COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

ICML
International Conference
On Machine Learning

# Causal Flow Q-Learning for Robust Offline Reinforcement Learning

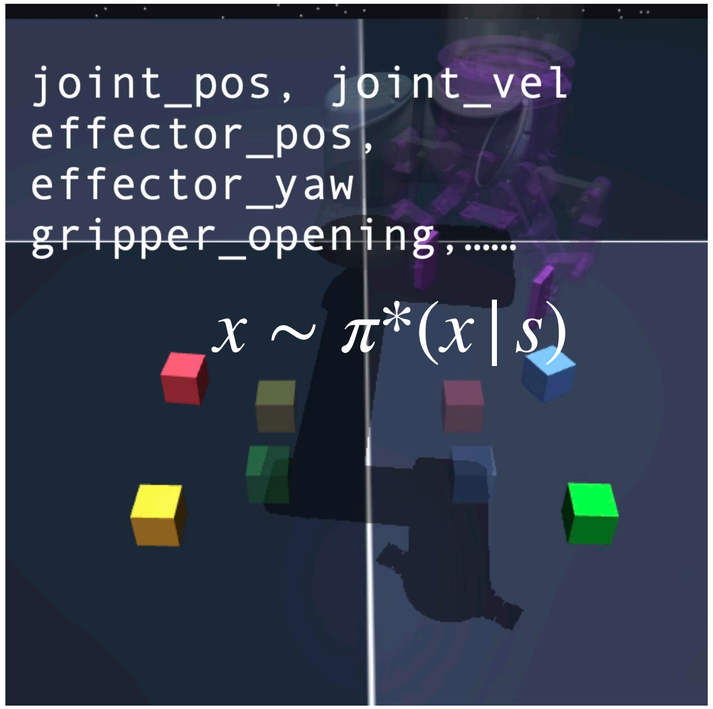

**Mingxuan Li**

**Junzhe Zhang**

**Elias Bareinboim**

# Challenges of Confounding in Offline RL

In pixel-based offline RL tasks, the dataset is composed of expert generated **actions using structured state observations** and **blurry image observations**.

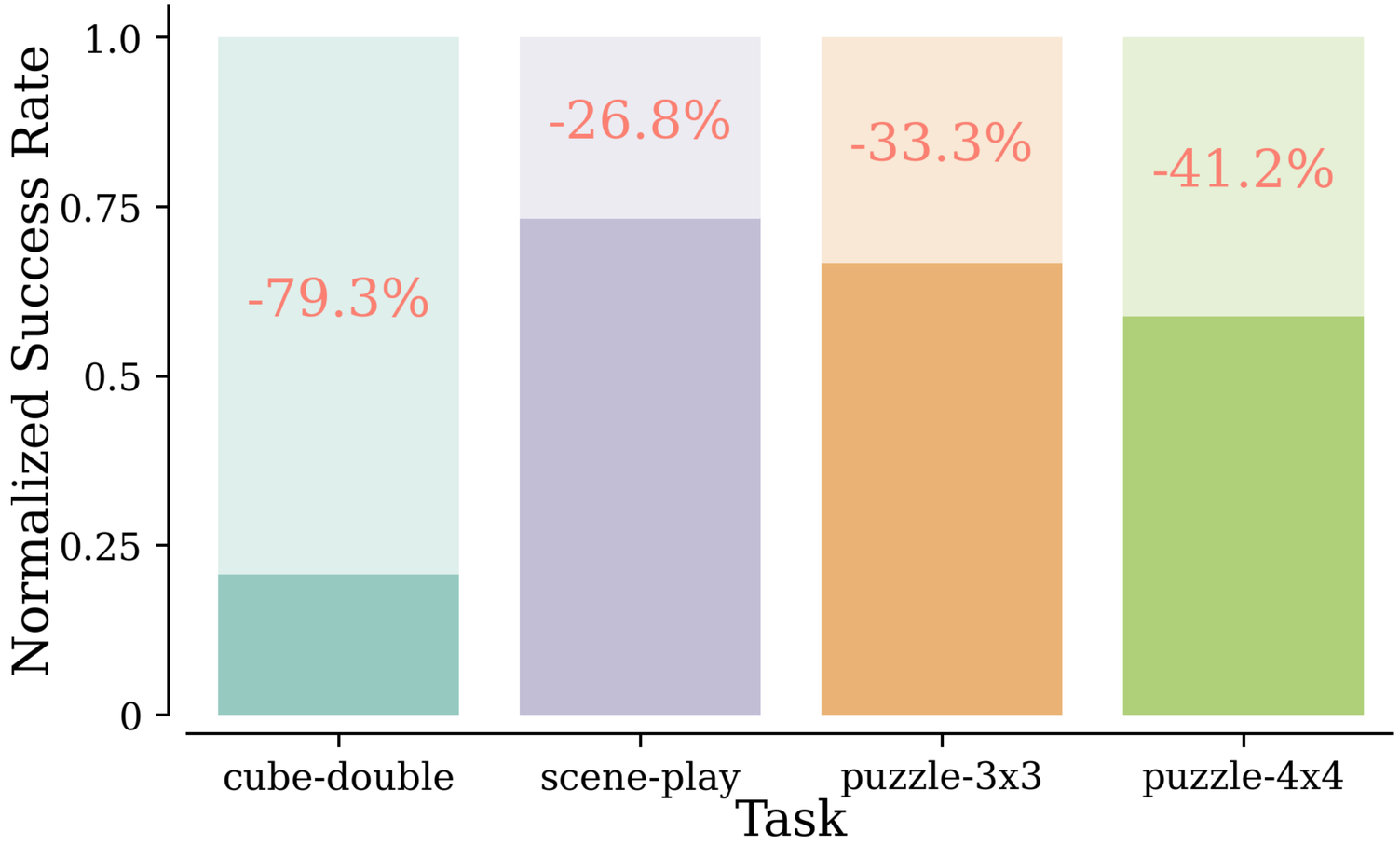



$+$

$= (o, x)?$

# Challenges of Confounding in Offline RL

This **mismatch (confounding)** caused SOTA offline RL algorithms' performance drop sharply in pixel-based tasks, even though they achieve high success rate in the corresponding state-based version.

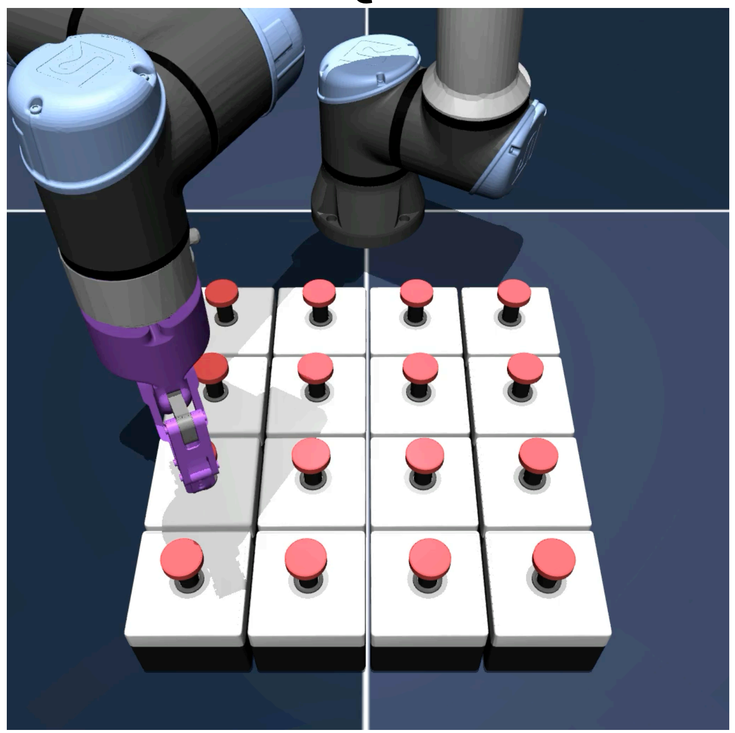


# Challenges of Confounding in Offline RL

We develop a confounding robust offline RL objective that,

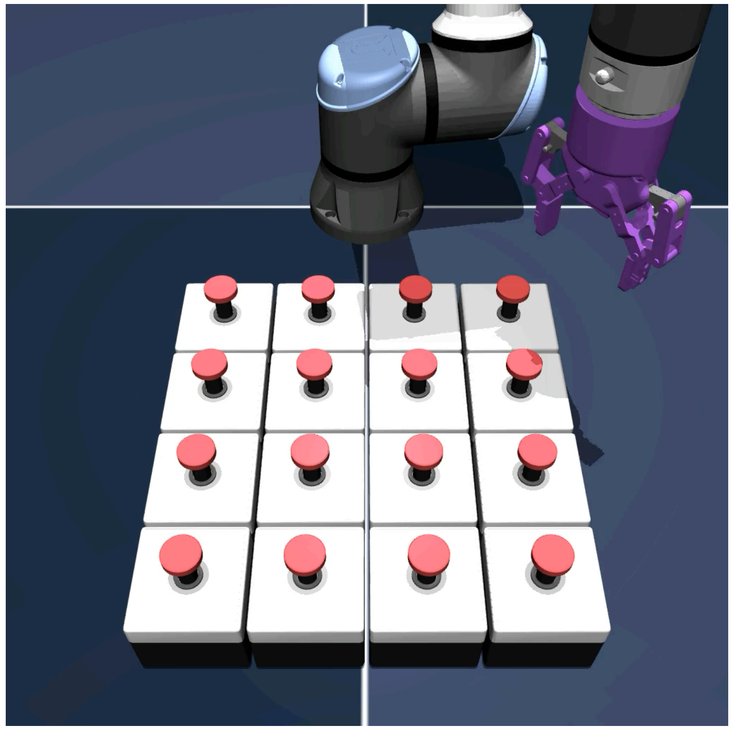
- is grounded in causal language, and
- improves success rate significantly in both offline & offline-to-online pixel tasks.

FQL



Avg. Success Rate +20 %

Causal-FQL



# Modeling the Problem with Confounded MDP (CMDP)

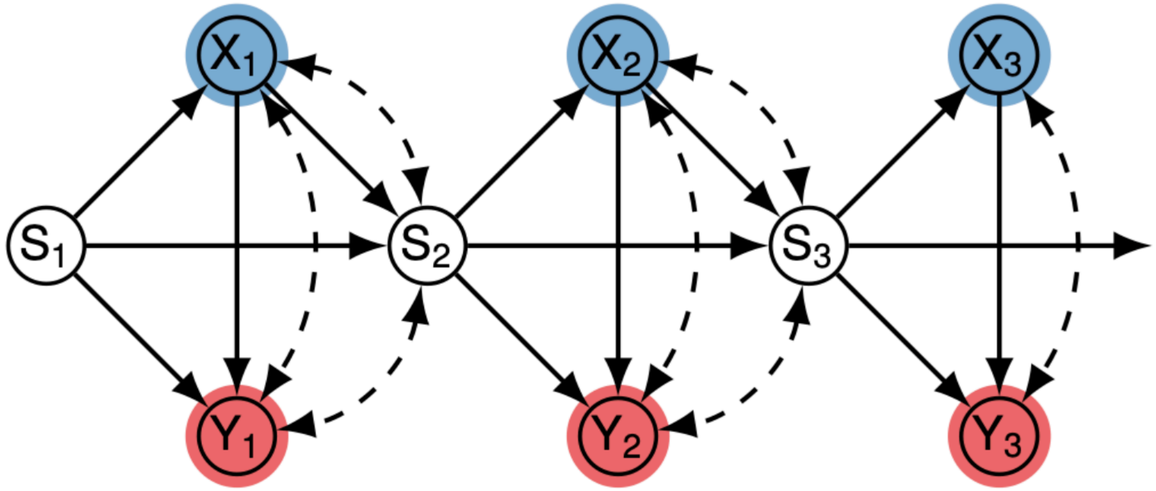
**Definition 1 (Confounded MDP).** A Confounded Markov Decision Process (CMDP) $\mathscr{M}$ is a tuple of $\langle \mathcal{S}, \mathcal{X}, \mathcal{Y}, \mathcal{U}, \mathbb{F}, \mathbb{P} \rangle$ where,

- $\mathcal{S}, \mathcal{X}, \mathcal{Y}$ are, respectively, the space of observed states, actions, and rewards;
- $\mathcal{U}$ is the space of unobserved exogenous noise;
- $\mathbb{F}$ is a set consisting of the transition function $\tau_h : \mathcal{S} \times \mathcal{X} \times \mathcal{U} \mapsto \mathcal{S}$, behavioral policy $\beta_h : \mathcal{S} \times \mathcal{U} \mapsto \mathcal{X}$, and reward function $r_h : \mathcal{S} \times \mathcal{X} \times \mathcal{U} \mapsto \mathcal{Y}$ for every time step $h = 1, \ldots$;
- $\mathbb{P}$ is a set of distributions $P_h$ over the unobserved domain $\mathcal{U}$ for every time step $h = 1, \ldots.$

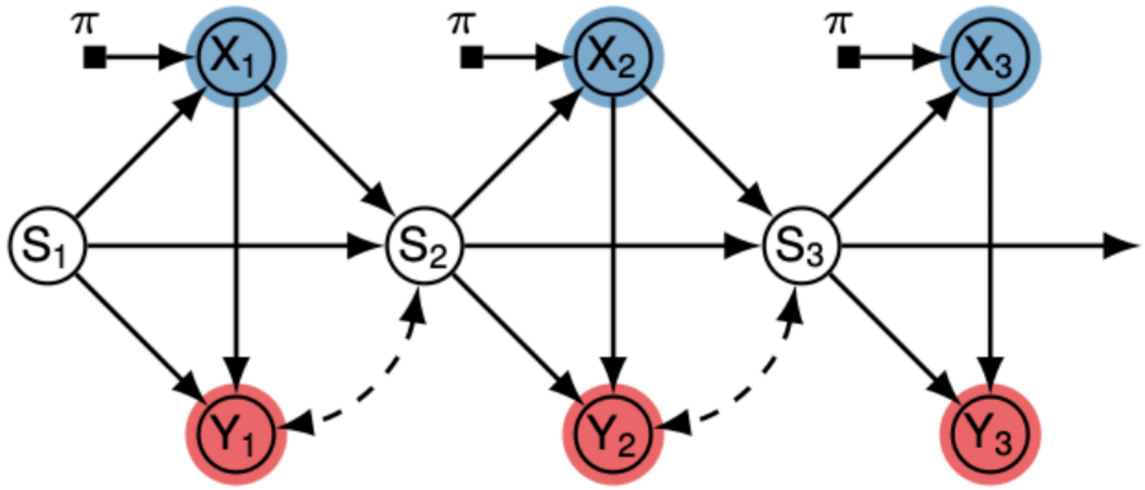
# CMDP - Graphical Model

Those bi-directed dashed arrows are confounders indicating the variables that are,

1. Observable to expert; and
2. Not observable to the learner.




(a) CMDP




(b) CMDP-Online

# Bounding Worst Case via Causal Bellman Equation

Directly maximizing the Q-values from such confounded data results in poor performance.

We propose the Causal Bellman Equation to lower bound the optimal Q-values.

**Theorem 1 (Causal Bellman Equation).** For a CMDP environment $\mathscr{M}$ with reward $Y_h \geq a, a \in \mathbb{R}$, the optimal value of interventional policies, $Q^*(\boldsymbol{s}, x), \forall \boldsymbol{s} \in \mathcal{S}$, is lower bounded by $\underline{Q_*}(\boldsymbol{s}, x)$ satisfying the Causal Bellman Optimality Equation, for every step $h = 1, \ldots, H$,

$$\underline{Q_*}(s,x) = P(x \mid s)\left(\widetilde{\mathscr{R}}(s,x) + \gamma \sum_{s',x'} \widetilde{\mathscr{T}}(s,x,s') \max_{x'} \underline{Q_*}(s',x')\right) + P(\neg x \mid s)\left(a + \gamma \min_{s'} \max_{x'} \underline{Q_*}(s',x')\right)$$

where $P(x \mid s) = P(X_t = x \mid S_t = s)$ is the propensity score calculated from the offline dataset, $\tilde{T}, \tilde{R}$ are estimated transition distribution and rewards, separately.

# Causal Bellman Equation in the Continuous Space

- There are two policies involved in estimating value bounds, one behavioral, one learner's policy we are trying to train.

$$\underline{Q_*}(s, x) = P(x \mid s)\left(\widetilde{\mathcal{R}}(s,x) + \gamma \sum_{s',x'} \widetilde{\mathcal{T}}(s,x,s') \max_{x'} \underline{Q_*}(s',x')\right) + P(\neg x \mid s)\left(a + \gamma \min_{s'} \max_{x'} \underline{Q_*}(s',x')\right)$$

$$x \sim \pi(\cdot \mid s)$$

- For continuous actions, P(x|s) is zero. Thus, we instead interpret P(x|s) as an action $x$ is sampled from the behavioral policy.

- And this can be thought of as a binary classification problem and can be learned by a neural network classifier,

$$P(x \mid s) = D_\theta(s, x)$$

# Approximate Causal Bellman Equation via Ensembles

- We propose to fit an ensemble of neural networks to represent different possible "realities" (SCMs) corresponding to the same offline dataset,

$$Q_{\theta_i}, i \in [1,K]$$

- Taking min over the ensembles and plug in the discriminator,

$$\underline{Q_*}(s,x) = P(x \mid s)\left(\widetilde{\mathscr{R}}(s,x) + \gamma \sum_{s',x'} \widetilde{\mathscr{T}}(s,x,s') \max_{x'} \underline{Q_*}(s',x')\right) + P(\neg x \mid s)\left(a + \gamma \min_{s'} \max_{x'} \underline{Q_*}(s',x')\right)$$

$$\underline{\hat{Q}_*}(s,x) = \frac{1}{K} D_\theta(x \mid s) \sum_{i=1}^{K} Q_{\theta_i}(s,x) + \left(1 - D_\theta(\ x \mid s)\right) \min_{i \in [1,K]} Q_{\theta_i}$$

# Causal Flow Q-Learning (Algorithm)

While not converged:

Sample a batch $\{(s, x, r, s')\} \sim \mathcal{D}$

Train critic ensemble $Q_{\theta_i}$

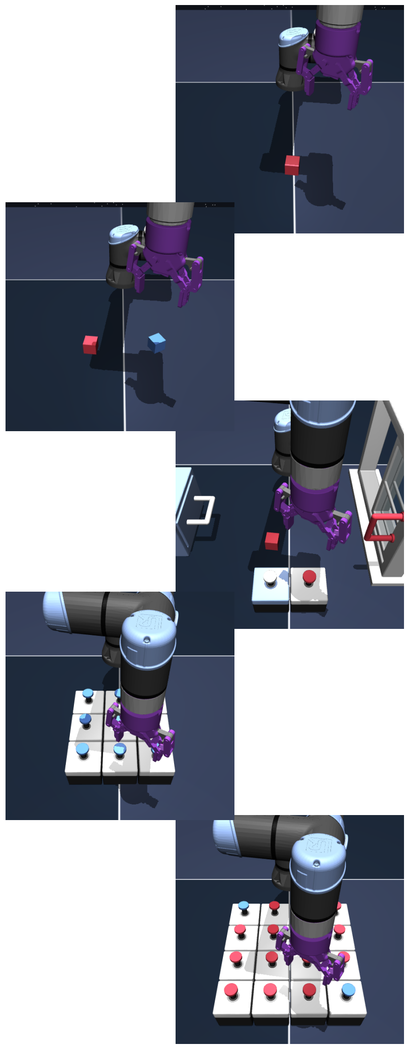
Train flow policy $\mu_\omega$ and discriminator $D_\theta$

Train one-step target policy:

$$\mathcal{L}_\pi(\theta) = \mathbb{E}_{s\sim\mathcal{D}, x^\pi\sim\pi_\theta}[-\underline{\hat{Q}_*}(s, x^\pi)] + \alpha\mathcal{L}_{\text{Distill}}(\omega)$$

$$\underline{\hat{Q}_*}(s, x) = \frac{1}{K} D_\theta(x \mid s) \sum_{i=1}^{K} Q_{\theta_i}(s, x) + \left(1 - D_\theta(x \mid s)\right) \min_{i\in[1,K]} Q_{\theta_i}$$

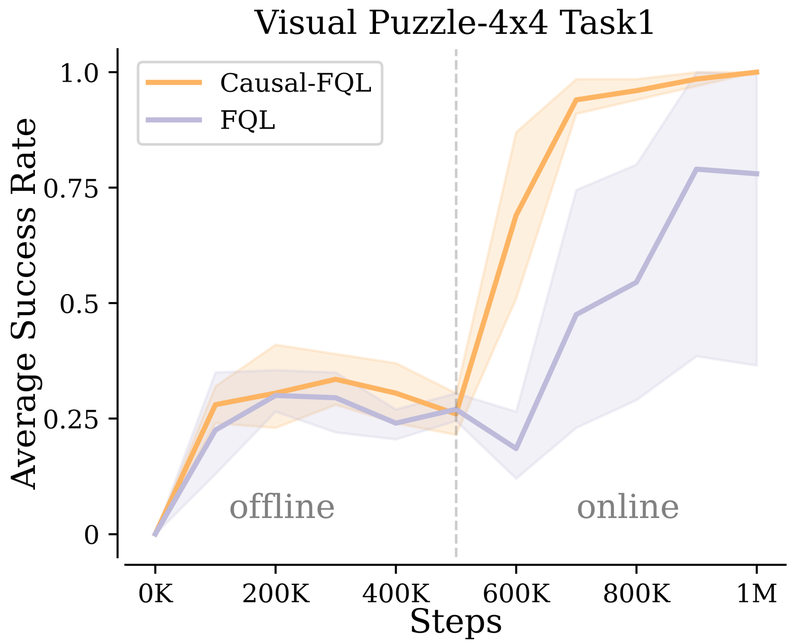
# Experimental Results - Offline

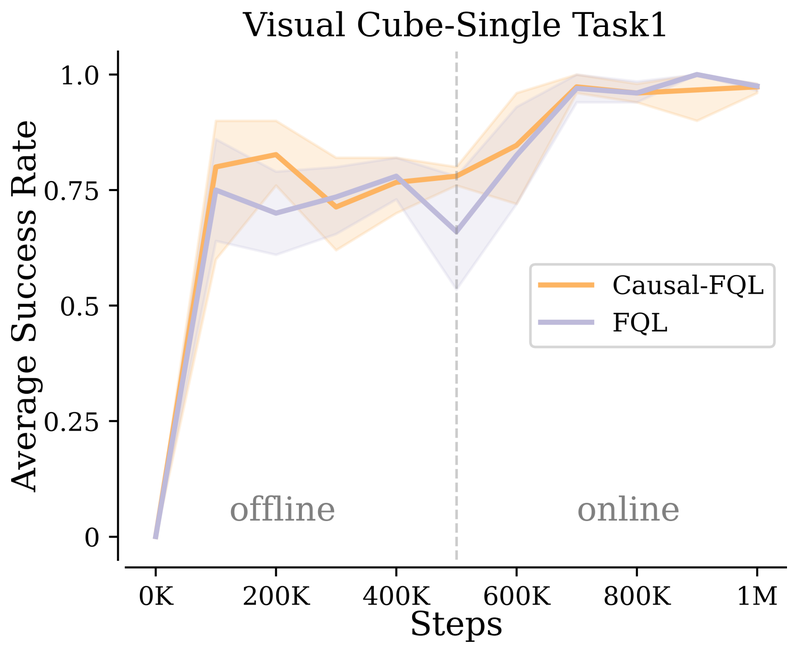


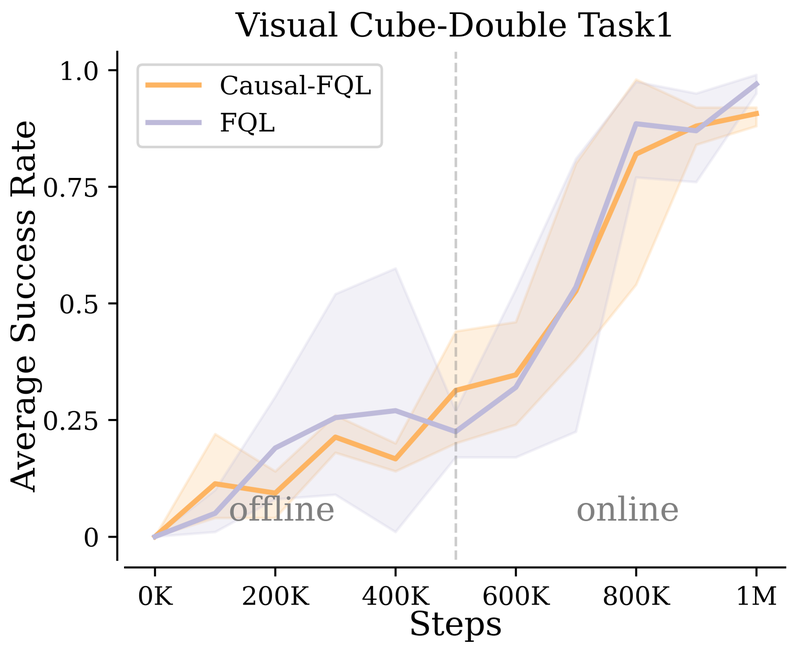
| | Gaussian Policies | | Flow Policies | | | | |
|---|---|---|---|---|---|---|---|
| | IQL | ReBRAC | FBRAC | IQN | IFQL | FQL | Causal-FQL |
| visual-cube-single-play-singletask-task1-v0 (*) | 70 ± 12 | **83 ± 6** | 55 ± 8 | 5 ± 4 | 49 ± 7 | 81 ± 12 | **85 ± 6** |
| visual-cube-single-play-singletask-task2-v0 | 59 ± 13 | **85 ± 6** | 70 ± 11 | 58 ± 9 | 45 ± 11 | 58 ± 12 | **81 ± 5** |
| visual-cube-single-play-singletask-task3-v0 | 64 ± 35 | **88 ± 4** | 81 ± 4 | 83 ± 11 | 67 ± 6 | 73 ± 16 | **84 ± 5** |
| visual-cube-single-play-singletask-task4-v0 | 68 ± 12 | **78 ± 9** | 56 ± 19 | **78 ± 3** | 34 ± 13 | 63 ± 8 | **74 ± 5** |
| visual-cube-single-play-singletask-task5-v0 | 59 ± 14 | **76 ± 9** | 59 ± 10 | 64 ± 9 | 27 ± 10 | 49 ± 12 | **79 ± 7** |
| visual-cube-double-play-singletask-task1-v0 (*) | 34 ± 23 | 4 ± 4 | 6 ± 2 | 4 ± 1 | 8 ± 6 | 23 ± 4 | **43 ± 11** |
| visual-cube-double-play-singletask-task2-v0 | **2 ± 1** | 0 ± 0 | 2 ± 2 | 0 ± 0 | 0 ± 0 | 0 ± 0 | **2 ± 1** |
| visual-cube-double-play-singletask-task3-v0 | **7 ± 4** | 2 ± 2 | 2 ± 1 | 0 ± 0 | 1 ± 1 | 4 ± 2 | 4 ± 2 |
| visual-cube-double-play-singletask-task4-v0 | **1 ± 1** | 0 ± 0 | 0 ± 0 | 0 ± 0 | 0 ± 0 | 0 ± 0 | **1 ± 1** |
| visual-cube-double-play-singletask-task5-v0 | **11 ± 2** | 0 ± 0 | 0 ± 0 | 1 ± 1 | 2 ± 1 | 4 ± 1 | 4 ± 2 |
| visual-scene-play-singletask-task1-v0 (*) | **97 ± 2** | **98 ± 4** | 46 ± 4 | **95 ± 2** | 86 ± 10 | **98 ± 3** | **100 ± 0** |
| visual-scene-play-singletask-task2-v0 | 21 ± 16 | 30 ± 15 | 0 ± 0 | 79 ± 15 | 0 ± 0 | 86 ± 8 | **92 ± 3** |
| visual-scene-play-singletask-task3-v0 | 12 ± 9 | 10 ± 7 | 10 ± 3 | **31 ± 14** | 19 ± 2 | 22 ± 6 | 22 ± 1 |
| visual-scene-play-singletask-task4-v0 | 1 ± 0 | 0 ± 0 | 0 ± 0 | 0 ± 0 | 0 ± 0 | 1 ± 1 | **2 ± 1** |
| visual-scene-play-singletask-task5-v0 | **0 ± 0** | **0 ± 0** | **0 ± 0** | **0 ± 0** | **0 ± 0** | **0 ± 0** | **0 ± 0** |
| visual-puzzle-3x3-play-singletask-task1-v0 (*) | 7 ± 15 | 88 ± 4 | 7 ± 2 | 84 ± 1 | **100 ± 0** | 94 ± 1 | **98 ± 1** |
| visual-puzzle-3x3-play-singletask-task2-v0 | 0 ± 0 | **12 ± 1** | 0 ± 0 | 6 ± 2 | 0 ± 0 | 0 ± 0 | 1 ± 0 |
| visual-puzzle-3x3-play-singletask-task3-v0 | 0 ± 0 | 1 ± 1 | 0 ± 0 | 1 ± 0 | 2 ± 1 | 0 ± 0 | **3 ± 2** |
| visual-puzzle-3x3-play-singletask-task4-v0 | 1 ± 1 | 0 ± 1 | 0 ± 0 | 3 ± 0 | 1 ± 0 | 5 ± 4 | **8 ± 3** |
| visual-puzzle-3x3-play-singletask-task5-v0 | 0 ± 0 | 0 ± 0 | 0 ± 0 | 1 ± 0 | 0 ± 0 | 1 ± 2 | **19 ± 4** |
| visual-puzzle-4x4-play-singletask-task1-v0 (*) | 0 ± 0 | 26 ± 6 | 0 ± 0 | 21 ± 3 | 8 ± 15 | 33 ± 6 | **37 ± 4** |
| visual-puzzle-4x4-play-singletask-task2-v0 | 0 ± 0 | 0 ± 0 | 1 ± 1 | **14 ± 2** | 1 ± 1 | 1 ± 2 | 4 ± 1 |
| visual-puzzle-4x4-play-singletask-task3-v0 | 0 ± 0 | 0 ± 0 | 0 ± 0 | 9 ± 3 | 9 ± 15 | 16 ± 6 | **18 ± 4** |
| visual-puzzle-4x4-play-singletask-task4-v0 | 0 ± 0 | 1 ± 1 | 0 ± 0 | **12 ± 4** | 2 ± 2 | 2 ± 1 | **6 ± 2** |
| visual-puzzle-4x4-play-singletask-task5-v0 | 0 ± 0 | 0 ± 0 | 1 ± 1 | **11 ± 4** | 0 ± 0 | 1 ± 1 | 1 ± 1 |

# Experimental Results - Offline to Online

On puzzle-4x4 task1, our method is the first to achieve 100% success rate.

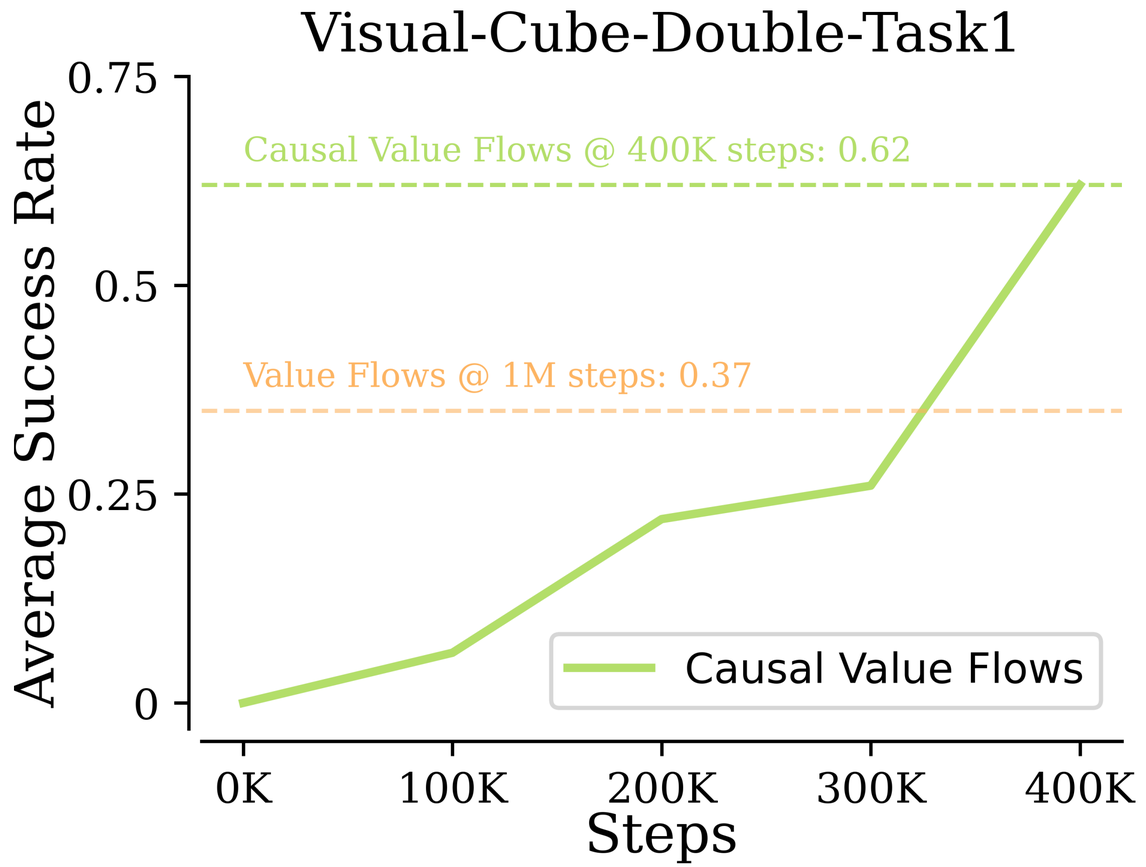


# Experimental Results - Causal Value Flows

Our proposed causal offline RL objective can also be implemented upon other algorithms. We use value flows[1], a more recent SOTA, to demonstrate this.




[1]Value Flows, ICLR 2026