Q-Flow:

Stable and Expressive Reinforcement Learning with Flow-Based Policy

JaeHyeok Doo | jdoo2@kaist.ac.kr

Background: Flow-based RL/offline RL

Behavior-regularized actor critic

$$\mathcal{L}_{\text{critic}}(\phi) = \mathbb{E}_{\substack{(s,a,r,s')\sim\mathcal{D} \\ a'\sim\pi(\cdot|s')}}\left[\left(Q_\phi(s,a) - (r + \gamma \bar{Q}_{\bar{\phi}}(s',a'))\right)^2\right]$$

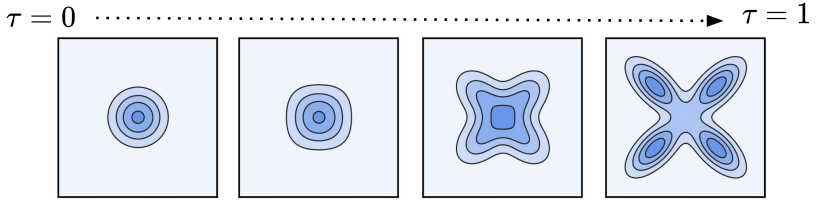
Regularization strength

$$\mathcal{L}_{\text{actor}}(\theta) = \mathbb{E}_{\substack{(s,a)\sim\mathcal{D} \\ a^\pi\sim\pi_\theta(\cdot|s)}}\left[-Q_\phi(s,a^\pi) - \alpha \log \pi_\theta(a \mid s)\right]$$

value maximization

behavior regularization

Flow-matching



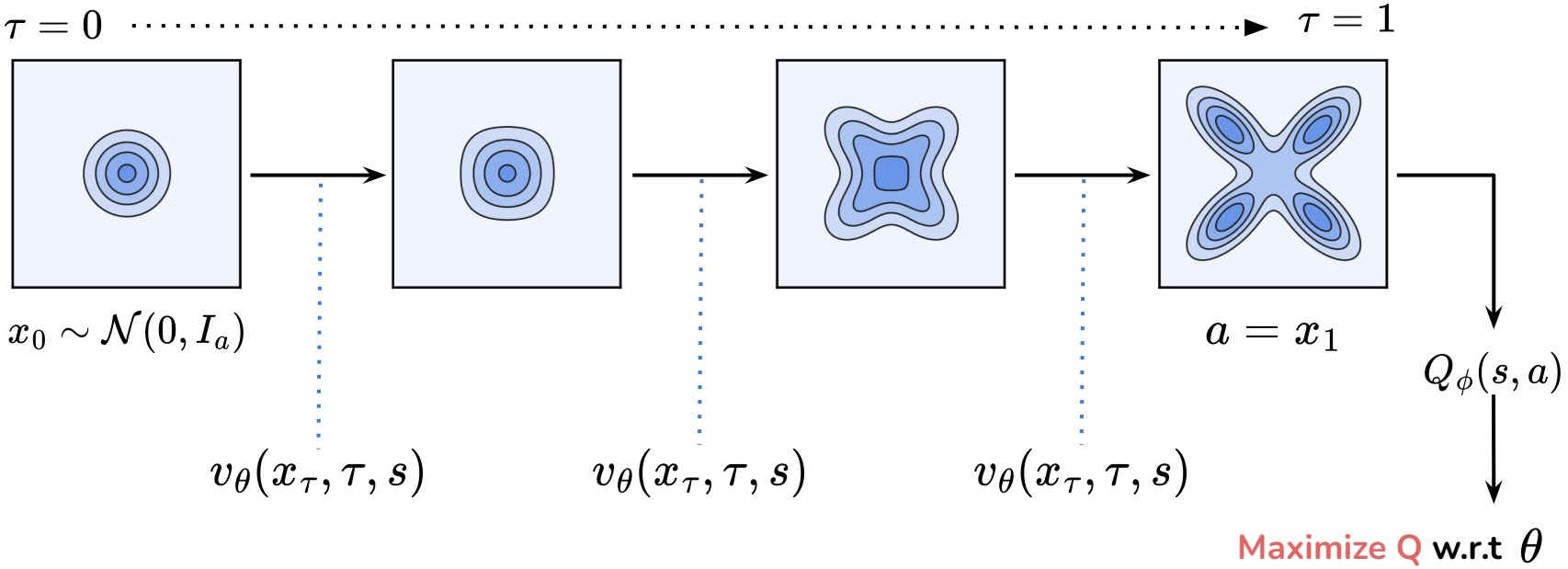
$$\mathcal{L}_{\text{CFM}}(\theta) = \mathbb{E}_{\substack{\tau\sim\mathcal{U}(0,1) \\ x_0\sim\mathcal{N}(0,I_a) \\ (s,a=x_1)\sim\mathcal{D}}}\left[||v_\theta(x_\tau,\tau,s) - (x_1 - x_0)||^2\right]$$

Flow-based RL/offline RL

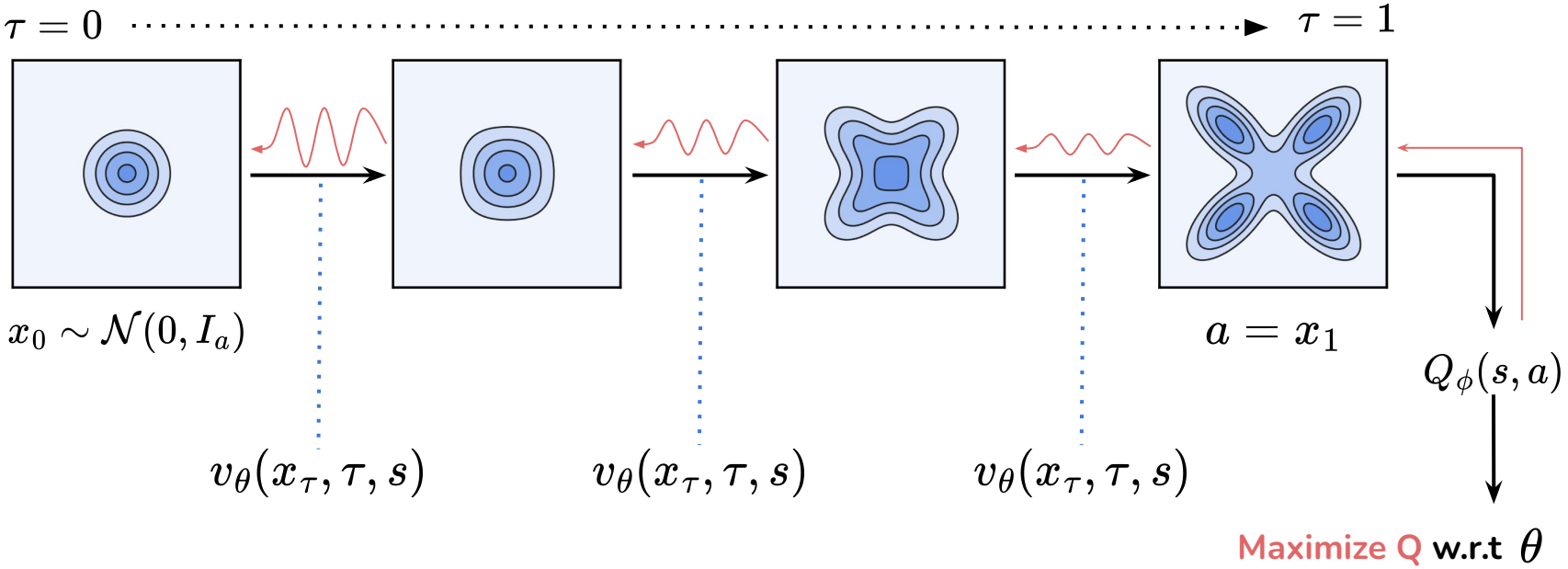
$$\mathcal{L}_{\text{actor}}(\theta) = \mathbb{E}_{\substack{(s,a)\sim\mathcal{D} \\ a^\pi\sim\pi_\theta(\cdot|s)}}\left[-Q_\phi(s,a^\pi)\right] + \alpha\mathcal{L}_{\text{CFM}}(\theta)$$

Flow-matching as behavior regularizer

What happens with Flow Policy?

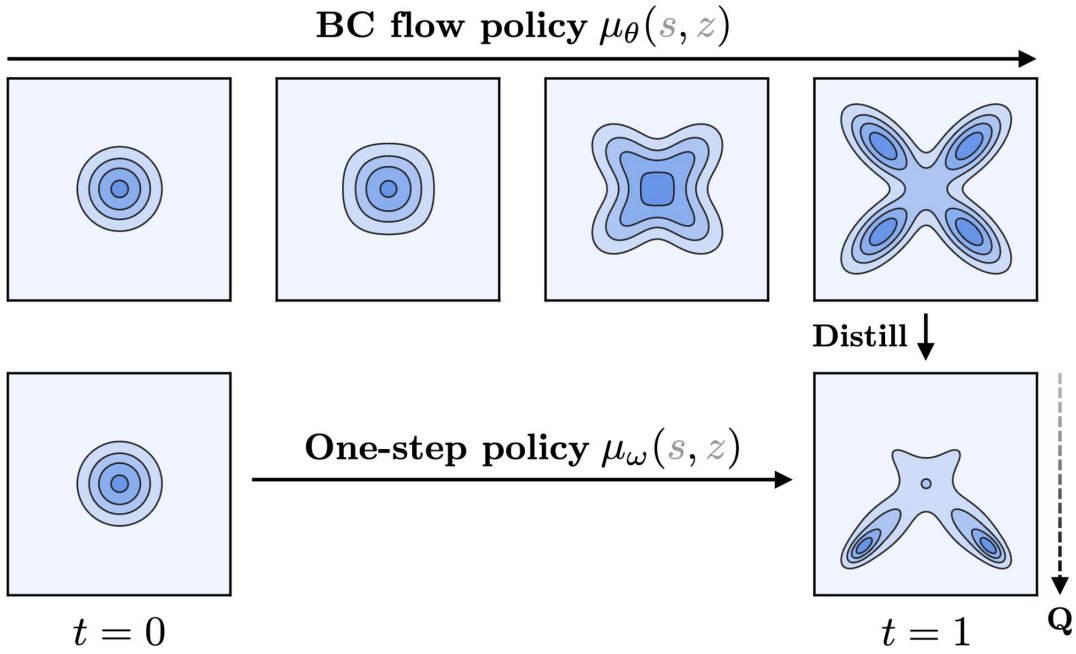


What happens with Flow Policy?



Gradient backpropagates through time!

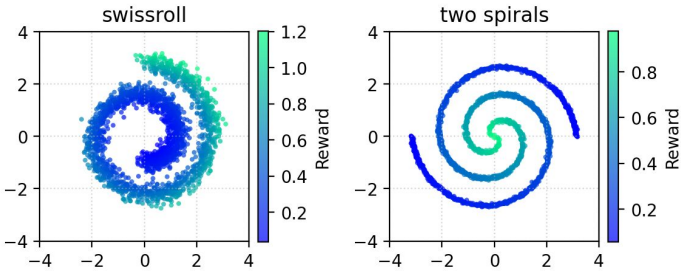
One-step Distillation avoids BPTT



Avoids gradient backpropgation,
But limited expressivity!

Stability-Expressivity dilemma in Flow-based RL

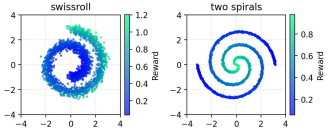
2D empirical validation

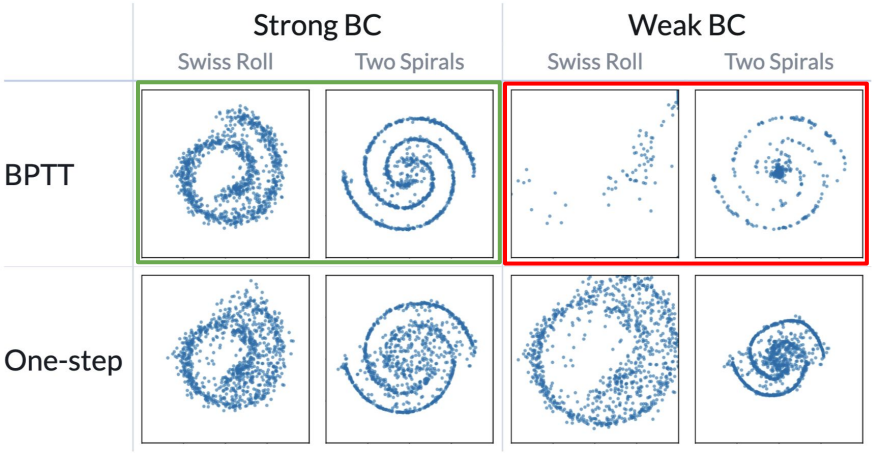


We tested **BPTT** and **One-step distillation**-based methods with
2D synthetic datasets across **different BC reg. strengths**

Stability-Expressivity dilemma in Flow-based RL

2D empirical validation

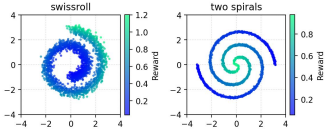


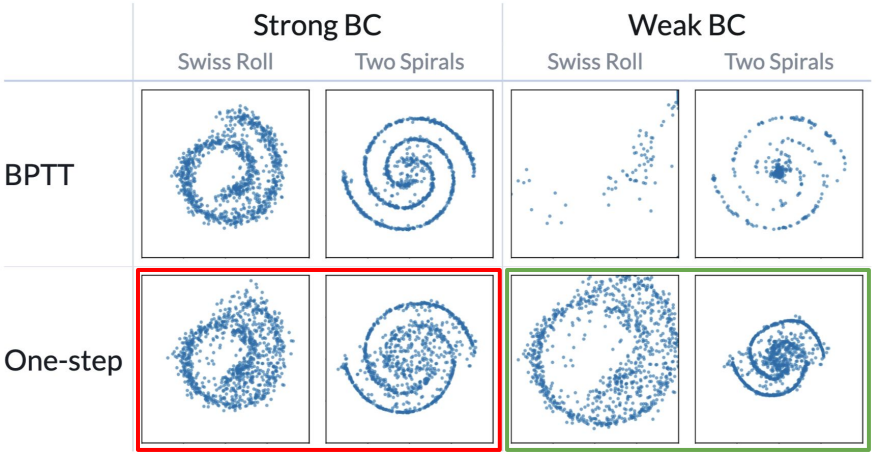


Expressive, but
unstable optimization

Stability-Expressivity dilemma in Flow-based RL

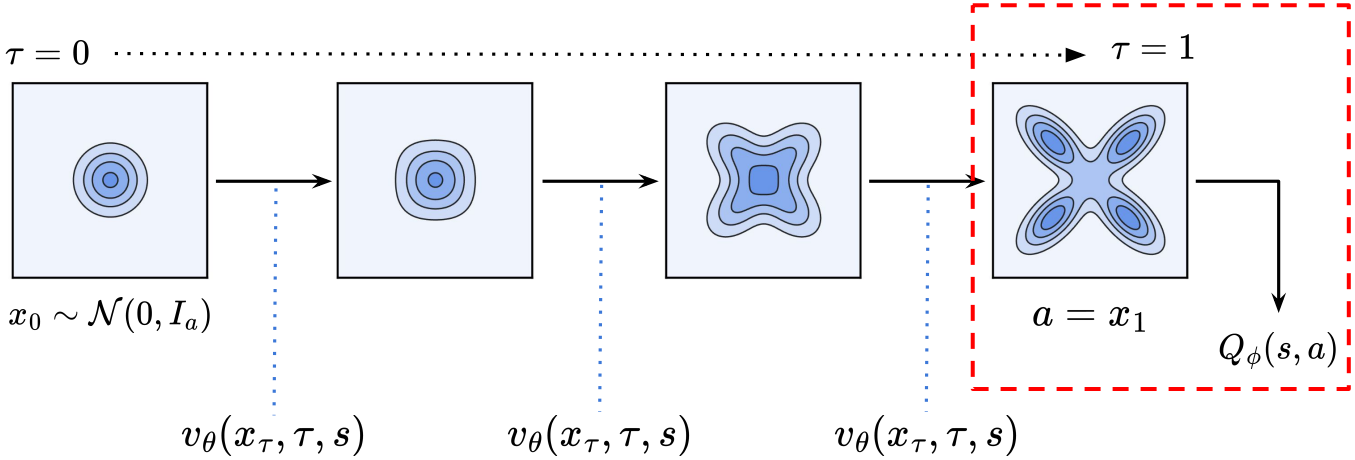
2D empirical validation





Gained stability, but
Limited expressivity

What's the underlying reason?



Requires **full policy rollout** for **value maximization**

$$\mathcal{L}_{\text{actor}}(\theta) = \mathbb{E}_{\substack{(s,a)\sim\mathcal{D} \\ a^\pi \sim \pi_\theta(\cdot|s)}} \left[-Q_\phi(s, a^\pi) \right] + \alpha \mathcal{L}_{\text{CFM}}(\theta)$$

Full rollout with flow policy

Q-function defined only over clean action

Solution: Learn the **Intermediate Value**

Flow policy is *deterministic*...

Flow policy gives a deterministic map from noise to action

$$x_1 := \Psi^{\pi}_{1,\tau}(x_\tau, s)$$

→ the terminal value can be attributed to the flow trajectory!

Flow-consistent Value: $V^{\pi}(s, x_\tau, \tau) := Q(s, \Psi^{\pi}_{1,\tau}(x_\tau, s))$

Construct **velocity field** steered by **inner value gradient**

$$v_{\text{target}}(x_1, x_0, \tau, s) := \underbrace{(x_1 - x_0)}_{\text{BC anchor}} + \underbrace{\frac{1}{\lambda}\nabla_{x_\tau} V^{\pi}_{\omega}(s, x_\tau, \tau)}_{\text{Value guidance}}$$

Full Algorithm: Q-Flow

Our Q-learning method with Flow policy:

1) **Outer critic** learns environmental reward

$$\mathcal{L}_Q(\phi) = \mathbb{E}_{\substack{(s,a,r,s')\sim\mathcal{D} \\ a'\sim\pi_\theta}}\left[\left(Q_\phi(s,a) - r - \gamma \bar{Q}_{\bar{\phi}}(s',a')\right)^2\right]$$

2) **Inner critic** learns **flow-consistent value**

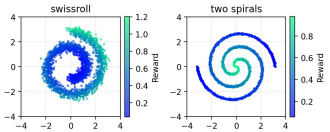
$$\mathcal{L}_V(\omega) = \mathbb{E}_{\substack{\tau\sim\mathcal{U}(0,1), x_0\sim\mathcal{N}(0,I_a), \\ (s,a=x_1)\sim\mathcal{D}}}\left[\left(V^\pi_\omega(s,x_\tau,\tau) - \bar{Q}_{\bar{\phi}}(s,\Psi^\pi_{1,\tau}(x_\tau,s))\right)^2\right]$$

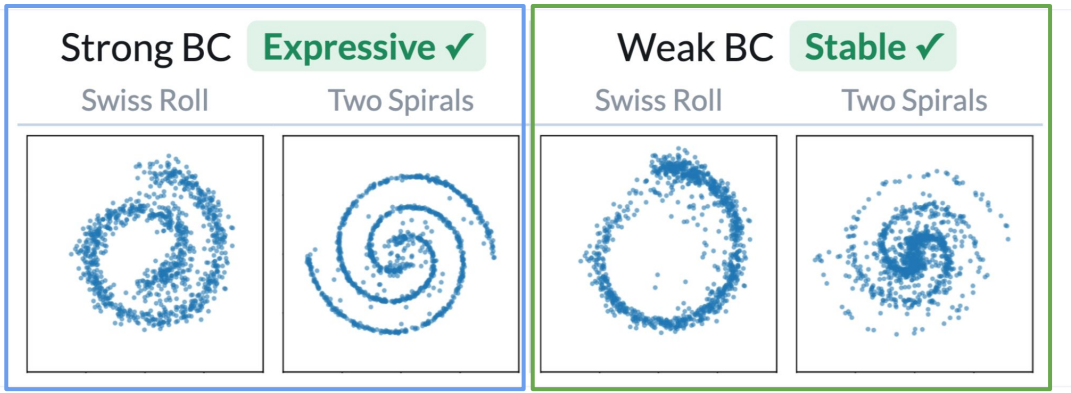
3) **Flow policy** learns velocity field steered by **inner value gradient**

$$\mathcal{L}_\pi(\theta) = \mathbb{E}_{\substack{\tau\sim\mathcal{U}(0,1), x_0\sim\mathcal{N}(0,I_a), \\ (s,a=x_1)\sim\mathcal{D}}}\left[\left\| v_\theta(x_\tau,\tau,s) - v_{\text{target}}(x_1,x_0,\tau,s)\right\|^2\right]$$

Q-Flow overcomes the dilemma

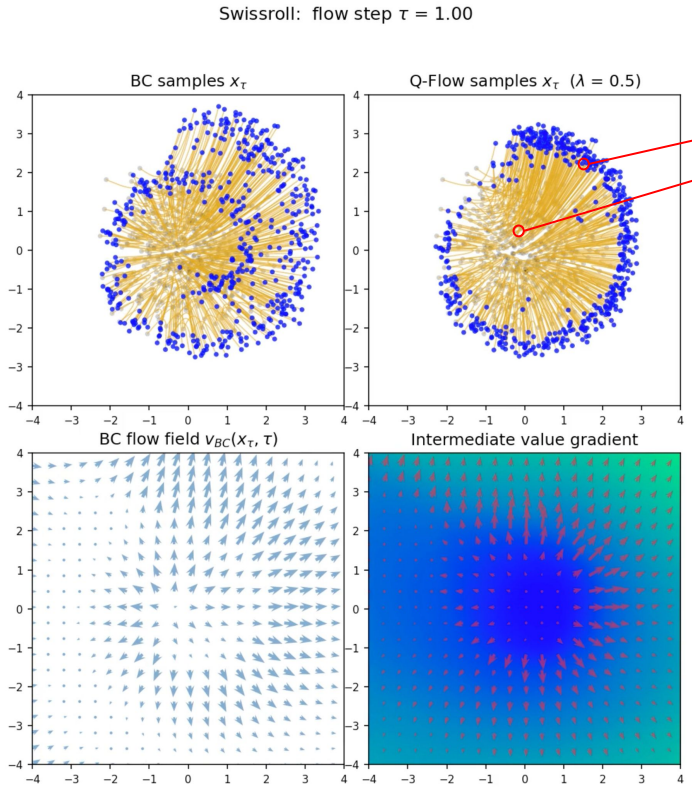
2D empirical validation



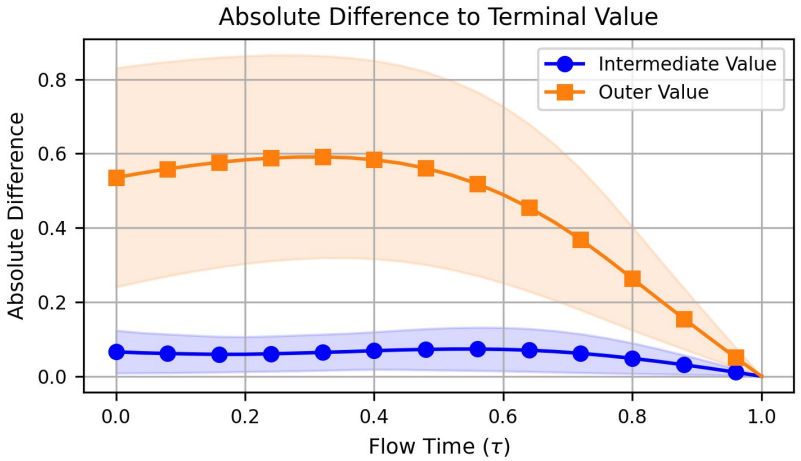


Preserves **full expressivity** while gaining **optimization stability**!

Sample Evolution over Flow



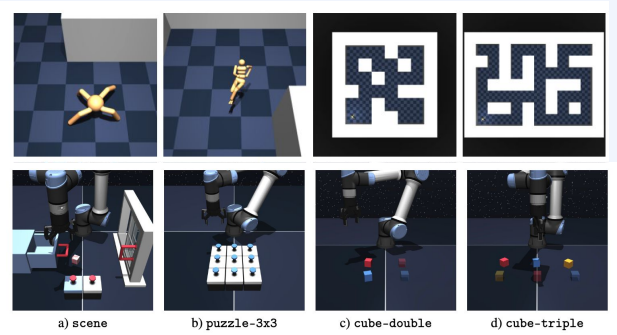
$$\left|V^\pi(s, x_1, 1) - V^\pi(s, x_\tau, \tau)\right|$$



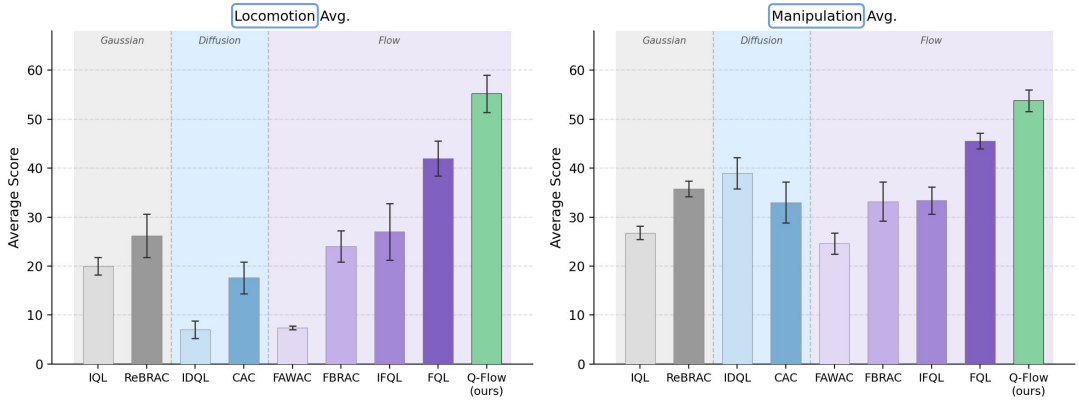
Flow-consistency is successfully learned!

Main Results:

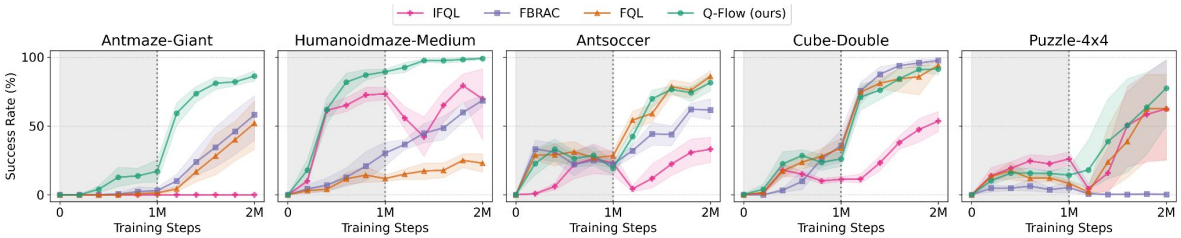
Offline & Offline-to-Online RL



OGBench task suite



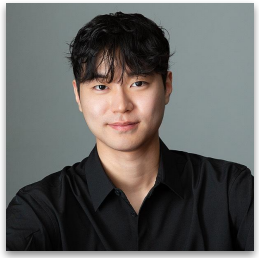
Offline RL results



Offline-to-Online RL results

Thank you!

Byeongguk Jeon



Seonghyeon Ye

Kimin Lee

Minjoon Seo

For more details,
please check out our **paper** and **project page!**

Paper

Project Page