Guided Flow Policy: Learning from High-Value Actions in Offline RL

Franki Nguimatsia Tiofack[1,*], Théotime Le Hellard[1,*], Fabian Schramm[1,3], Nicolas Perrin-Gilbert[3] and Justin Carpentier[1]

[1]Willow, Inria - Département d'Informatique de l'École normale supérieure ENS, PSL Research University,
[2]ISIR, Sorbonne University, franki.nguimatsia-tiofack@inria.fr, theotime.le-hellard@inria.fr

Offline RL

Objective: train an agent from a static dataset $\mathcal{D} = \{(s, a, r, s')\}$.
Challenge: avoid extrapolation, without the ability to evaluate actions.

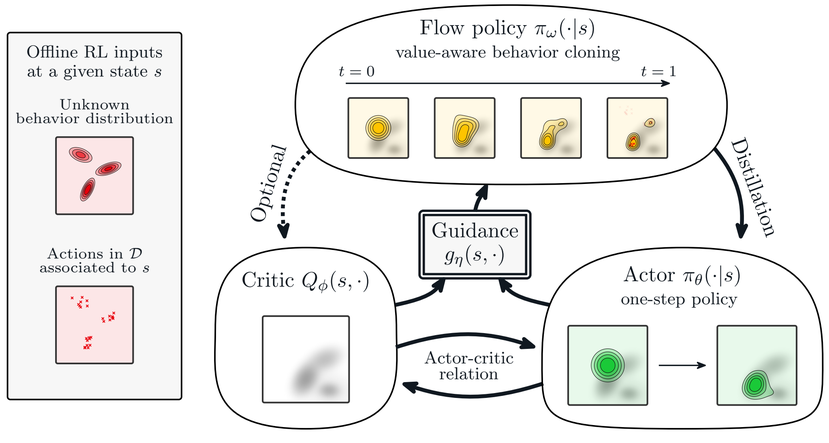
[1] and [2] add a behavior cloning term in the actor-critic framework

$$\text{TD3+BC / ReBRAC} \begin{cases} \mathcal{L}^{\mathcal{A}}(\theta) = \mathbb{E}_{(s,a)\sim\mathcal{D}, a_\theta\sim\pi_\theta(\cdot|s)}\Big[-Q_\phi(s, a_\theta) + \alpha \overbrace{\|a_\theta - a\|^2}^{\text{BC term}} \Big], \\ \mathcal{L}^{\mathcal{C}}(\phi) = \mathbb{E}_{(s,a,r,s')\sim\mathcal{D}, a'\sim\pi_\theta(\cdot|s')}\Big[\big(Q_\phi(s,a) - r - \gamma Q_{\bar{\phi}}(s', a')\big)^2 \Big] \end{cases}$$

[3] trains a flow-matching model to clone $\mathcal{D}$, and then distill it into the actor, replacing the BC term.

$\Rightarrow$ These methods risk regularizing with suboptimal dataset actions.

Overview of our approach



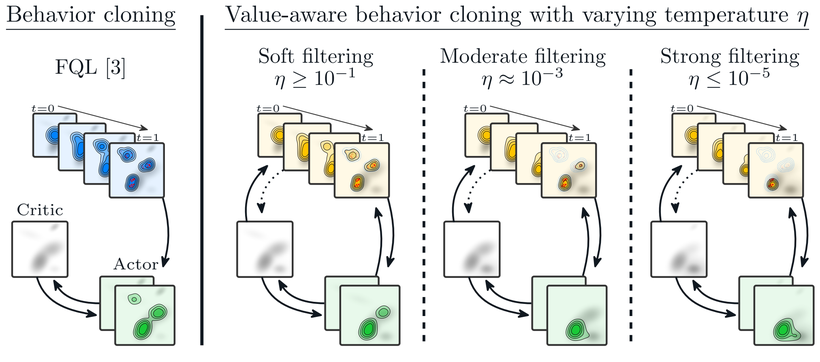
We introduce **Guided Flow Policy (GFP)**, a dual-policy framework with a bidirectional guidance mechanism between a multi-step flow-matching policy, termed Value-aware Behavior Cloning (VaBC), and a distilled one-step actor. VaBC acts as a regularizer for the actor, but, unlike standard BC, VaBC leverages the actor and its critic to prioritize cloning high-value actions from the dataset.

Guided flow policy

Algorithm 1: Guided Flow Policy (GFP)

function *Integrate* $\mu_\omega(s, z)$
  // Explicit discrete Euler integration with M steps
  for $t = 0, 1, \ldots, M-1$ **do**
    $z \leftarrow z + \frac{1}{M} v_\omega(t/M, s, z)$
  return z
while *not converged* **do**
  Sample $\{(s, a, r, s')\} \sim \mathcal{D}$
  // Step 1 - Train critic Q_ϕ
  $z' \sim \mathcal{N}(0, I_d), \quad a' = \mu_{\bar{\theta}}(s', z')$
  Update ϕ to minimize $\mathbb{E}[(Q_\phi(s, a) - r - \gamma Q_{\bar{\phi}}(s', a'))^2]$
  // Step 2 - Train the distilled one-step actor π_θ
  $z \sim \mathcal{N}(0, I_d), \quad a^{\pi_\theta} = \mu_\theta(s, z),$
  $a^{\pi_\omega} = \mu_{\bar{\omega}}(s, z)$ // Using Integrate-μ_ω, Line. 1
  Compute $\lambda = \frac{1}{\frac{1}{N}\sum |Q_\phi(s, a^{\pi_\theta})|}$ // Stop gradient a^{π_θ}
  Update θ to minimize $\mathbb{E}[-\lambda Q_\phi(s, a^{\pi_\theta}) + \alpha \|a^{\pi_\theta} - a^{\pi_{\bar{\omega}}}\|_2^2]$
  // Step 3 - Train the value-aware BC policy π_ω
  Compute $g_\eta(s, a) = \frac{\exp\left(\frac{\lambda}{\eta} Q_\phi(s,a)\right)}{\exp\left(\frac{\lambda}{\eta} Q_\phi(s,a^{\pi_\theta})\right) + \exp\left(\frac{\lambda}{\eta} Q_\phi(s,a)\right)}$ // Stop gradient a^{π_θ}
  $a_t = (1-t)\epsilon + ta$, with $\epsilon \sim \mathcal{N}(0, I_d)$ and $t \sim \mathcal{U}([0, 1))$
  Update θ to minimize $\mathbb{E}\left[g(s, a) \|v_\omega(t, s, a_t) - (a - \epsilon)\|_2^2\right]$
Output: π_θ, π_ω, Q_ϕ

Analysis of the guidance



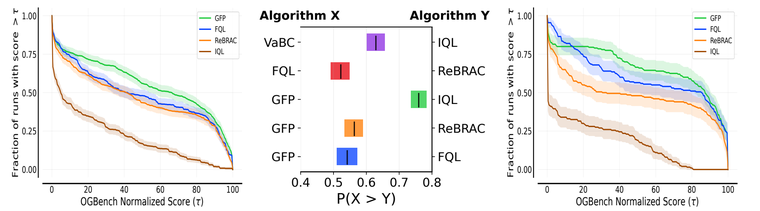
Left: Prior work uses no filtering, their BC term is reward agnostic.

Right: Our method introduces a temperature-controlled guidance mechanism, the lower the temperature, the sharper the filtering.

Extensive benchmarks

We evaluate GFP across 129 locomotion and manipulation tasks

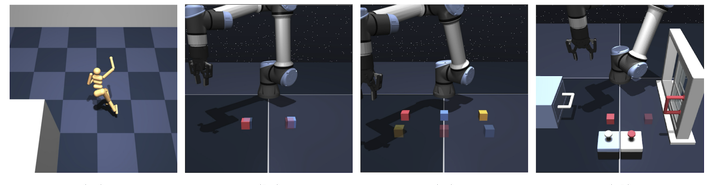
	IQL [4]	ReBRAC [2]	FQL [3]	**GFP**
OGBench (90 tasks)	20.7	46.4	48.9	**56.0**
D4RL (18 tasks)	54.0	**64.8**	**62.1**	**63.0**
Minari (21 tasks)	–	–	65.9	**74.1**



(a) (b) (c)

OGBench analysis: (a) Performance profiles for 90 tasks showing the fraction of tasks where each algorithm achieves a score above threshold τ (b) Probability of improvement P(X>Y) on the same 90 tasks (c) Focusing on 25 tasks from noisy and explore environments.

Focus on some OGBench environments, from [5]



(a) (b) (c) (d)

	IQL	ReBRAC	FQL	**GFP**
(a) humanoidmaze-large-navigate	2	12.9	6.5	**17.8**
(b) cube-double-noisy	4.5	19.6	38.2	**63.1**
(c) cube-triple-play	0.1	2.9	3.9	**15.9**
(d) scene-noisy	16.0	39.9	**59.3**	**57.5**

References

[1] Scott Fujimoto and Shixiang Shane Gu. "A minimalist approach to offline reinforcement learning". In: *Advances in neural information processing systems* 34 (2021), pp. 20132–20145.

[2] Denis Tarasov et al. "Revisiting the minimalist approach to offline reinforcement learning". In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 11592–11620.

[3] Seohong Park, Qiyang Li, and Sergey Levine. "Flow Q-Learning". In: *arXiv preprint* (2025).

[4] Ilya Kostrikov, Ashvin Nair, and Sergey Levine. "Offline reinforcement learning with implicit q-learning". In: *arXiv preprint* (2021).

[5] Seohong Park et al. "Ogbench: Benchmarking offline goal-conditioned rl". In: *arXiv preprint* (2024).