Flow Matching for Offline Reinforcement Learning with Discrete Actions

Accepted at NeurIPS 2026

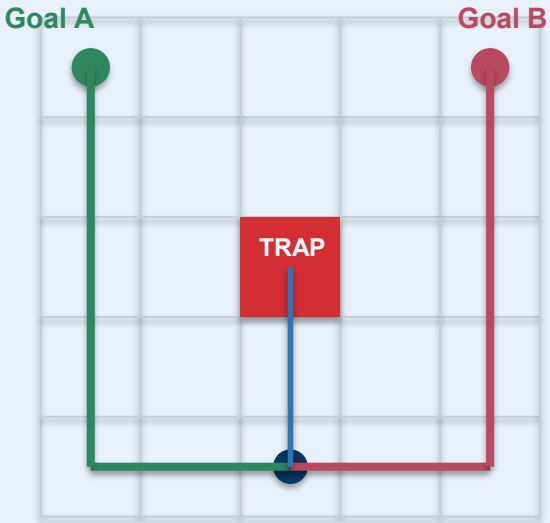
Fairoz Nower Khan[1], Nabuat Zaman Nahim[1], Ruiquan Huang[1], Haibo Yang[2], Peizhong Ju[1]

[1] University of Kentucky, Lexington, KY, USA
[2] Rochester Institute of Technology, Rochester, NY, USA

Why Discrete Flow Matching for Offline RL?

Continuous interpolation can be inappropriate for distinct, discrete decisions

A motivating example: two valid behaviors



Discrete routes preserve two modes

averaging can enter the trap.

Continuous flows

ODEs use a vector field and action-space derivatives

Discrete actions

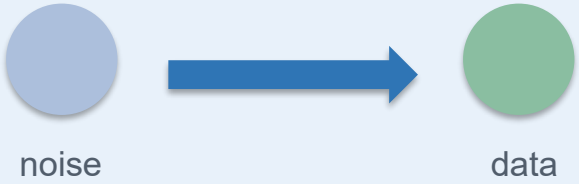
No continuous action gradient
We need jumps between actions

Central question: How can flow matching learn value-guided policies directly over discrete actions?

From Continuous flows to Discrete jumps

Replace the ODE velocity field with a continuous-time Markov chain (CTMC)

Continuous flow matching



learn velocity $v_\theta(x, t)$

ODE transport in Euclidean space

QDFM: discrete rate field



learn jump rates $u_{\theta,t}(a', a \mid s, \omega)$

probability mass moves between valid actions

Conditional path: $p_{t|Z}(a) = (1-t)\delta(a, A_0) + t\delta(a, A_1)$ | Ideal endpoint jump rate: $\frac{1}{1-t}$

How Q-guided training works

Three training phases move the learned policy toward high value dataset supported actions

1 Flow warm-up

Match unweighted discrete flows
to learn a behavior-supported policy

2 Critic learning

Fit a vector-valued Q critic
using offline transitions

3 Policy improvement

Sample endpoints, give high-Q
endpoints larger training weights

Q-weighted endpoints: $w_j \propto \exp\left(\hat{\beta}\, Q_\omega(s, A_1^{(j)})\right)$ where $Q_\omega(s,a) = \langle \omega, Q_\psi(s,a) \rangle$

Theory with precise scope

Fixed weights: conditional and guided-marginal gradients agree
Oracle single-step theory: under idealized assumptions, the oracle update recovers the KL-regularized Boltzmann policy

One model for multiple objectives and agents

Preference conditioning and factorized CTMC dynamics broaden the framework

Multi-objective offline RL

Preference $\omega = (\omega_1, \ldots, \omega_K)$



Vector critic $Q_\psi(s, a)$

$$Q_\omega = \langle \omega, Q_\psi \rangle$$

Change preferences at test time
without training separate policies

Multi-agent offline RL

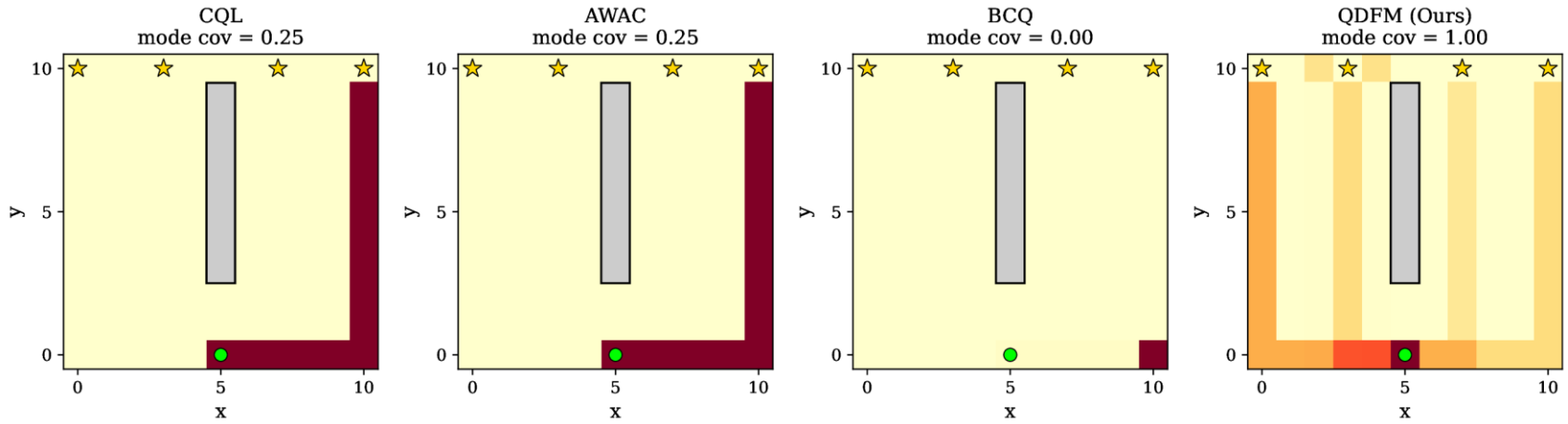


Asynchronous one-agent jumps

Rate complexity scales with $\Sigma_i |A_i|$
instead of the joint-action product

Result 1: preserve multiple successful behaviors

Four-goal gridworld: average return alone does not reveal mode collapse



Standard baselines 0.00 - 0.25

QDFM mode coverage 1.00

Strong baselines can achieve similar return while selecting fewer distinct successful goals

Result 2: preference trade-offs and benchmarks

Preference coverage and competitive returns across distinct evaluation settings

Multi-objective preference coverage

Deep Sea Treasure

15 vs 1

distinct non-dominated solutions
QDFM vs scalarized CQL

Resource Gathering

HV ratio: 0.82 vs 0.67

General benchmark performance

Quantized MuJoCo

Competitive across six tasks
using 256 discretized actions

SMAC multi-agent

Competitive return but not
overall superiority to baselines

MO-Hopper: QDFM HV $= 2.085 \times 10^7$ vs MODULI $= 2.025 \times 10^7$

Results: Quantized MuJoCo and SMAC

Return comparisons across five seeds; the best baseline in each row is selected by mean return

Quantized MuJoCo (256 actions)

Task / dataset	Baseline	Base return	QDFM
HalfCheetah expert	BCQ	10.33 ± 5.49	**1.32 ± 1.09**
HalfCheetah medium	IQL	22.83 ± 19.32	**6.02 ± 2.00**
Hopper expert	IQL	16.94 ± 13.94	**17.74 ± 7.61**
Hopper medium	BCQ	36.18 ± 26.22	**37.79 ± 16.11**
Walker2d expert	ACTIVE-I	4.77 ± 1.82	**4.85 ± 2.34**
Walker2d medium	IQL	13.11 ± 9.42	**14.01 ± 9.88**

SMAC multi-agent

Task / dataset	Baseline	Base return	QDFM
2s3z medium-replay	CFCQL	16.79 ± 0.27	**17.06 ± 0.39**
5m_vs_6m expert	CFCQL	15.05 ± 0.67	**14.99 ± 0.19**
5m_vs_6m medium	CFCQL	11.59 ± 1.23	**11.58 ± 1.55**
5m_vs_6m medium-replay	BC	6.89 ± 0.11	**6.94 ± 0.25**
5m_vs_6m poor	OMAR	3.43 ± 0.15	**1.67 ± 0.29**

Best baseline means per task; values are mean ± standard deviation. Source: paper Tables 1 and 4.

Key takeaways

What QDFM contributes and what remains to be improved

Discrete generative policies

1

CTMC jumps replace continuous ODE action flows.

Value-guided learning

2

Q-weighted endpoint training with an oracle theory result.

Diverse, flexible decisions

3

Multimodal choices, changing preferences, and multi-agent RL.

Limitations: CTMC sampling has extra inference cost; exact policy recovery is only an oracle single-step result.