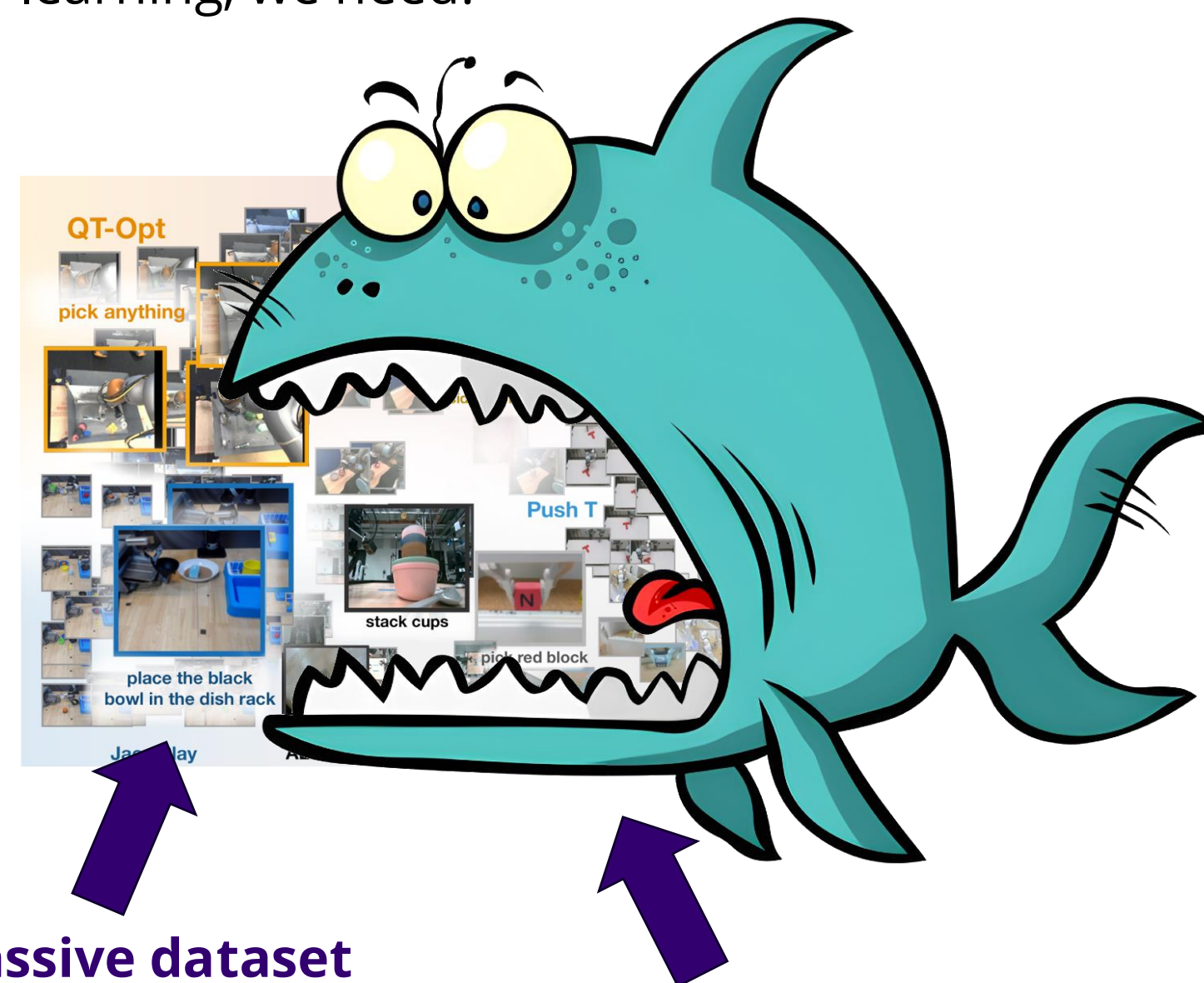


Free from Bellman Completeness: Trajectory Stitching via Model-Based Return-Conditioned Supervised Learning

Zhaoyi Zhou, Chuning Zhu, Runlong Zhou, Qiwen Cui, Abhishek Gupta, Simon S. Du



To enable scalable off-policy reinforcement learning, we need:



An algorithm that can consume the data

What algorithm should we use?

Dynamic Programming?

✓ Learn optimal policy from suboptimal data via trajectory stitching.

✗ Can diverge in practice, due to **Bellman completeness** requirement.

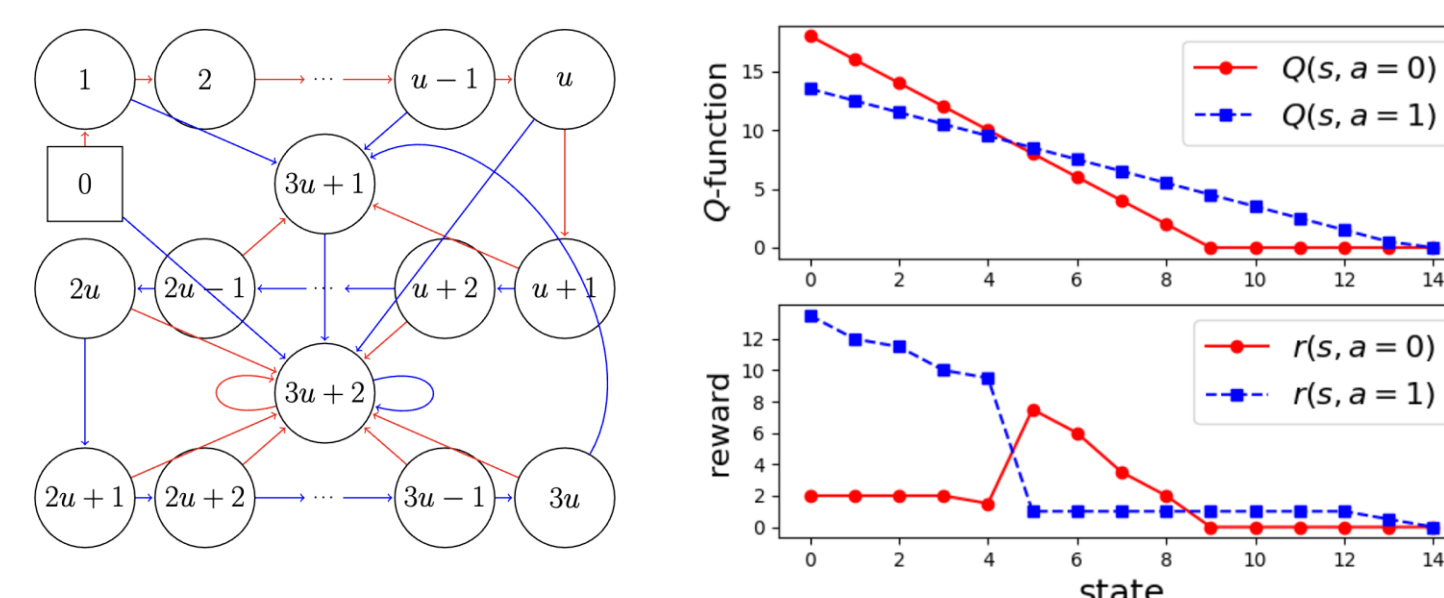
A function approximation class $\mathcal{F}$ is Bellman complete under Bellman operator $\mathcal{B}$ if

$$\max_{Q \in \mathcal{F}} \min_{Q' \in \mathcal{F}} \|Q' - \mathcal{B}Q\| = 0$$

For any Q function, we can find a Q' function that achieves zero Bellman error

Return-Conditioned Supervised Learning?

✓ Does not require Bellman completeness, only realizability.



Example: construct a class of MDPs such that using 2-layer MLPs,

- > Q-Learning needs $\Omega(|\mathcal{S}^u|)$ hidden neurons to satisfy Bellman completeness.
- > RCSL needs $O(1)$ hidden neurons to represent the optimal policy.

✗ Only recovers trajectories in the dataset. No stitching.

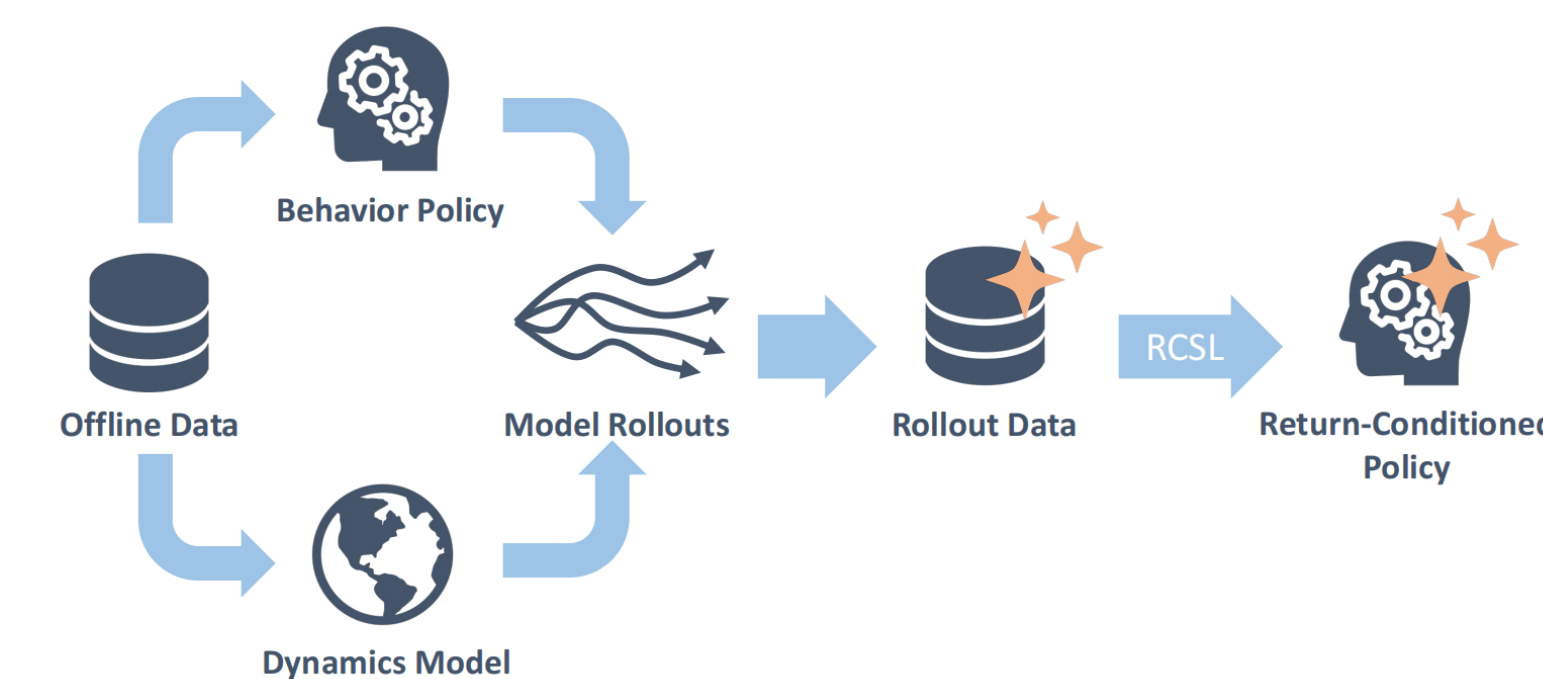
- > Holds for Markovian policy and Decision Transformer,
- > even if the environment has deterministic transition and the dataset has uniform coverage.

Can we combine the best of both worlds?

Free from Bellman completeness

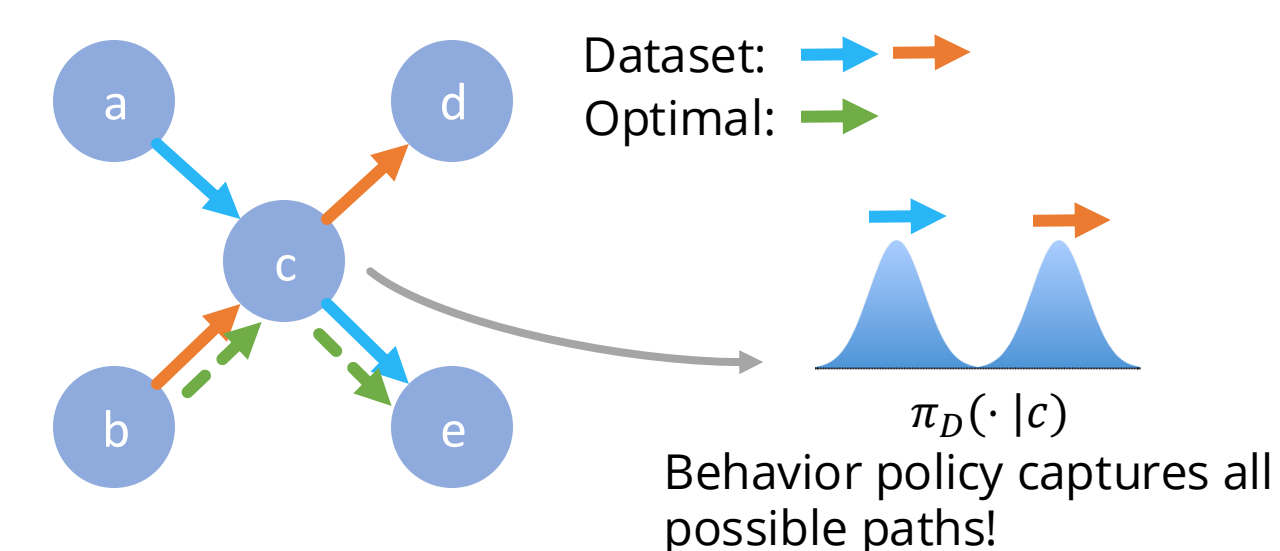
Trajectory stitching

Model-Based RCSL



- > Learn a dynamics model and a (multimodal) behavior policy from the dataset.
- > Rollout the behavior policy under the model to generate potentially optimal trajectories.
- > Add trajectories to dataset and do RCSL.

Why does this do stitching?



Experiments

Point Maze

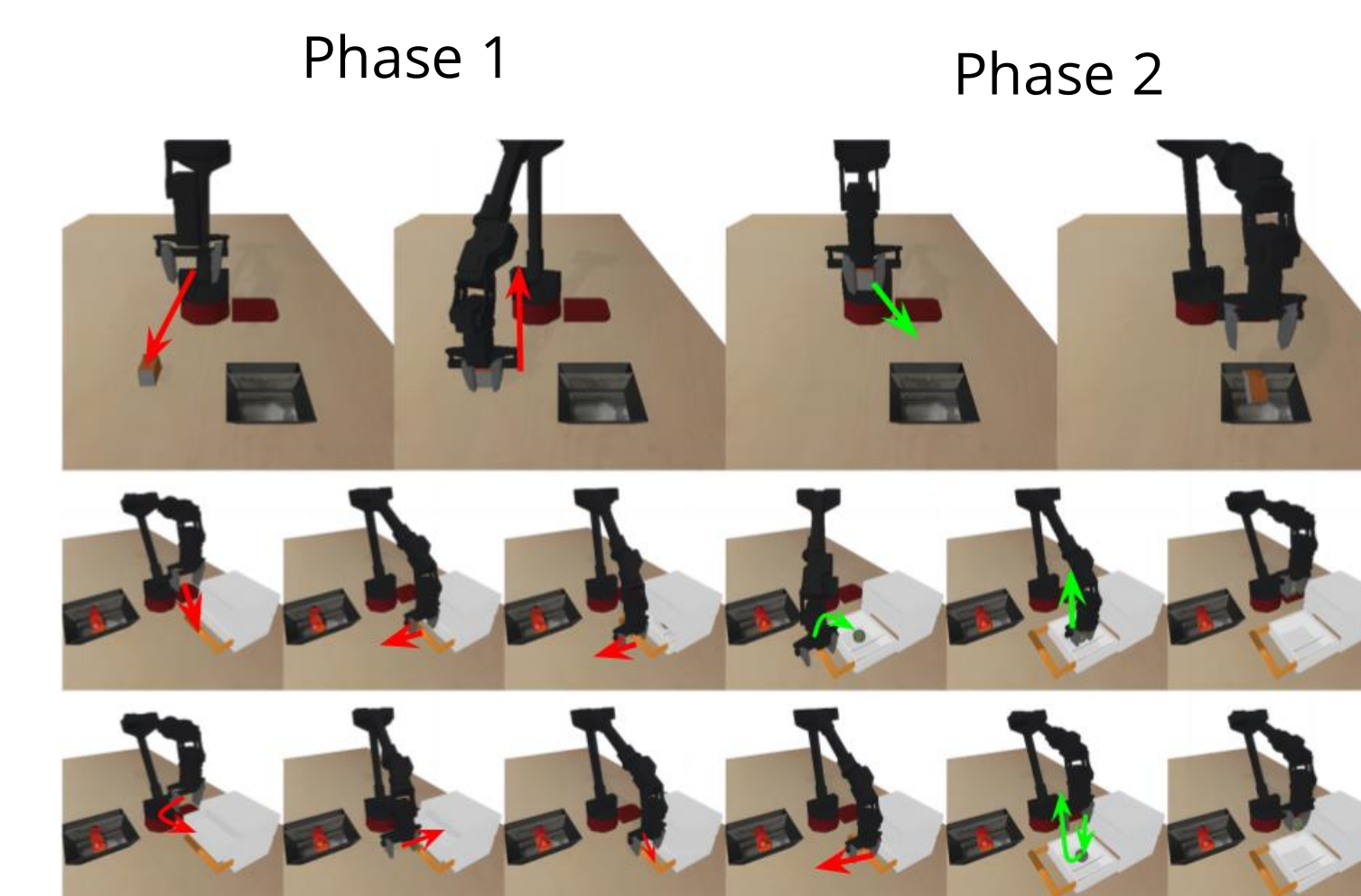
- > Optimal policy stitches together two suboptimal trajectories in the dataset.



Dataset	MBRCSL (ours)	COMBO	MOPO	CQL	DT	%BC
Pointmaze	91.5±7.1	56.6±50.1	77.9±41.9	34.8±24.9	57.2±4.1	54.0±9.2

Simulated Robotic Tasks

- > Each task consists of two phases of actions.
- > Dataset only contains trajectory from each phase, but not both.



Task	MBRCSL (ours)	CQL	COMBO	DT	BC	MBCQL
PickPlace	0.40±0.16	0.22±0.35	0±0	0±0	0.07±0.03	0.08±0.05
ClosedDrawer	0.51±0.12	0.11±0.08	0±0	0±0	0.38±0.02	0±0
BlockedDrawer	0.68±0.09	0.34±0.23	0±0	0±0	0.61±0.02	0±0

TL; DR

- > Off-policy RL via dynamic programming can diverge due to Bellman incompleteness.
- > Return-condition supervised learning cannot recover optimal behavior by stitching suboptimal trajectories.
- > We propose MBRCSL, augmenting the dataset with model-based rollouts of the behavior policy to enable trajectory stitching for RCSL.