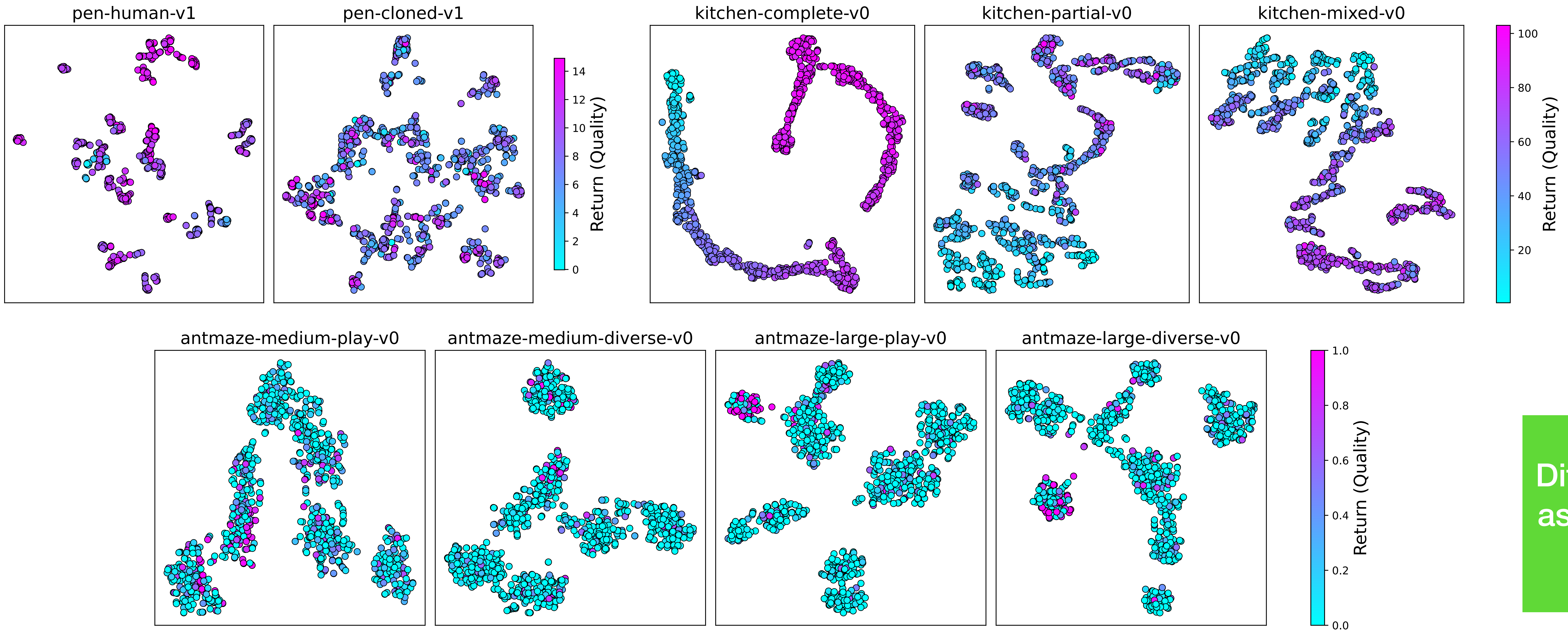


Figure 1. A t-SNE visualization of randomly selected 1000 states from Antmaze, Adroit and Kitchen domain. The color coding represents the return of the trajectory associated with each state.

Diffusion model as the behavior policy

- **Diffusion-QL [1]**: Diffusion Model as the behaviour policy

$$\pi = \arg\min_{\pi_\phi} L(\phi) = \mathscr{L}_d(\phi) + \mathscr{L}_q(\phi) = \mathscr{L}_d(\phi) - \lambda \cdot \mathbb{E}_{s\sim\mathscr{D}, a^0\sim\pi_\phi}[Q_\psi(s, a^0)]$$

[1] *Wang, Z., Hunt, J. J., & Zhou, M. (2022). Diffusion policies as an expressive policy class for offline reinforcement learning. arXiv preprint arXiv:2208.06193.*
[2] *Haarnoja, T., Zhou, A., Abbeel, P., & Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. ICML 2018.*

Out-of-distribution state & actions?

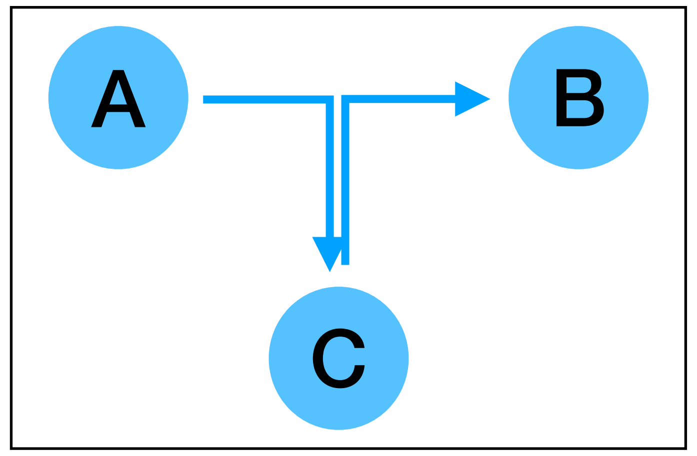
- **Diffusion-QL [1]**: Diffusion Model as the behaviour policy

$$\pi = \arg\min_{\pi_\phi} L(\phi) = \mathcal{L}_d(\phi) + \mathcal{L}_q(\phi) = \mathcal{L}_d(\phi) - \lambda \cdot \mathbb{E}_{s\sim\mathcal{D},a^0\sim\pi_\phi}[Q_\psi(s, a^0)]$$

[1] *Wang, Z., Hunt, J. J., & Zhou, M. (2022). Diffusion policies as an expressive policy class for offline reinforcement learning. arXiv preprint arXiv:2208.06193.*

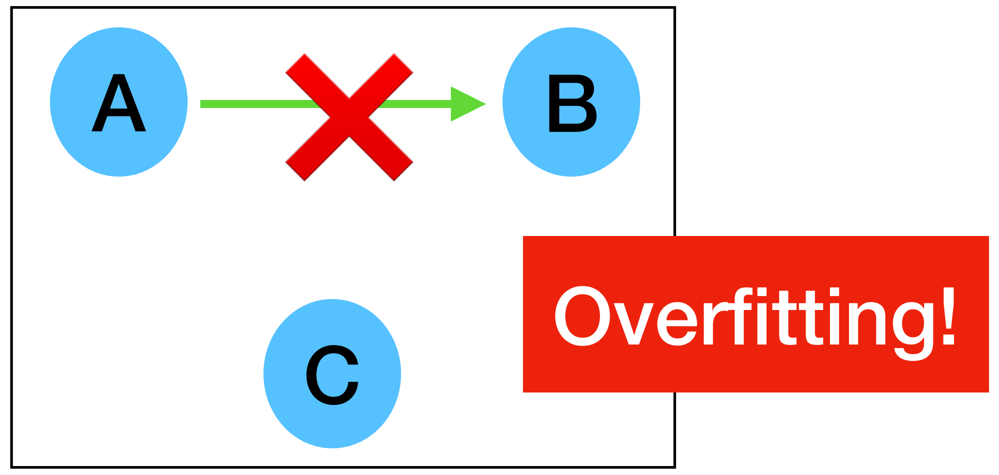
[2] *Haarnoja, T., Zhou, A., Abbeel, P., & Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. ICML 2018.*

Out-of-distribution state & actions?



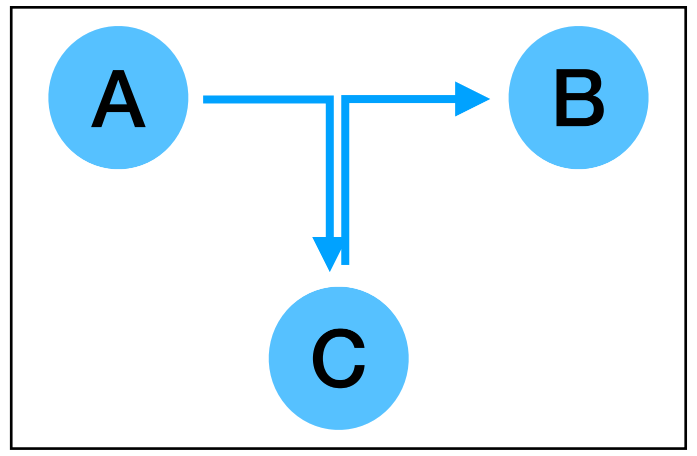
- **Diffusion-QL [1]**: Diffusion Model as the behaviour policy

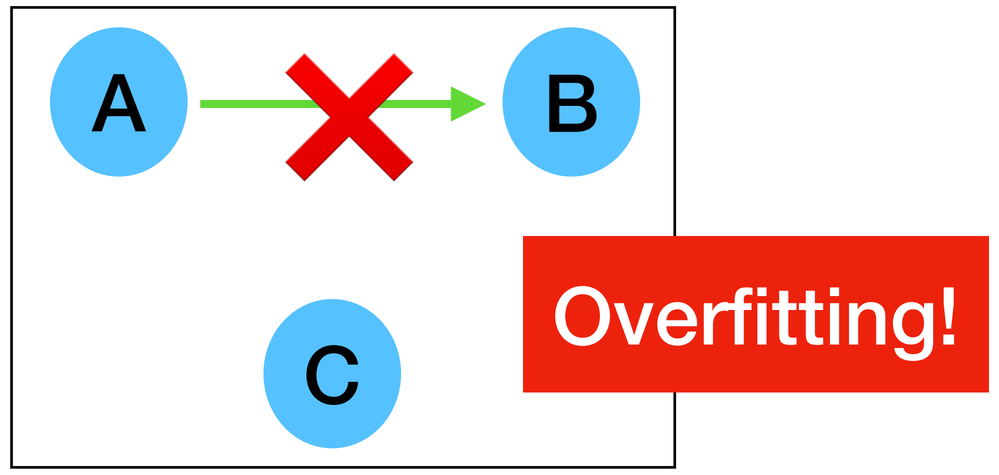
$$\pi = \arg\min_{\pi_\phi} L(\phi) = \mathscr{L}_d(\phi) + \mathscr{L}_q(\phi) = \mathscr{L}_d(\phi) - \lambda \cdot \mathbb{E}_{s\sim\mathscr{D}, a^0\sim\pi_\phi}[Q_\psi(s, a^0)]$$



[1] *Wang, Z., Hunt, J. J., & Zhou, M. (2022). Diffusion policies as an expressive policy class for offline reinforcement learning. arXiv preprint arXiv:2208.06193.*
[2] *Haarnoja, T., Zhou, A., Abbeel, P., & Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. ICML 2018.*

Out-of-distribution state & actions?





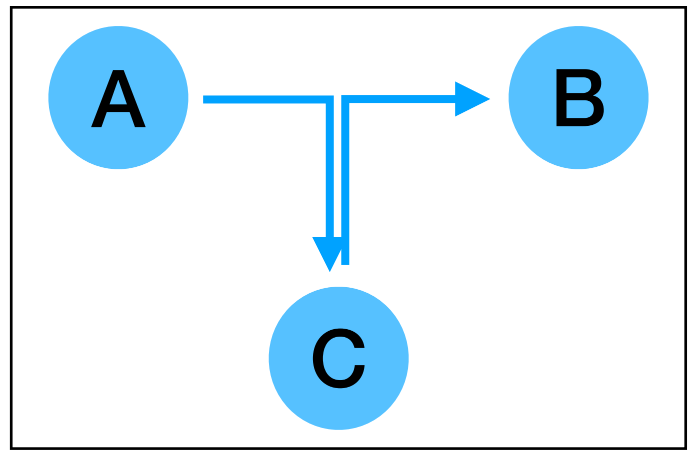
- **Diffusion-QL [1]**: Diffusion Model as the behaviour policy

$$\pi = \arg\min_{\pi_\phi} L(\phi) = \mathscr{L}_d(\phi) + \mathscr{L}_q(\phi) = \mathscr{L}_d(\phi) - \lambda \cdot \mathbb{E}_{s\sim\mathscr{D}, a^0\sim\pi_\phi}[Q_\psi(s, a^0)]$$

- **Our method:**

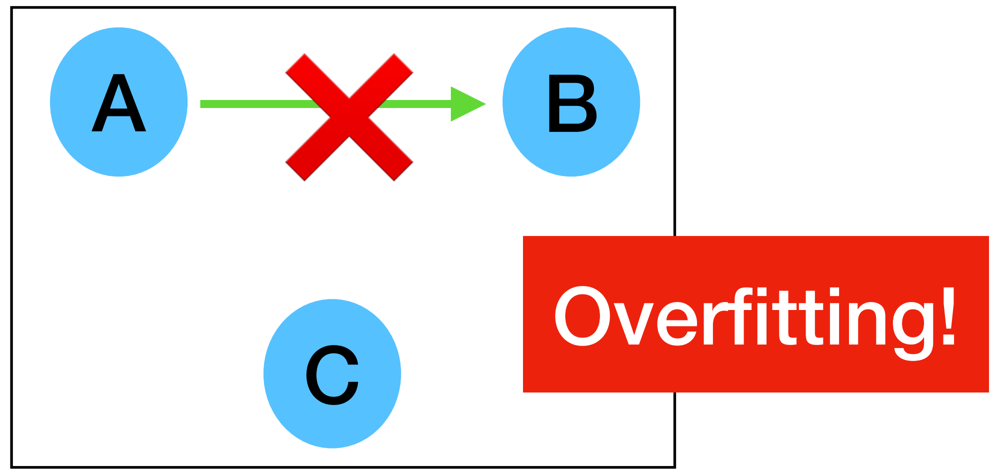
[1] *Wang, Z., Hunt, J. J., & Zhou, M. (2022). Diffusion policies as an expressive policy class for offline reinforcement learning. arXiv preprint arXiv:2208.06193.*
[2] *Haarnoja, T., Zhou, A., Abbeel, P., & Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. ICML 2018.*

Out-of-distribution state & actions?



- **Diffusion-QL [1]**: Diffusion Model as the behaviour policy

$$\pi = \arg\min_{\pi_\phi} L(\phi) = \mathcal{L}_d(\phi) + \mathcal{L}_q(\phi) = \mathcal{L}_d(\phi) - \lambda \cdot \mathbb{E}_{s\sim\mathcal{D}, a^0\sim\pi_\phi}[Q_\psi(s, a^0)]$$



- **Our method:**

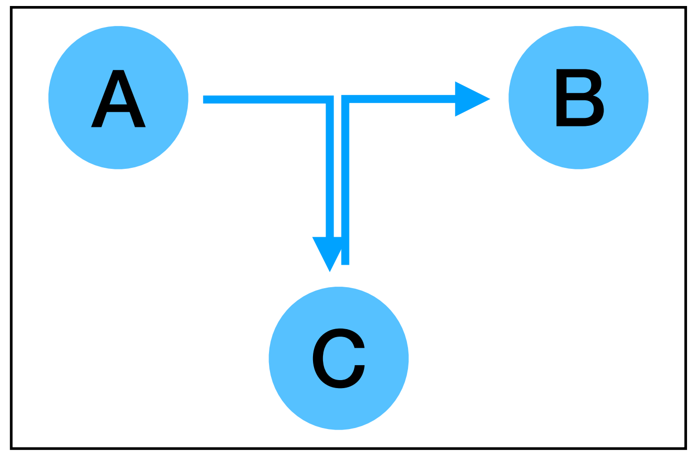
$$\pi = \arg\min_{\pi_\phi} L(\phi) = \mathcal{L}_d(\phi) + \mathcal{L}_q(\phi)$$

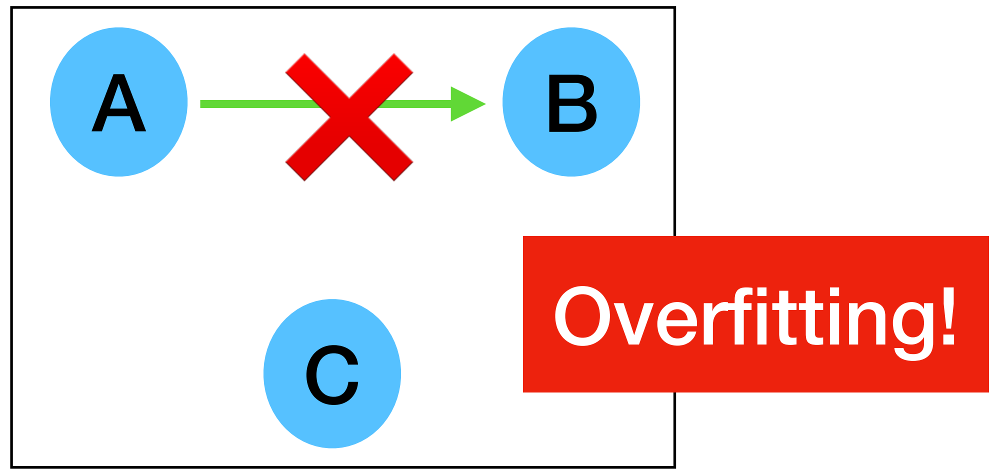
$$= \mathcal{L}_d(\phi) - \lambda \cdot \mathbb{E}_{s\sim\mathcal{D}, (a^0, \hat{a}_i^1)\sim\pi_\phi}[Q_\psi(s, a^0) - \alpha \log(p(\hat{a}_i^1 \,|\, a_i^T, s_i))]$$

Entropy Regularization

[1] Wang, Z., Hunt, J. J., & Zhou, M. (2022). Diffusion policies as an expressive policy class for offline reinforcement learning. arXiv preprint arXiv:2208.06193.
[2] Haarnoja, T., Zhou, A., Abbeel, P., & Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. ICML 2018.

Out-of-distribution state & actions?





- **Diffusion-QL [1]**: Diffusion Model as the behaviour policy

$$\pi = \arg\min_{\pi_\phi} L(\phi) = \mathcal{L}_d(\phi) + \mathcal{L}_q(\phi) = \mathcal{L}_d(\phi) - \lambda \cdot \mathbb{E}_{s\sim\mathcal{D}, a^0\sim\pi_\phi}[Q_\psi(s, a^0)]$$

- **Our method:**

SAC[2]: $\pi^* = \arg\max_{\pi} \sum_t \mathbb{E}_{(\mathbf{s}_t,\mathbf{a}_t)\sim\rho_\pi}\left[r\left(\mathbf{s}_t, \mathbf{a}_t\right) + \alpha\mathcal{H}\left(\pi\left(\cdot \mid \mathbf{s}_t\right)\right)\right]$

$= \arg\min J_\pi(\phi) = \mathbb{E}_{\mathbf{s}_t\sim\mathcal{D}}\left[\mathbb{E}_{\mathbf{a}_t\sim\pi_\phi}\left[\alpha\log\left(\pi_\phi\left(\mathbf{a}_t \mid \mathbf{s}_t\right)\right) - Q_\psi\left(\mathbf{s}_t, \mathbf{s}_t\right)\right]\right]$

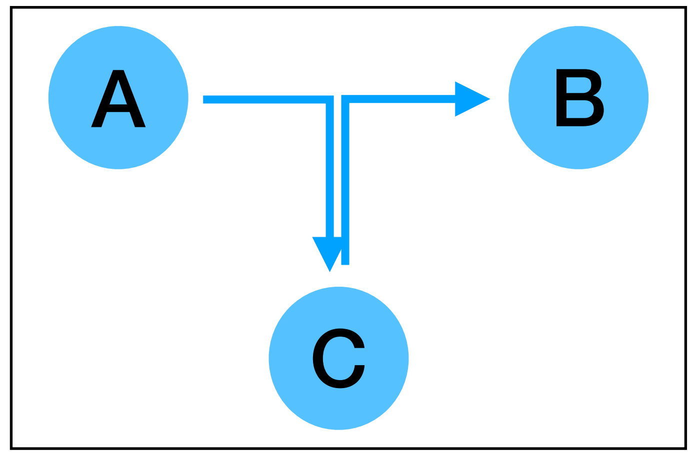
$$\pi = \arg\min_{\pi_\phi} L(\phi) = \mathcal{L}_d(\phi) + \mathcal{L}_q(\phi)$$

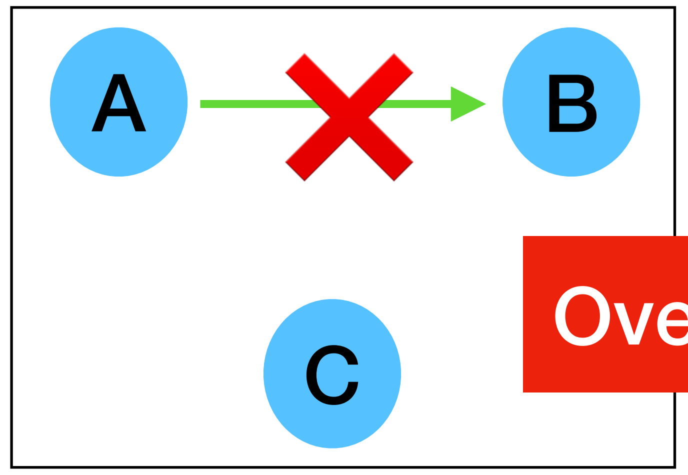
$$= \mathcal{L}_d(\phi) - \lambda \cdot \mathbb{E}_{s\sim\mathcal{D},(a^0,\hat{a}_i^1)\sim\pi_\phi}[Q_\psi(s, a^0) - \alpha\log(p(\hat{a}_i^1 \mid a_i^T, s_i))]$$

Entropy Regularization

[1] *Wang, Z., Hunt, J. J., & Zhou, M. (2022). Diffusion policies as an expressive policy class for offline reinforcement learning. arXiv preprint arXiv:2208.06193.*
[2] *Haarnoja, T., Zhou, A., Abbeel, P., & Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. ICML 2018.*

Out-of-distribution state & actions?





Overfitting!

- **Diffusion-QL [1]**: Diffusion Model as the behaviour policy

$$\pi = \arg\min_{\pi_\phi} L(\phi) = \mathscr{L}_d(\phi) + \mathscr{L}_q(\phi) = \mathscr{L}_d(\phi) - \lambda \cdot \mathbb{E}_{s\sim\mathscr{D}, a^0\sim\pi_\phi}[Q_\psi(s, a^0)]$$

- **Our method:**

$$\pi = \arg\min_{\pi_\phi} L(\phi) = \mathscr{L}_d(\phi) + \mathscr{L}_q(\phi)$$

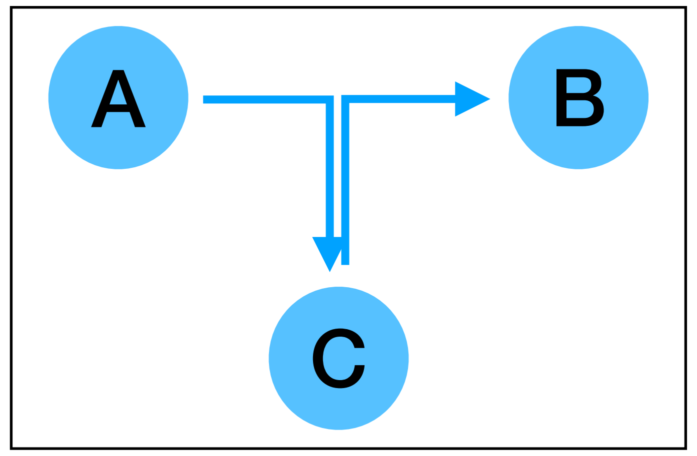
$$= \mathscr{L}_d(\phi) - \lambda \cdot \mathbb{E}_{s\sim\mathscr{D}, (a^0, \hat{a}_i^1)\sim\pi_\phi}[Q_\psi(s, a^0) - \alpha \log(p(\hat{a}_i^1 \mid a_i^T, s_i))]$$

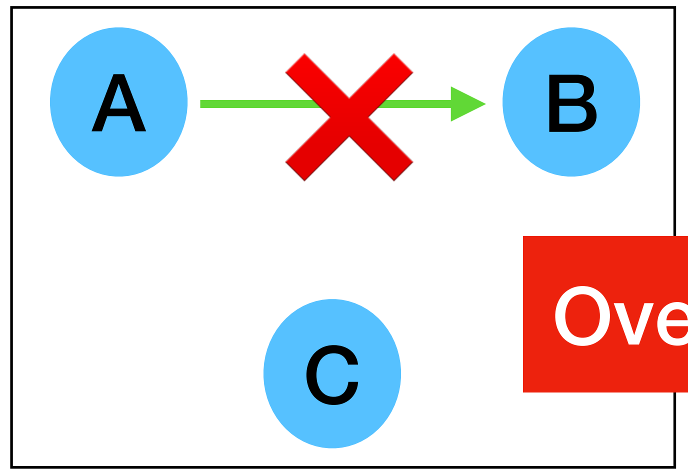
Entropy Regularization

LCB of Q-ensemble

[1] *Wang, Z., Hunt, J. J., & Zhou, M. (2022). Diffusion policies as an expressive policy class for offline reinforcement learning. arXiv preprint arXiv:2208.06193.*
[2] *Haarnoja, T., Zhou, A., Abbeel, P., & Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. ICML 2018.*

Out-of-distribution state & actions?





Overfitting!

- **Diffusion-QL [1]**: Diffusion Model as the behaviour policy

$$\pi = \arg\min_{\pi_\phi} L(\phi) = \mathscr{L}_d(\phi) + \mathscr{L}_q(\phi) = \mathscr{L}_d(\phi) - \lambda \cdot \mathbb{E}_{s\sim\mathscr{D}, a^0\sim\pi_\phi}[Q_\psi(s, a^0)]$$

- **Our method:**

$$\pi = \arg\min_{\pi_\phi} L(\phi) = \mathscr{L}_d(\phi) + \mathscr{L}_q(\phi)$$

$$= \mathscr{L}_d(\phi) - \lambda \cdot \mathbb{E}_{s\sim\mathscr{D}, (a^0, \hat{a}_i^1)\sim\pi_\phi}[Q_\psi^{\text{LCB}}(s, a^0) - \alpha \log(p(\hat{a}_i^1 \mid a_i^T, s_i))]$$

Entropy Regularization

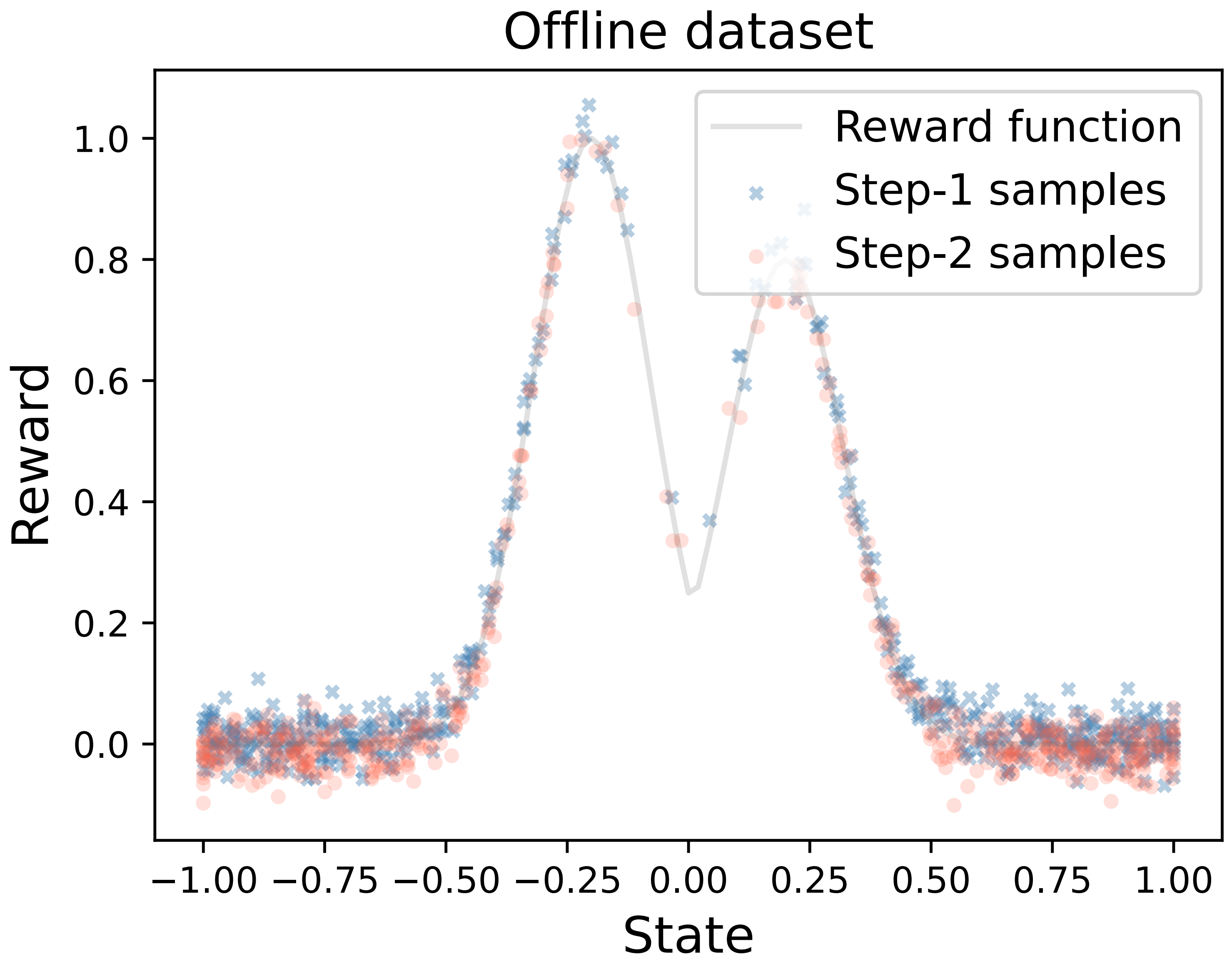
$$Q_\psi^{\text{LCB}} = \mathbb{E}_{\text{ens}}\left[Q_{\psi^m}(s, a)\right] - \beta\left[\sqrt{\mathbb{V}_{\text{ens}}[Q_{\psi^m}(s, a)]}\right]$$

LCB of Q-ensemble

[1] Wang, Z., Hunt, J. J., & Zhou, M. (2022). Diffusion policies as an expressive policy class for offline reinforcement learning. arXiv preprint arXiv:2208.06193.
[2] Haarnoja, T., Zhou, A., Abbeel, P., & Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. ICML 2018.

Diffusion Policy

Task: Starting from 0, take two steps to seek a state with the highest reward.



(a)

Diffusion Policy

Task: Starting from 0, take two steps to seek a state with the highest reward.

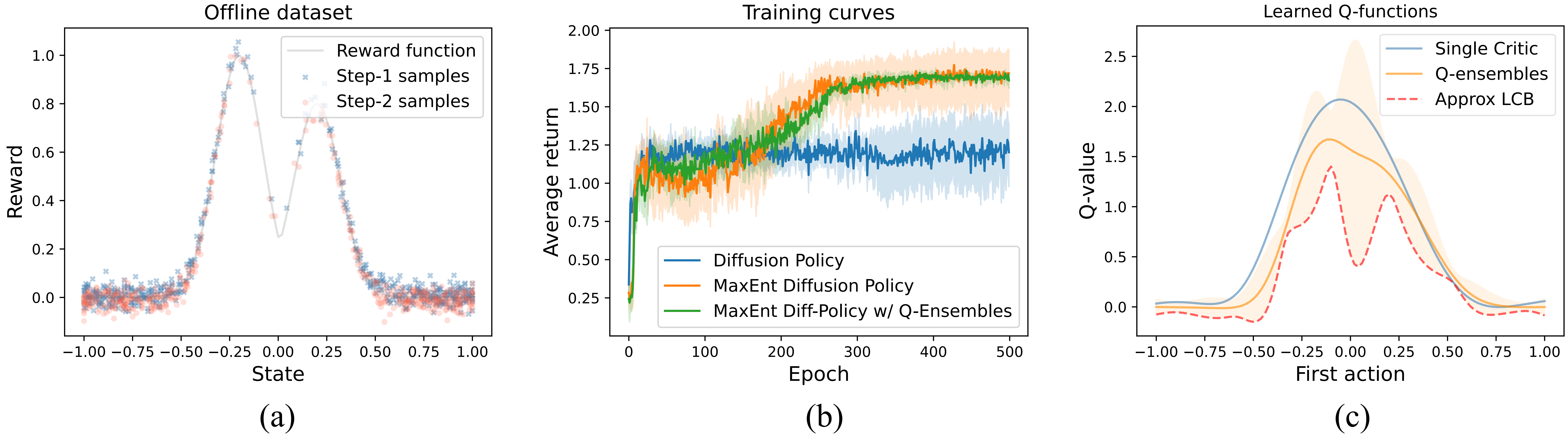
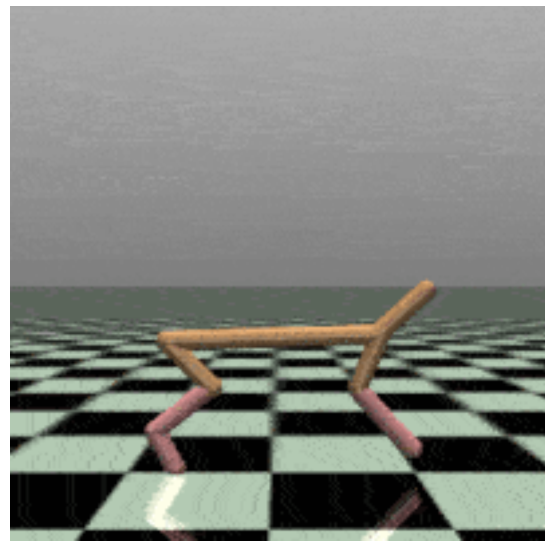
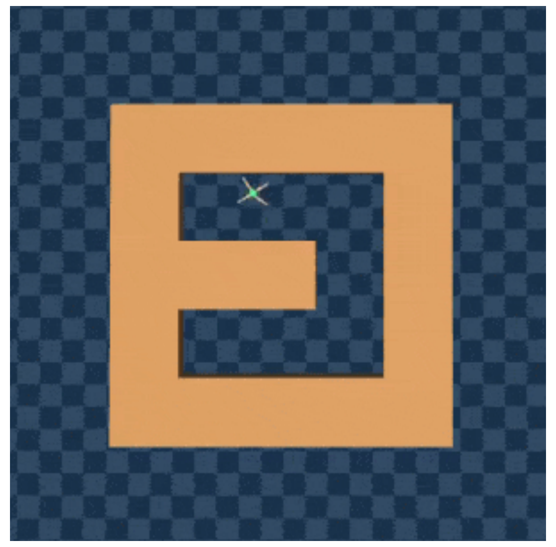
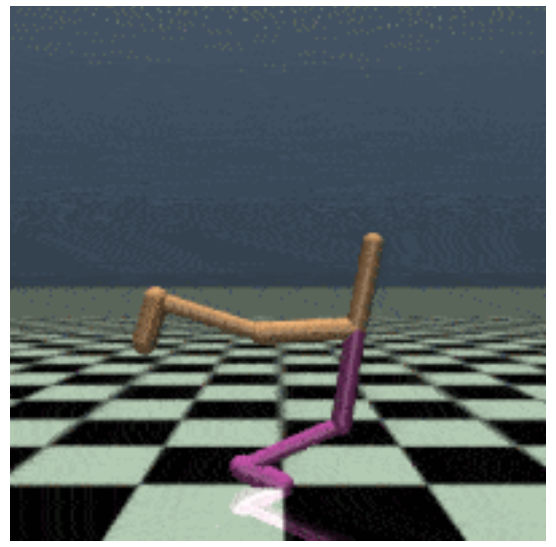


Figure 3. **Left:** Reward function and training samples **Center:** Training progress comparison **Right:** Learned Q-values curve in state 0 **Take-away:** Only combined entorpy+diffusion+ensembles learn a better policy and accurate Q-values. **[2]**

[2] *Zhang, R., Luo, Z., Sjölund, J., Schön, T. B., & Mattsson, P. (2024).* *Entropy-regularized diffusion policy with q-ensembles for offline reinforcement learning. Neurips 2024 (Accepted).*

UPPSALA UNIVERSITET

Experiments: D4RL

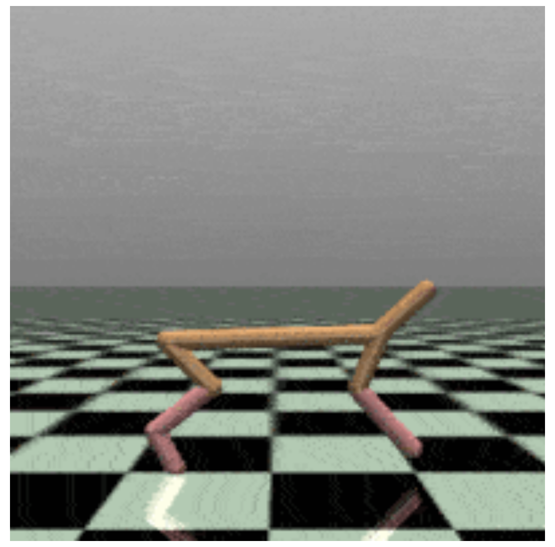
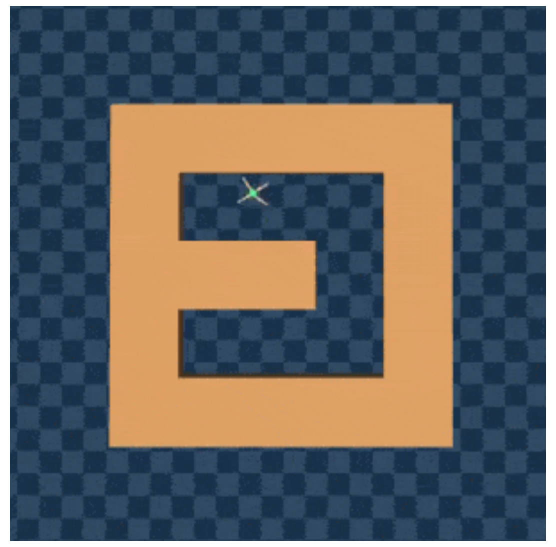
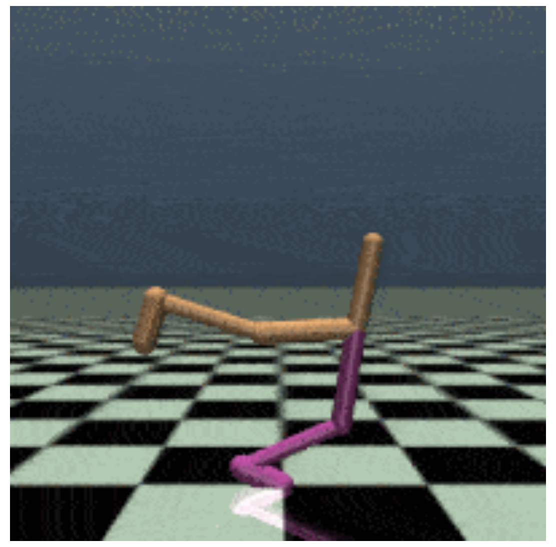





GYM TASKS	BC	DT	CQL	IQL	IDQL-A	IQL+EDP	DIFF-QL	OURS
HALFCHEETAH-MEDIUM-V2	42.6	42.6	44.0	47.4	51.0	48.1	51.1	**54.9**
HOPPER-MEDIUM-V2	52.9	67.6	58.5	66.3	65.4	63.1	90.5	**94.2**
WALKER2D-MEDIUM-V2	75.3	74.0	72.5	78.3	82.5	85.4	87.0	**92.5**
HALFCHEETAH-MEDIUM-REPLAY-V2	36.6	36.6	45.5	44.2	45.9	43.8	47.8	**57.0**
HOPPER-MEDIUM-REPLAY-V2	18.1	82.7	95.0	94.7	92.1	99.1	101.3	**102.7**
WALKER2D-MEDIUM-REPLAY-V2	26.0	66.6	77.2	73.9	85.1	84.0	**95.5**	94.20
HALFCHEETAH-MEDIUM-EXPERT-V2	55.2	86.8	91.6	86.7	95.9	86.7	**96.8**	90.32
HOPPER-MEDIUM-EXPERT-V2	52.5	107.6	105.4	91.5	108.6	99.6	111.1	**111.9**
WALKER2D-MEDIUM-EXPERT-V2	107.5	108.1	108.8	109.6	**112.7**	109.0	110.1	111.2
AVERAGE	51.9	74.7	77.6	77.0	82.1	79.9	88.0	**89.9**

ANTMAZE TASKS	BC	DT	CQL	IQL	MSG	IDQL-A	IQL+EDP	DIFF-QL	OURS
ANTMAZE-UMAZE-V0	54.6	59.2	74	87.5	97.8	94.0	87.5	93.4	**100**
ANTMAZE-UMAZE-DIVERSE-V0	45.6	53.0	84.0	62.2	**81.8**	80.2	62.2	66.2	79.8
ANTMAZE-MEDIUM-PLAY-V0	0.0	0.0	61.2	71.2	89.6	84.5	71.2	76.6	**91.4**
ANTMAZE-MEDIUM-DIVERSE-V0	0.0	0.0	53.7	70.0	88.6	84.8	70.0	78.6	**91.6**
ANTMAZE-LARGE-PLAY-V0	0.0	0.0	15.8	39.6	72.6	63.5	39.6	46.4	**81.2**
ANTMAZE-LARGE-DIVERSE-V0	0.0	0.0	14.9	47.5	71.4	67.9	47.6	56.6	**76.4**
AVERAGE	16.7	18.7	50.6	63.0	83.6	79.1	63.0	69.6	**86.7**

UPPSALA
UNIVERSITET

Experiments: D4RL

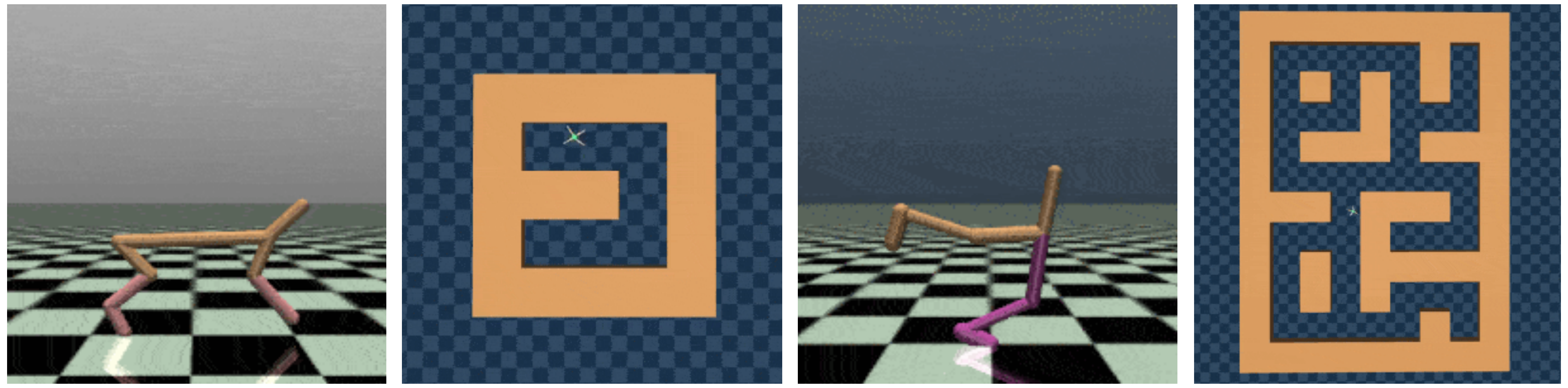





Gym Tasks	BC	DT	CQL	IQL	IDQL-A	IQL+EDP	Diff-QL	Ours
halfcheetah-medium-v2	42.6	42.6	44.0	47.4	51.0	48.1	51.1	**54.9**
hopper-medium-v2	52.9	67.6	58.5	66.3	65.4	63.1	90.5	**94.2**
walker2d-medium-v2	75.3	74.0	72.5	78.3	82.5	85.4	87.0	**92.5**
halfcheetah-medium-replay-v2	36.6	36.6	45.5	44.2	45.9	43.8	47.8	**57.0**
hopper-medium-replay-v2	18.1	82.7	95.0	94.7	92.1	99.1	101.3	**102.7**
walker2d-medium-replay-v2	26.0	66.6	77.2	73.9	85.1	84.0	**95.5**	94.20
halfcheetah-medium-expert-v2	55.2	86.8	91.6	86.7	95.9	86.7	**96.8**	90.32
hopper-medium-expert-v2	52.5	107.6	105.4	91.5	108.6	99.6	111.1	**111.9**
walker2d-medium-expert-v2	107.5	108.1	108.8	109.6	**112.7**	109.0	110.1	111.2
Average	51.9	74.7	77.6	77.0	82.1	79.9	88.0	**89.9**

AntMaze Tasks	BC	DT	CQL	IQL	MSG	IDQL-A	IQL+EDP	Diff-QL	Ours
antmaze-umaze-v0	54.6	59.2	74	87.5	97.8	94.0	87.5	93.4	**100**
antmaze-umaze-diverse-v0	45.6	53.0	84.0	62.2	**81.8**	80.2	62.2	66.2	79.8
antmaze-medium-play-v0	0.0	0.0	61.2	71.2	89.6	84.5	71.2	76.6	**91.4**
antmaze-medium-diverse-v0	0.0	0.0	53.7	70.0	88.6	84.8	70.0	78.6	**91.6**
antmaze-large-play-v0	0.0	0.0	15.8	39.6	72.6	63.5	39.6	46.4	**81.2**
antmaze-large-diverse-v0	0.0	0.0	14.9	47.5	71.4	67.9	47.6	56.6	**76.4**
Average	16.7	18.7	50.6	63.0	83.6	79.1	63.0	69.6	**86.7**

UPPSALA
UNIVERSITET

Experiments: D4RL



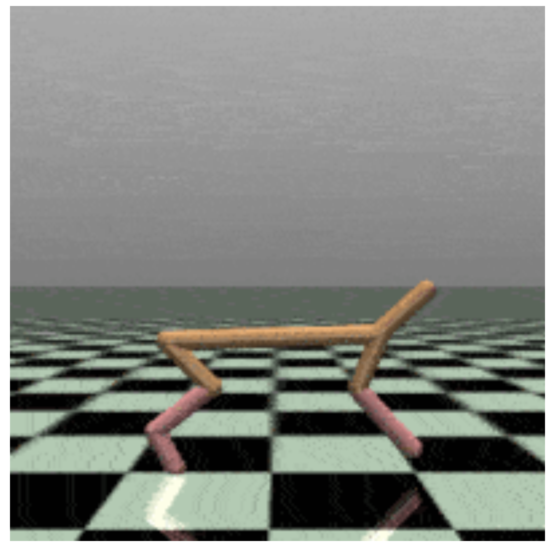
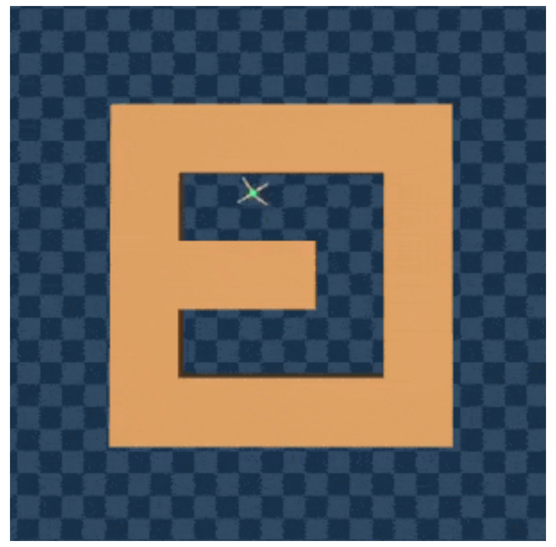
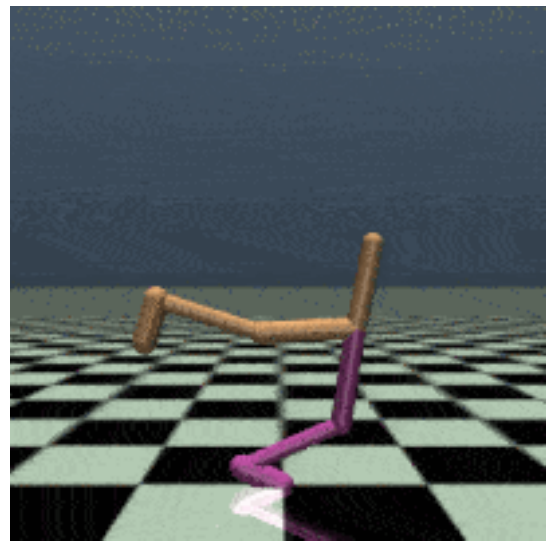
Gym Tasks	BC	DT	CQL	IQL	IDQL-A	IQL+EDP	Diff-QL	Ours
halfcheetah-medium-v2	42.6	42.6	44.0	47.4	51.0	48.1	51.1	**54.9**
hopper-medium-v2	52.9	67.6	58.5	66.3	65.4	63.1	90.5	**94.2**
walker2d-medium-v2	75.3	74.0	72.5	78.3	82.5	85.4	87.0	**92.5**
halfcheetah-medium-replay-v2	36.6	36.6	45.5	44.2	45.9	43.8	47.8	**57.0**
hopper-medium-replay-v2	18.1	82.7	95.0	94.7	92.1	99.1	101.3	**102.7**
walker2d-medium-replay-v2	26.0	66.6	77.2	73.9	85.1	84.0	**95.5**	94.20
halfcheetah-medium-expert-v2	55.2	86.8	91.6	86.7	95.9	86.7	**96.8**	90.32
hopper-medium-expert-v2	52.5	107.6	105.4	91.5	108.6	99.6	111.1	**111.9**
walker2d-medium-expert-v2	107.5	108.1	108.8	109.6	**112.7**	109.0	110.1	111.2
Average	51.9	74.7	77.6	77.0	82.1	79.9	88.0	**89.9**

Suboptimal + Sparse Reward

AntMaze Tasks	BC	DT	CQL	IQL	MSG	IDQL-A	IQL+EDP	Diff-QL	Ours
antmaze-umaze-v0	54.6	59.2	74	87.5	97.8	94.0	87.5	93.4	**100**
antmaze-umaze-diverse-v0	45.6	53.0	84.0	62.2	**81.8**	80.2	62.2	66.2	79.8
antmaze-medium-play-v0	0.0	0.0	61.2	71.2	89.6	84.5	71.2	76.6	**91.4**
antmaze-medium-diverse-v0	0.0	0.0	53.7	70.0	88.6	84.8	70.0	78.6	**91.6**
antmaze-large-play-v0	0.0	0.0	15.8	39.6	72.6	63.5	39.6	46.4	**81.2**
antmaze-large-diverse-v0	0.0	0.0	14.9	47.5	71.4	67.9	47.6	56.6	**76.4**
Average	16.7	18.7	50.6	63.0	83.6	79.1	63.0	69.6	**86.7**

UPPSALA UNIVERSITET

Experiments: D4RL


GYM TASKS	BC	DT	CQL	IQL	IDQL-A	IQL+EDP	DIFF-QL	OURS
HALFCHEETAH-MEDIUM-V2	42.6	42.6	44.0	47.4	51.0	48.1	51.1	**54.9**
HOPPER-MEDIUM-V2	52.9	67.6	58.5	66.3	65.4	63.1	90.5	**94.2**
WALKER2D-MEDIUM-V2	75.3	74.0	72.5	78.3	82.5	85.4	87.0	**92.5**
HALFCHEETAH-MEDIUM-REPLAY-V2	36.6	36.6	45.5	44.2	45.9	43.8	47.8	**57.0**
HOPPER-MEDIUM-REPLAY-V2	18.1	82.7	95.0	94.7	92.1	99.1	101.3	**102.7**
WALKER2D-MEDIUM-REPLAY-V2	26.0	66.6	77.2	73.9	85.1	84.0	**95.5**	94.20
HALFCHEETAH-MEDIUM-EXPERT-V2	55.2	86.8	91.6	86.7	95.9	86.7	**96.8**	90.32
HOPPER-MEDIUM-EXPERT-V2	52.5	107.6	105.4	91.5	108.6	99.6	111.1	**111.9**
WALKER2D-MEDIUM-EXPERT-V2	107.5	108.1	108.8	109.6	**112.7**	109.0	110.1	111.2
AVERAGE	51.9	74.7	77.6	77.0	82.1	79.9	88.0	**89.9**

Suboptimal + Sparse Reward

ANTMAZE TASKS	BC	DT	CQL	IQL	MSG	IDQL-A	IQL+EDP	DIFF-QL	OURS
ANTMAZE-UMAZE-V0	54.6	59.2	74	87.5	97.8	94.0	87.5	93.4	**100**
ANTMAZE-UMAZE-DIVERSE-V0	45.6	53.0	84.0	62.2	**81.8**	80.2	62.2	66.2	79.8
ANTMAZE-MEDIUM-PLAY-V0	0.0	0.0	61.2	71.2	89.6	84.5	71.2	76.6	**91.4**
ANTMAZE-MEDIUM-DIVERSE-V0	0.0	0.0	53.7	70.0	88.6	84.8	70.0	78.6	**91.6**
ANTMAZE-LARGE-PLAY-V0	0.0	0.0	15.8	39.6	72.6	63.5	39.6	46.4	**81.2**
ANTMAZE-LARGE-DIVERSE-V0	0.0	0.0	14.9	47.5	71.4	67.9	47.6	56.6	**76.4**
AVERAGE	16.7	18.7	50.6	63.0	83.6	79.1	63.0	69.6	**86.7**

↑24%

UPPSALA UNIVERSITET