Flow Actor-Critic for Offline Reinforcement Learning

Jongseong Chae[1] Jongeui Park[1] Yongjae Shin[1]
Gyeongmin Kim[1] Seungyul Han[2] Youngchul Sung[1]

[1]KAIST [2]UNIST

ICLR 2026

Offline RL

- **Offline RL** aims to seek an optimal policy from a pre-collected dataset without environment interactions
- Main Challenge: offline RL suffers from value overestimation for out-of-distribution (OOD) actions due to its limited dataset coverage.
- Previous methods based on Support-Constrained Policy Optimization:

$$\max_{\pi} \quad \mathbb{E}_{s\sim\mathcal{D},a\sim\pi}\left[Q(s,a)\right] - \lambda\mathbb{E}_{s\sim\mathcal{D}}\left[D(\pi(\cdot|s),\beta(\cdot|s)\right], \tag{1}$$

where D is some distance or divergence measure, $\mathcal{D}$ is the dataset, and β is an unknown underlying behavior policy.

Offline RL with Expressive Policies

- As offline RL datasets accumulate and grow diverse, their behavioral distributions $\beta(\cdot|s)$ become complicated and sometimes multi-modal.
- To capture such behavioral distributions, some existing offline RL methods adopted expressive models as their policies.
- Flow Q-learning (FQL[1]) uses a flow matching model as a proxy model $\hat{\beta}_\psi$ of behavior policy β:

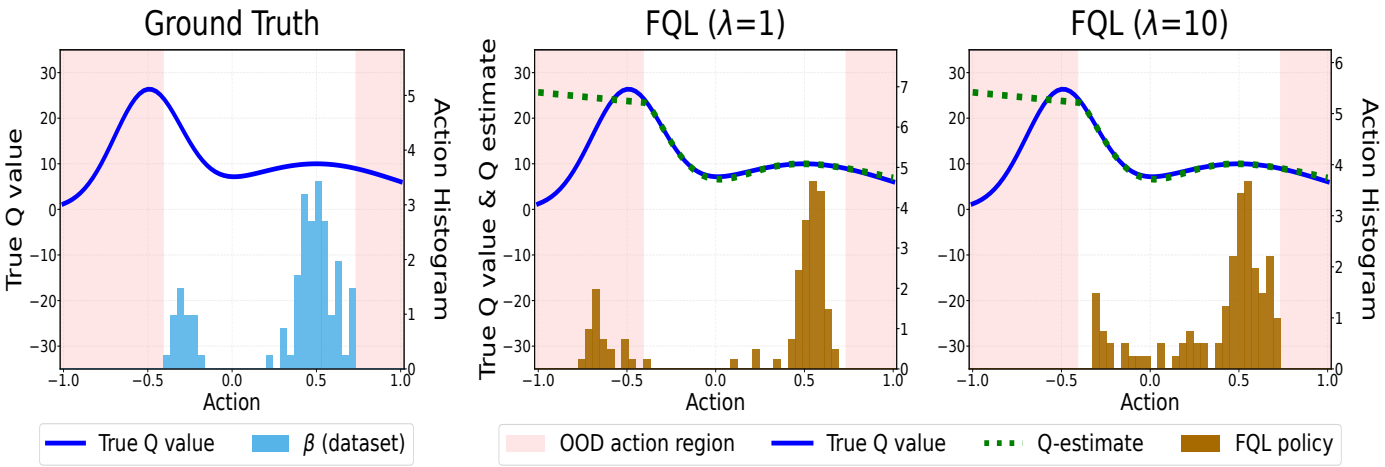
$$\min_{\hat{\beta}_\psi} \quad \mathbb{E}_{(s,a)\sim\mathcal{D},z\sim\mathcal{N}(0,I),u\sim\text{Unif}([0,1])}\left[\|v_\psi((1-u)z+ua;s,u)-(a-z)\|_2^2\right] \tag{2}$$

$$\max_{\pi_\theta} \quad \mathbb{E}_{s\sim\mathcal{D},a\sim\pi_\theta}\left[Q(s,a)\right]-\lambda\mathbb{E}_{s\sim\mathcal{D},z\sim\mathcal{N}(0,I)}\left[\|a_\theta(s,z)-a_\psi(s,z)\|_2^2\right], \tag{3}$$

where $a_\theta(s,z)$ and $a_\psi(s,z)$ denotes actions generated by π_θ and $\hat{\beta}_\psi$ given (s,z), respectively.

[1] "Flow Q-learning", Seohong Park, Qiyang Li, and Sergey Levine, ICML 2025.

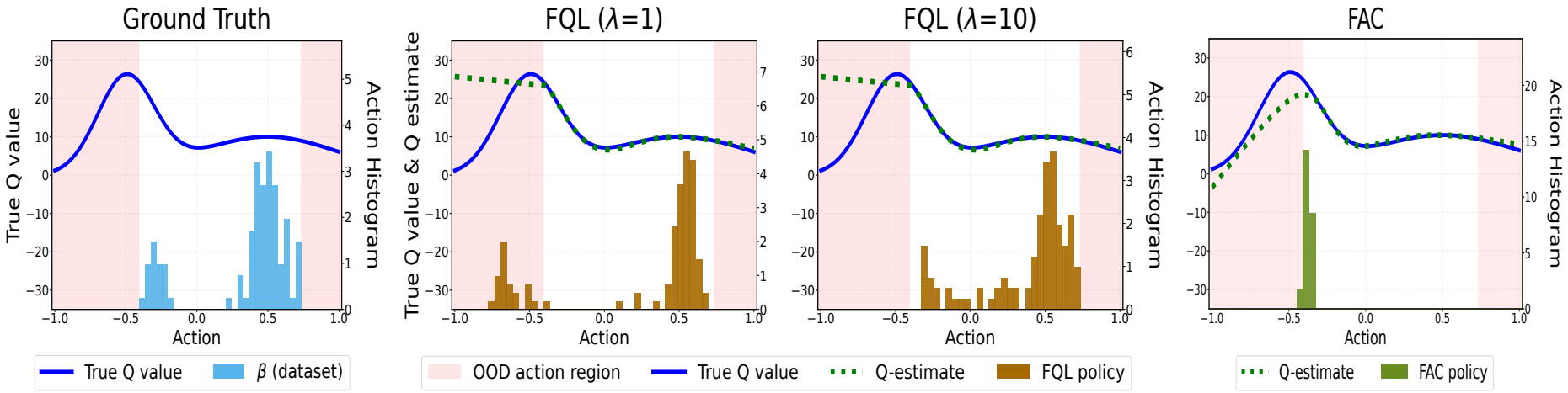
Flow Policies May Suffer from the Overestimation Issue



(a) Dataset and True Q value (b) Sampled actions and Q-estimates of FQL under different λ

- On a continuous-action bandit problem, FQL tends to overestimate in the OOD region.
  - Weak actor regularization ($\lambda = 1$): it leads to sampling OOD actions with low true Q values.
  - Strong actor regularization ($\lambda = 10$): it over-imitates sub-optimal actions.

Critic Penalization Might Address the Overestimation Issue



(a) Dataset and True Q value

(b) FQL under different $\lambda = \{1, 10\}$

(c) FAC with $\lambda = 0.1$

- Identifying OOD action region and gradually suppressing value in the OOD region
  $\Rightarrow$ We can obtain a near-optimal policy even with weaker actor regularization.

Estimating Behavior Density to Identify OOD Samples

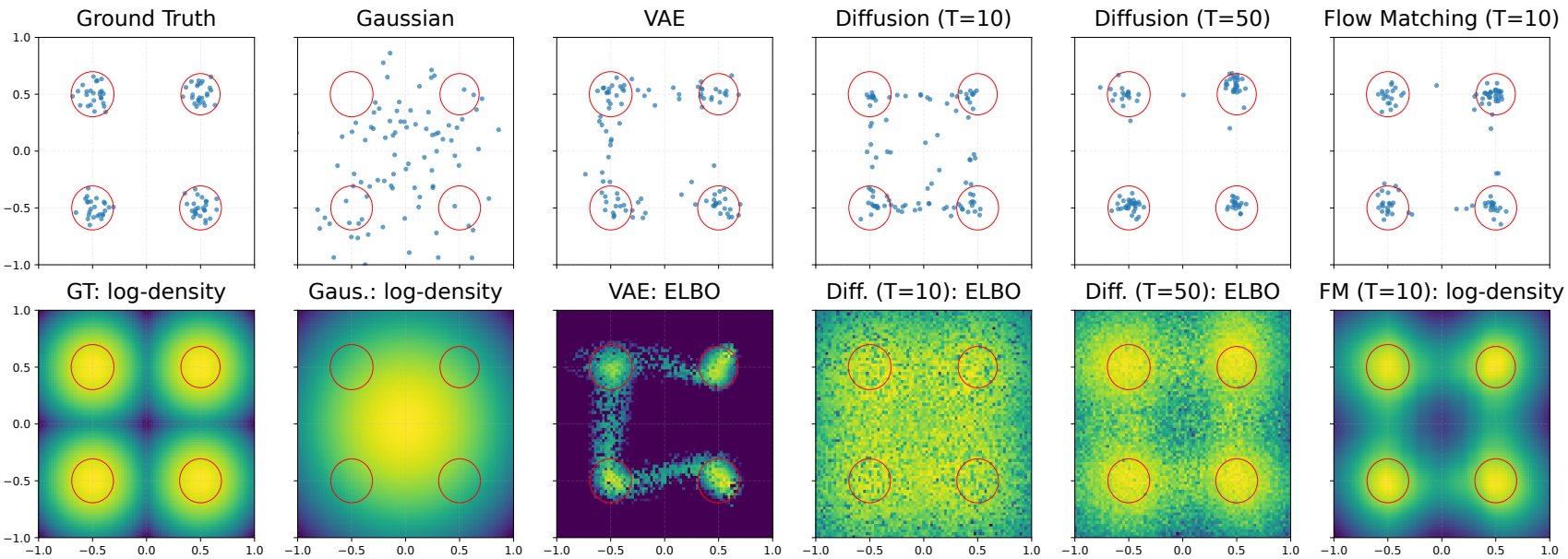


Figure 3: BC models on a synthetic four-component Gaussian mixture dataset. Top row: samples from each models. Bottom row: log-density or ELBO plot.

- Behavior cloning diffusion model with 50 denoising steps and flow matching model with 10 Euler step counts tend to generate samples within in-distribution (ID) region.
- The log-density of flow matching model yields a strong density contrast to distinguish ID regions from OOD ones.

FAC: Flow-based Critic Penalization

- Using the BC flow matching model as the proxy $\hat{\beta}_\psi$ for behavior policy β, we define

$$w^{\hat{\beta}_\psi}(s,a) = \max\left(0,\ 1 - (\hat{\beta}_\psi(a|s)/\epsilon)\right) \quad \text{for some } \epsilon > 0. \tag{4}$$

- With the weight $w^{\hat{\beta}_\psi}(s,a)$, we propose the following objective for critic learning:

$$\min_Q \quad \underbrace{\alpha \mathbb{E}_{s\sim\mathcal{D}, a\sim\pi}\left[w^{\hat{\beta}_\psi}(s,a) Q(s,a)\right]}_{\text{flow-based critic penalization}} + \underbrace{\mathbb{E}_{(s,a,s')\sim\mathcal{D}}\left[\left(Q(s,a) - \mathcal{T}^\pi Q(s,a)\right)^2\right]}_{\text{TD loss}} \tag{5}$$

FAC: Flow-based Critic Penalization

Proposition

Let β be the underlying behavior policy, $\hat{\beta}$ be our proxy for β, π be the learned actor, and Q be the value function of π. Consider the original Bellman operator $\mathcal{T}^\pi Q(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a),\, a' \sim \pi(\cdot|s')} [Q(s',a')]$. Then, in the tabular setting without function approximation, the proposed objective yields the following operator:

$$\mathcal{T}^\pi_{FAC} Q(s,a) = \begin{cases} \mathcal{T}^\pi Q(s,a) & \text{if } \hat{\beta}(a|s) \geq \epsilon,\ \beta(a|s) > 0 \\ \mathcal{T}^\pi Q(s,a) - \frac{\alpha}{2} \left(\frac{w^{\hat{\beta}}(s,a)\pi(a|s)}{\beta(a|s)} \right) & \text{if } \hat{\beta}(a|s) < \epsilon,\ \beta(a|s) > 0 \\ -\infty & \text{if } \beta(a|s) = 0. \end{cases}$$

unless $\beta(a|s) = 0$ and $w^{\hat{\beta}}(s,a) = 0$ simultaneously.

- The operator of FAC is a γ-contraction operator on the support of behavior policy β.
- The fixed point of the operator provides unbiased Q-values for actions with $\hat{\beta}_\psi \geq \epsilon$ and under-estimated (conservative) Q-values for actions with $\hat{\beta}_\psi < \epsilon$.

FAC: Flow-based Critic Penalization

- Determining ϵ in the proposed critic penalization is important, and it can be challenging to determine a fixed and appropriate ϵ across different datasets.
- We consider two dataset-driven methods for ϵ design:
  1. Batch-adaptive threshold: $\hat{\beta}_\psi(a|s)$ of mini-batch samples (s, a)
     $\rightarrow$ This threshold adapts to local coverage and multi-modality.
  2. Dataset-wide constant threshold: $\min_{(s,a)\in\mathcal{D}} \hat{\beta}_\psi(a|s)$
     $\rightarrow$ This threshold does not exclude actual ID region of the dataset.

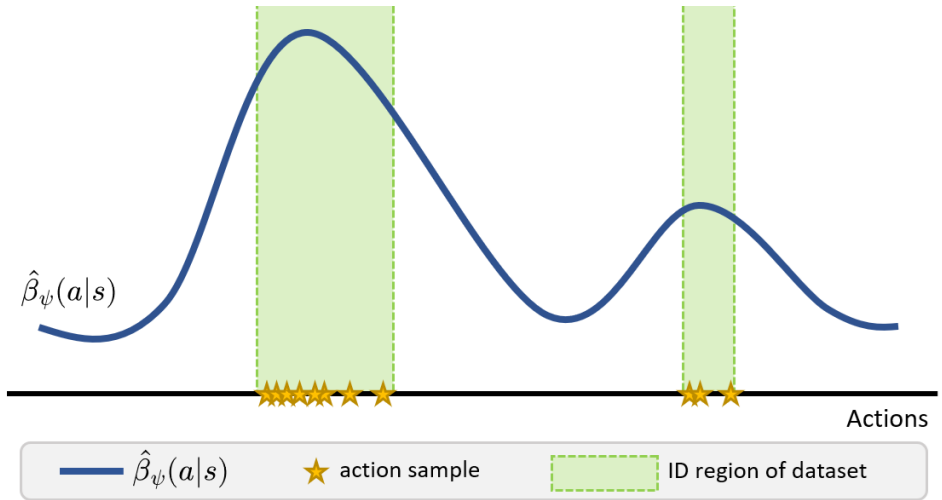


Figure 4: Selecting fixed and appropriate ϵ is non-trivial.

FAC: Flow-based Actor Regularization

- We adopt the one-step flow policy π_θ optimization of FQL:

$$\min_{\hat{\beta}_\psi} \quad \mathbb{E}_{(s,a)\sim\mathcal{D},z\sim\mathcal{N}(0,I),u\sim\text{Unif}([0,1])}\left[\|v_\psi((1-u)z+ua;s,u)-(a-z)\|_2^2\right] \tag{6}$$

$$\max_{\pi_\theta} \quad \mathbb{E}_{s\sim\mathcal{D},a\sim\pi_\theta}\left[Q(s,a)\right]-\lambda\mathbb{E}_{s\sim\mathcal{D},z\sim\mathcal{N}(0,I)}\left[\|a_\theta(s,z)-a_\psi(s,z)\|_2^2\right], \tag{7}$$

  where $a_\theta(s,z)$ and $a_\psi(s,z)$ denotes actions generated by π_θ and $\hat{\beta}_\psi$ given (s,z), respectively.

- Therefore, we leverage the expressive flow behavior proxy $\hat{\beta}_\psi$ not only for actor regularization but also for conservative critic penalization.

Performance on OGBench

Table 1: Evaluation on OGBench. For each task category, we report the final performance of our method.

	Gaussian Policies			Diffusion Policies			Flow Policies				
Task Category	**BC**	**IQL**	**ReBRAC**	**IDQL**	**SRPO**	**CAC**	**FAWAC**	**FBRAC**	**IFQL**	**FQL**	**FAC (Ours)**
antmaze-large-navigate	10.6	53.4	80.8	20.8	10.6	32.8	6.4	60.2	28.0	78.6	**92.6**±2.5
antmaze-giant-navigate	0.2	4.0	**26.2**	0.0	0.0	0.0	0.0	3.8	2.6	8.6	23.0±5.1
humanoidmaze-medium-navigate	2.0	32.8	21.8	0.8	1.4	52.8	19.4	38.4	60.4	57.4	**75.6**±3.6
humanoidmaze-large-navigate	0.4	2.4	2.6	0.6	0.2	0.6	0.2	2.2	**11.0**	4.2	8.3±5.5
antsoccer-arena-navigate	1.0	8.4	0.0	11.8	1.0	1.8	12.4	16.0	33.2	60.2	**67.7**±2.9
cube-single-play	5.4	83.0	90.6	94.6	79.6	85.2	81.2	78.6	79.2	95.8	**98.8**±1.4
cube-double-play	1.6	6.4	12.2	14.6	1.4	5.8	5.2	15.0	14.0	28.6	**33.1**±5.7
scene-play	4.6	27.6	40.6	46.2	20.0	39.8	29.8	44.8	30.4	55.8	**71.3**±7.6
puzzle-3x3-play	1.8	9.0	21.6	10.4	17.8	19.4	6.4	14.0	19.0	29.6	**100.0**±0.0
puzzle-4x4-play	0.2	7.4	14.0	29.2	10.6	14.8	0.4	13.2	25.2	17.2	**32.3**±6.5
visual manipulation (pixel-based)	-	41.6	59.8	-	-	-	-	22.8	50.2	65.4	**76.0**±3.9
Average (state-based)	2.8	23.4	31.0	22.9	14.3	25.3	16.1	28.6	30.3	43.6	**60.3**

Performance on D4RL

Table 2: Evaluation on D4RL. We report the final performance of our method.

	Gaussian Policies							Diffusion Policies			Flow Policies		
MuJoCo Tasks	**TD3+BC**	**IQL**	**CQL**	**MCQ**	**EPQ**	**SPOT**	**ReBRAC**	**IDQL**	**SRPO**	**CAC**	**QIPO-OT**	**FQL**	**FAC (Ours)**
halfcheetah-m	48.3	47.4	44.0	64.3	67.3	58.4	65.6	51.0	60.4	**69.1**	54.2	60.3±1.1	65.0±1.5
hopper-m	59.3	66.3	58.5	78.4	101.3	86.0	**102.0**	65.4	95.5	80.7	94.1	68.1±3.4	91.9±3.9
walker2d-m	83.7	78.3	72.5	**91.0**	87.8	86.4	82.5	82.5	84.4	83.1	87.6	77.2±2.5	85.2±0.9
halfcheetah-mr	44.6	44.2	45.5	56.8	**62.0**	52.2	51.0	45.9	51.4	58.7	48.0	49.3±0.5	55.4±2.7
hopper-mr	60.9	94.7	95.0	**101.6**	97.8	100.2	98.1	92.1	101.2	99.7	101.3	49.8±7.2	99.1±0.9
walker2d-mr	81.8	73.9	77.2	91.3	85.3	**91.6**	77.3	85.1	84.6	79.5	78.6	53.1±7.9	83.0±5.8
halfcheetah-me	90.7	86.7	91.6	87.5	95.7	86.9	101.1	95.9	92.2	84.3	94.5	99.6±6.5	**101.9**±5.6
hopper-me	98.0	91.5	105.4	**111.2**	108.8	99.3	107.0	108.6	100.1	100.4	108.0	83.1±17.0	104.2±4.6
walker2d-me	110.1	109.6	108.8	**114.2**	112.0	112.0	111.6	112.7	114.0	110.4	110.9	106.1±1.8	108.4±0.6
Average	75.3	77.0	77.6	88.5	**90.9**	85.9	88.5	82.1	87.1	85.1	86.4	71.8	88.2

	Gaussian Policies							Diffusion Policies			Flow Policies		
Antmaze Tasks	**TD3+BC**	**IQL**	**CQL**	**MCQ**	**EPQ**	**SAC-RND**	**ReBRAC**	**IDQL**	**SRPO**	**CAC**	**QIPO-OT**	**FQL**	**FAC (Ours)**
umaze	78.6	87.5	74.0	98.3	**99.4**	97.0	97.8	94.0	97.1	75.8	93.6	96.0	98.5±3.0
umaze-diverse	71.4	62.2	84.0	80.0	78.3	66.0	88.3	80.2	82.1	77.6	76.1	89.0	**93.5**±6.0
medium-play	10.6	71.2	61.2	52.5	85.0	38.5	84.0	84.5	80.7	56.8	80.0	78.0	**88.0**±9.6
medium-diverse	3.0	70.0	53.7	37.5	**86.7**	74.7	76.3	84.8	75.0	0.0	86.4	71.0	85.0±7.3
large-play	0.2	39.6	15.8	2.5	40.0	43.9	60.4	63.5	53.6	0.0	55.5	84.0	**90.0**±4.3
large-diverse	0.0	47.5	14.9	7.5	36.7	45.7	54.4	67.9	53.6	0.0	32.1	83.0	**88.0**±6.0
Average	27.3	63.0	50.6	46.4	71.0	61.0	76.9	79.2	73.7	35.0	70.6	83.5	**90.5**

	Gaussian Policies							Diffusion Policies			Flow Policies	
Adroit Tasks	**TD3+BC**	**IQL**	**CQL**	**MCQ**	**EPQ**	**SAC-RND**	**ReBRAC**	**IDQL**	**SRPO**	**CAC**	**FQL**	**FAC (Ours)**
pen-human	81.8	81.5	37.5	68.5	83.9	5.6	**103.5**	76.0	69.0	64.0	53.0	73.9±14.7
pen-cloned	61.4	77.2	39.2	49.4	91.8	2.5	91.8	64.0	61.0	56.0	74.0	**103.2**±11.1
door-human	-0.1	3.1	9.9	2.3	**13.2**	0.0	0.0	6.0	3.0	5.0	0.0	5.5±3.3
door-cloned	0.1	0.8	0.4	1.3	**5.8**	0.2	1.1	0.0	0.0	1.0	2.0	4.1±3.9
hammer-human	0.4	2.5	4.4	0.3	3.9	-0.1	0.2	2.0	1.0	2.0	1.0	**8.6**±5.4
hammer-cloned	0.8	1.1	2.1	1.4	**22.8**	0.1	6.7	2.0	2.0	1.0	11.0	11.1±11.2
relocate-human	-0.2	0.1	0.2	0.1	0.3	0.0	0.0	0.0	0.0	0.0	0.0	**0.6**±0.5
relocate-cloned	-0.1	0.2	-0.1	0.0	0.1	0.0	**0.9**	0.0	0.0	0.0	0.0	0.5±0.4
Average	18.0	20.8	11.7	15.4	**27.7**	1.0	25.5	18.8	17.0	16.1	17.6	25.9

FAC can Fine-tuned with Online Rollouts

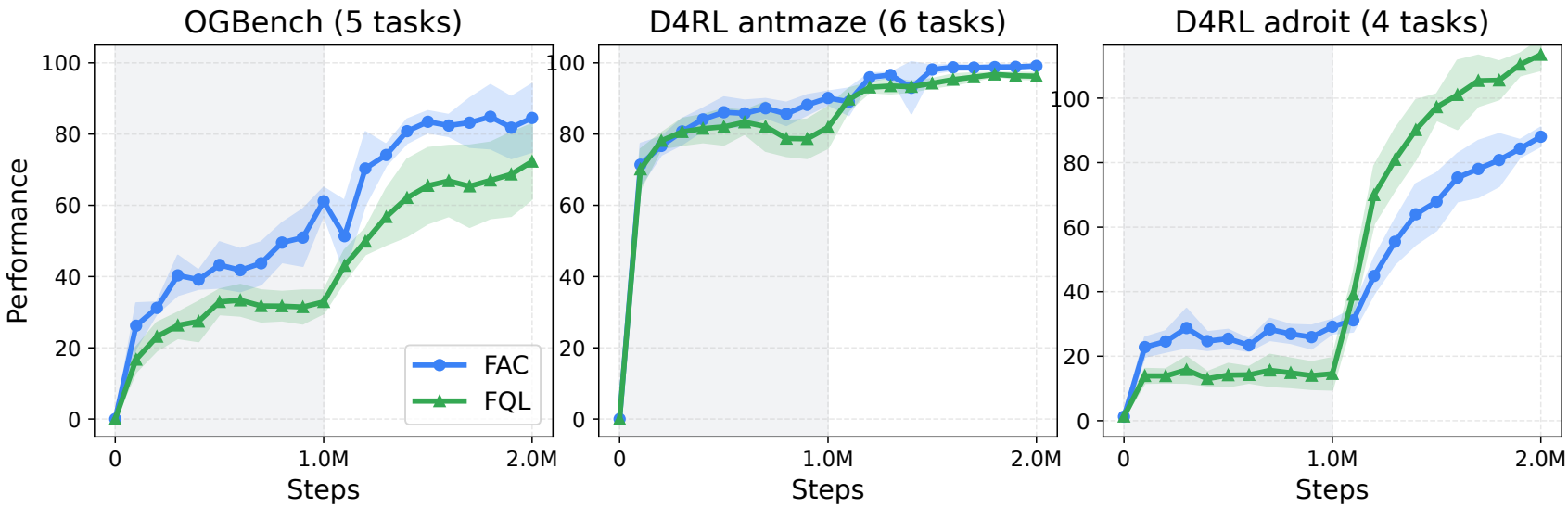


Figure 5: Offline-to-online results.

- FAC avoids excessively suppressing high Q-value neighborhoods near the datatset support boundary.
- This may enable exploration in such high Q-value regions.

Final Remarks

Takeaways

- **Joint actor-critic design:** FAC leverages the flow behavior proxy for both actor regularization and critic penalization.
- **Density-aware critic learning:** The flow behavior proxy (BC flow matching model) provides strong density estimates, enabling identification of poorly supported actions.
- **Support-aware optimization:** By suppressing Q-values in low-density regions while preserving them in well-supported regions, FAC achieves reliable policy optimization.

Possible Limitations

- **Proxy fidelity dependence:** The effectiveness of FAC depends on the quality of the learned flow behavior proxy and its density estimation.
- **Lack of Explicit Exploration:** No mechanism for beneficial exploration.
- **Additional computation:** Compared to actor regularization-only flow-based methods, FAC introduces extra cost for behavior density evaluation.

Thank you!