Zero-Shot Reinforcement Learning from Low Quality Data

Scott Jeen {srj38@cam.ac.uk}[1], Tom Bewley[2] & Jonathan M. Cullen[1]

[1] University of Cambridge, [2] University of Bristol

UNIVERSITY OF CAMBRIDGE

1 Motivation

- Behaviour Foundation Models (BFMs) based on forward-backward representations (FB) [1] and universal successor features (USF) [2] provide principled mechanisms for performing zero-shot task generalization.
- However, BFMs assume access to idealised (large & diverse) pre-training datasets that we can't expect for real problems.
- **Can we pre-train BFMs on realistic (small & narrow) datasets?**

2 Background

Forward-backward (FB) BFMs model the environment dynamics using *successor measures* which are the expected discounted time spent in subsets of future states:

$$M^{\pi}(s_0, a_0, S_+) := \sum_{t=0}^{T-1} \gamma^t \Pr(s_{t+1} \in S_+ | (s_0, a_0), \pi), \ \forall \ S_+ \subset \mathcal{S}.$$

Together, a forward model F and backward model B approximate successor measures for all policies

$$\begin{cases} M^{\pi_z}(s_0, a_0, X) \approx \int_X F(s_0, a_0, z)^\top B(s)\rho(\mathrm{d}s) & \forall \ s_0 \in \mathcal{S}, a_0 \in \mathcal{A}, X \subset \mathcal{S}, z \in \mathbb{R}^d, \\ \pi(s, z) \approx \max_a F(s, a, z)^\top z & \forall \ (s, a) \in \mathcal{S} \times \mathcal{A}, z \in \mathbb{R}^d \end{cases}$$

Zero-shot RL:

Pre-train on reward-free dataset $\mathcal{D} = \{(s_i, a_i, s_{i+1})\}_{i=1}^{|\mathcal{D}|}$

Infer task from $\mathcal{D}_{\text{labelled}} = \{(s_i, R_{\text{eval}}(s_i))\}_{i=1}^{10{,}000}$

3 Failure Mode on Low Quality Datasets

The FB loss relies on actions sampled from the policy, and these may not exist in the dataset (*i.e.* they can be out-of-distribution (OOD)).

$$\mathcal{L}_{\text{FB}} = \mathbb{E}_{(s_t, a_t, s_{t+1}, s_+) \sim \mathcal{D}, z \sim \mathcal{Z}}[(F(s_t, a_t, z)^\top B(s_+) - \gamma \bar{F}(s_{t+1}, \underbrace{\pi_z(s_{t+1})}_{\text{OOD}}, z)^\top \bar{B}(s_+))^2 - 2F(s_t, a_t, z)^\top B(s_{t+1})]$$

This leads to value function overestimation at OOD state-action pairs:

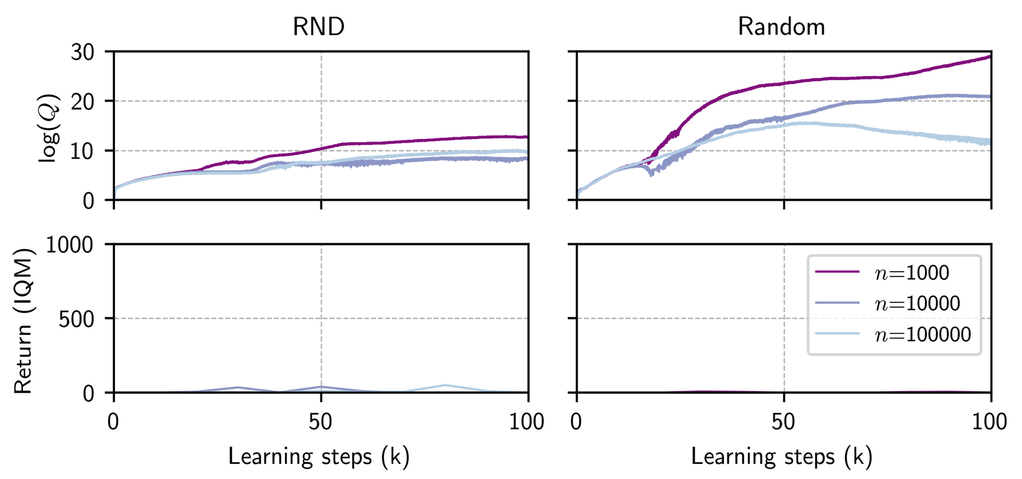


Figure 1: **FB value overestimation with respect to dataset size and quality.** Log Q values and IQM of rollout performance on all Maze tasks for RND and Random datasets.

4 *Conservative* Behaviour Foundation Models

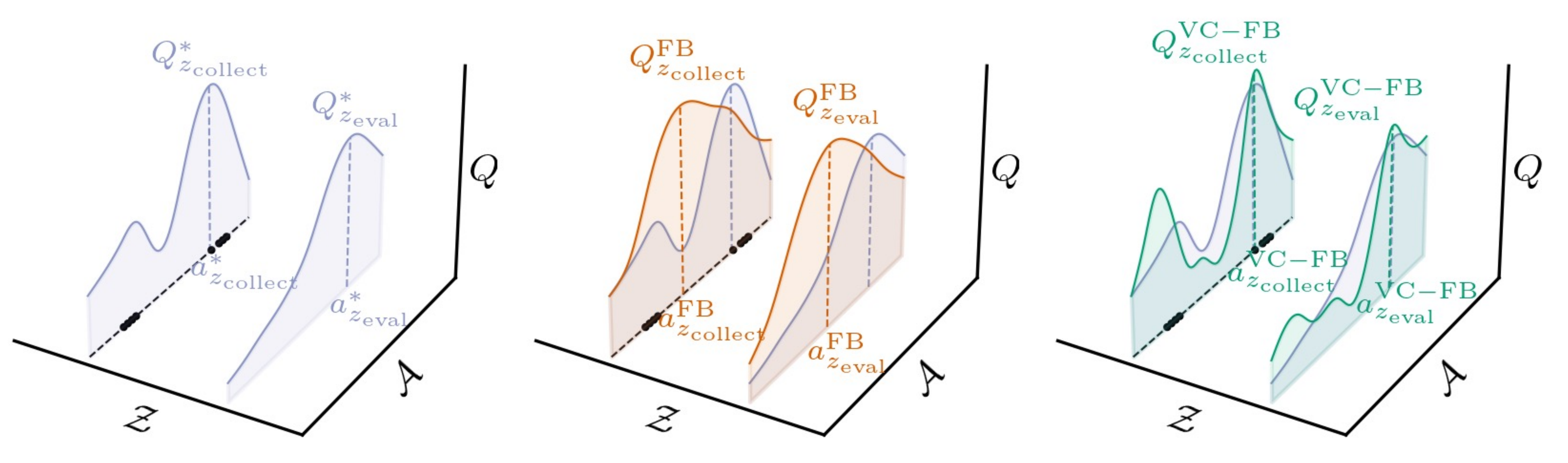


Figure 2: ***Conservative*** **BFMs.** (*Left*) Zero-shot RL methods must generalize to any task in z-space. (*Middle*) FB overestimates the value of actions not in the dataset. (*Right*) VC-FB suppresses the value of actions not in the dataset.

Value-Conservative Forward Backward Representations

$$\mathcal{L}_{\text{VC-FB}} = \alpha \cdot (\mathbb{E}_{s \sim \mathcal{D}, a \sim \mu(a|s), z \sim \mathcal{Z}}[F(s, a, z)^\top z] - \mathbb{E}_{(s,a) \sim \mathcal{D}, z \sim \mathcal{Z}}[F(s, a, z)^\top z] - \mathcal{H}(\mu)) + \mathcal{L}_{\text{FB}}$$

Measure-Conservative Forward Backward Representations

$$\mathcal{L}_{\text{MC-FB}} = \alpha \cdot (\mathbb{E}_{s \sim \mathcal{D}, a \sim \mu(a|s), z \sim \mathcal{Z}, s_+ \sim \mathcal{D}}[F(s, a, z)^\top B(s_+)] - \mathbb{E}_{(s,a) \sim \mathcal{D}, z \sim \mathcal{Z}, s_+ \sim \mathcal{D}}[F(s, a, z)^\top B(s_+)] - \mathcal{H}(\mu)) + \mathcal{L}_{\text{FB}}$$

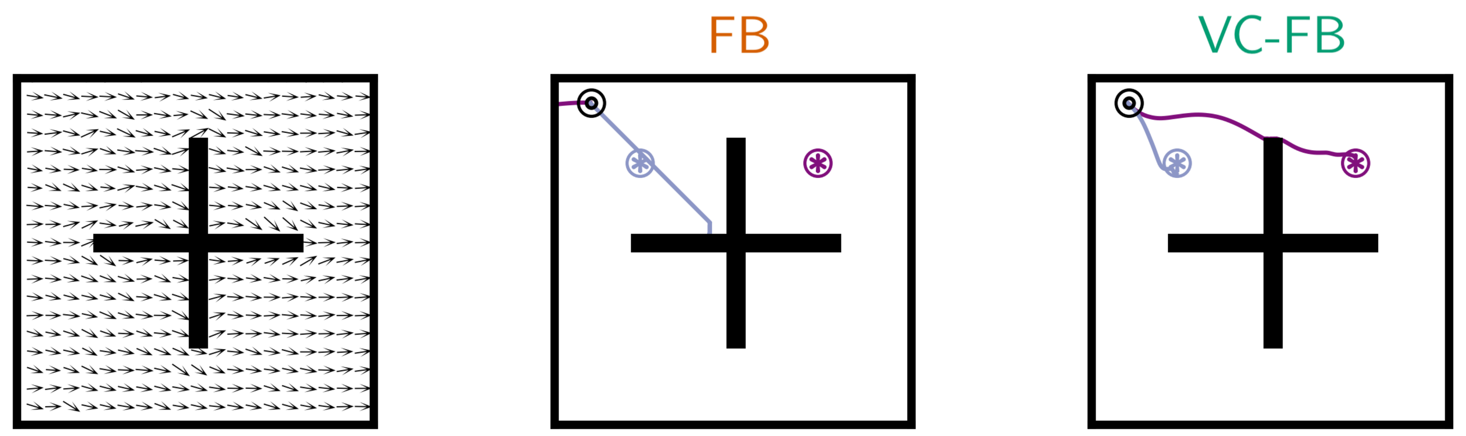


Figure 3: **Didactic example.** The agents are tasked with learning separate policies for reaching ⊛ and ⊛. (a) RND dataset with all "left" actions removed (b) Best FB rollout after 1 million steps. (c) Best VC-FB performance after 1 million learning steps.

5 Setup

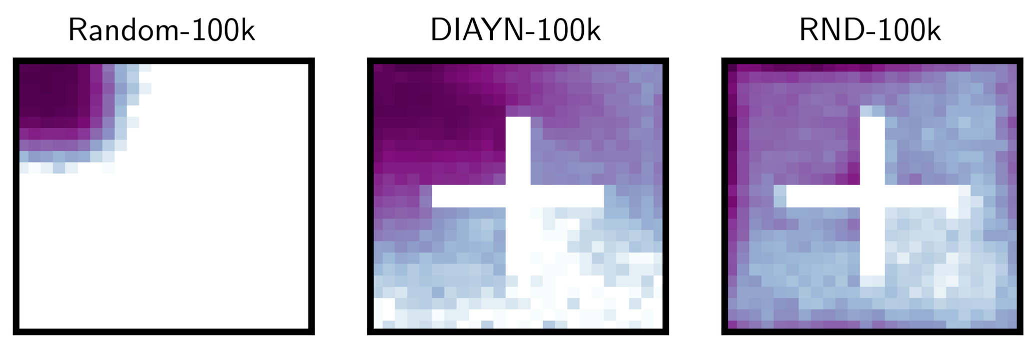
Baselines

- **Zero-shot RL**: FB, SF-LAP [5]
- **Goal-conditioned RL**: GC-IQL [6]
- **Offline RL**: CQL [7]

Environments

- ExORL & D4RL

Datasets



6 Results

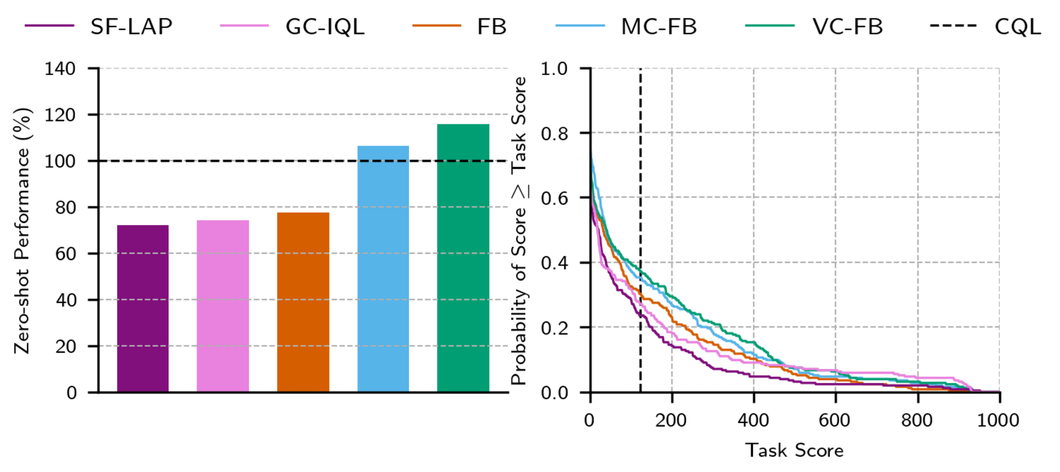


Figure 4: **Aggregate ExORL Performance.** Both conservative BFM variants stochastically dominate vanilla FB.

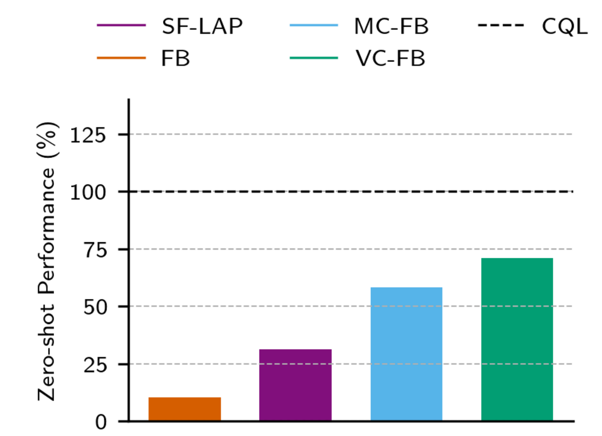


Figure 5: **D4RL Performance.** Conservative BFMs outperform vanilla BFMs, but do not match the performance of the single-task baseline.

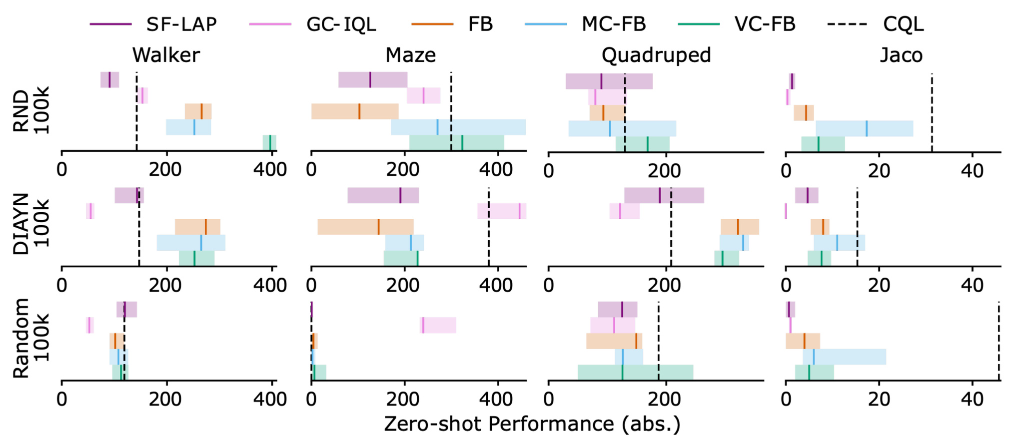


Figure 6: **ExORL Performance by dataset/domain.**

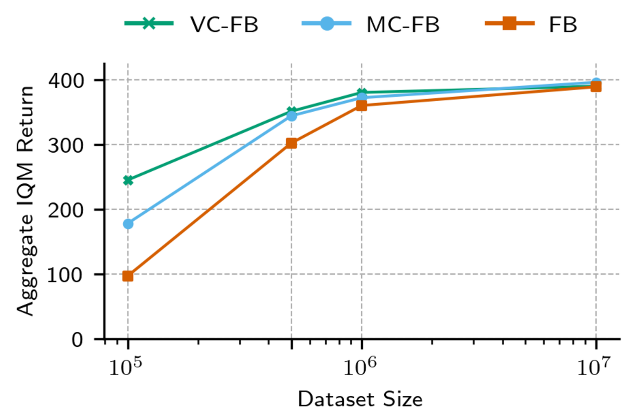


Figure 7: **Performance by RND dataset size.** The delta between vanilla FB and the conservative variants increases as dataset size decreases.

7 Limitations

- Absolute ExORL performance remains poor compared to methods trained on large/diverse datasets.
- Performance is sensitive the choice of τ which selects the degree of conservatism. IQL-style regularization would likely mitigate this. (*c.f.* D4RL performance)

References

[1] Robert Kirk, Amy Zhang, Edward Grefenstette, and Tim Rocktäschel. *A survey of zero-shot generalisation in deep reinforcement learning*. JAIR 2023

[2] Ahmed Touati, Jérémy Rapin, and Yann Ollivier. *Does zero-shot reinforcement learning exist?* ICLR 2023

[3] Seohong Park, Tobias Kreiman, and Sergey Levine. *Foundation Policies with Hilbert Representations*. ICML 2024

[4] Aviral Kumar, Justin Fu, Matthew Soh, George Tucker, and Sergey Levine. *Stabilizing off-policy q-learning via bootstrapping error reduction*. NeurIPS 2019

[5] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. *Conservative q-learning for offline reinforcement learning.* NeurIPS 2020

[6] Seohong Park, Dibya Ghosh, Benjamin Eysenbach, and Sergey Levine. *Hiql: Offline goalconditioned rl with latent states as actions.* NeurIPS 2023.

NEURAL INFORMATION PROCESSING SYSTEMS