RvS: What is Essential for Offline RL via Supervised Learning?

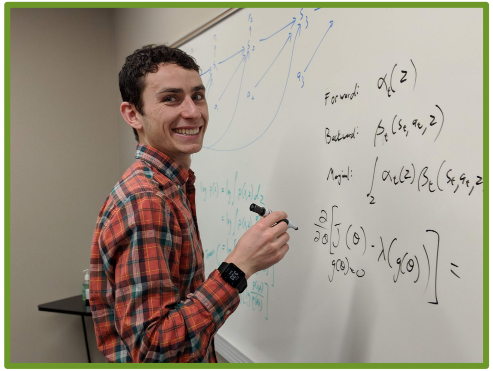
Scott Emmons

July 19, 2023

Center for Human-Compatible Artificial Intelligence

BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Acknowledgments



Ben Eysenbach

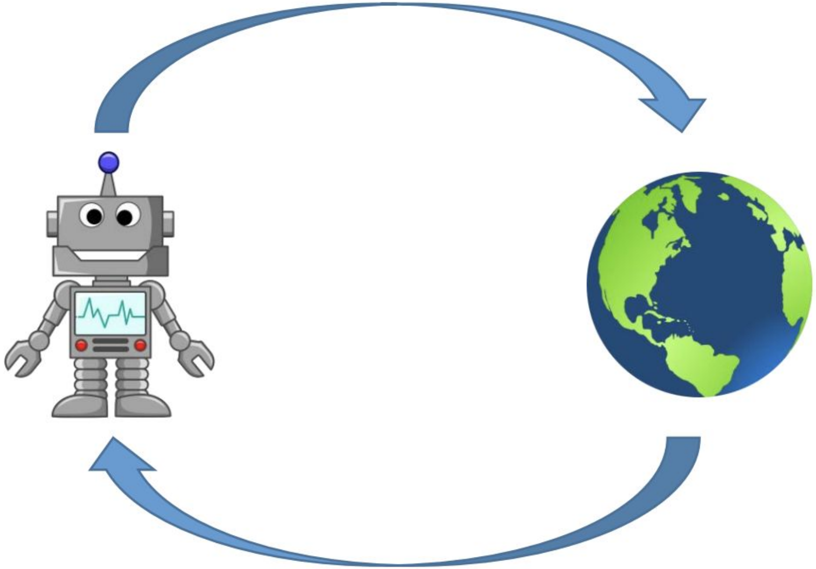
Ilya Kostrikov

Sergey Levine

DOE CSGF

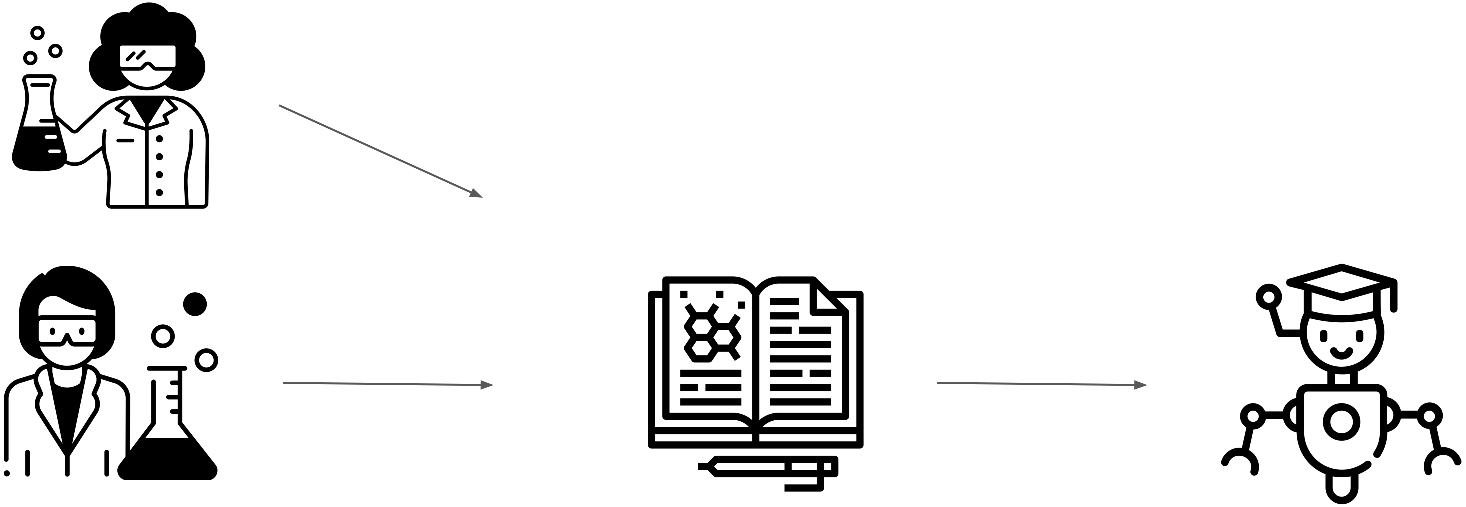
Reinforcement Learning

decisions (actions)



observations (states)
rewards

Offline Reinforcement Learning

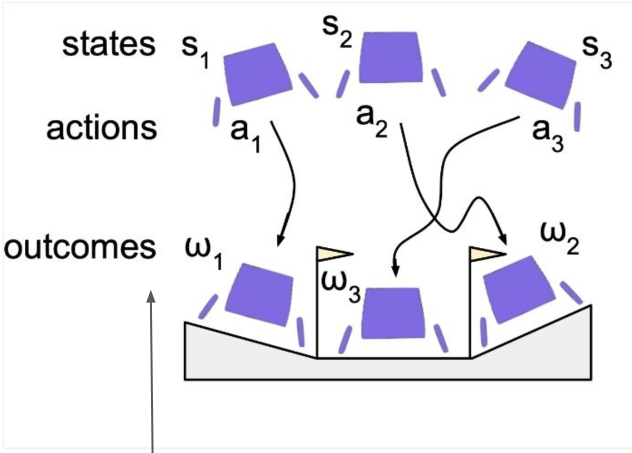


states

actions

rewards

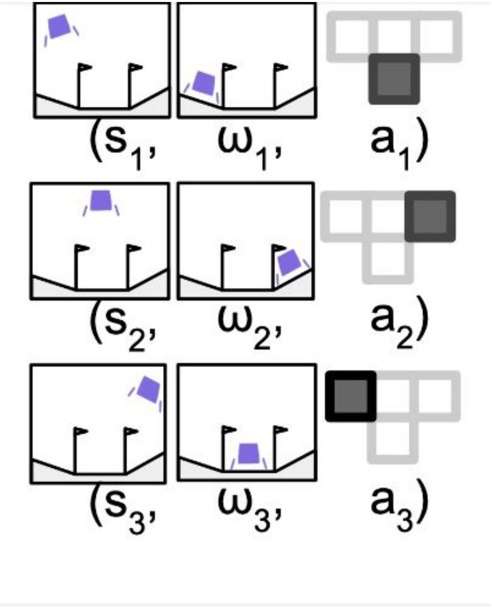
(Offline) RL via Supervised Learning



replay buffer

goal state (RvS-G)
reward-to-go (RvS-R)
language
etc.

hindsight relabeling



training dataset

$\pi(s, \omega) = a$

conditional policy

[Schmidhuber *et al.*, 2019; Kumar *et al.*, 2019; Ghosh *et al.*, 2021; Chen *et al.*, 2021]

Potential Benefits of Supervised Learning

More stable than RL

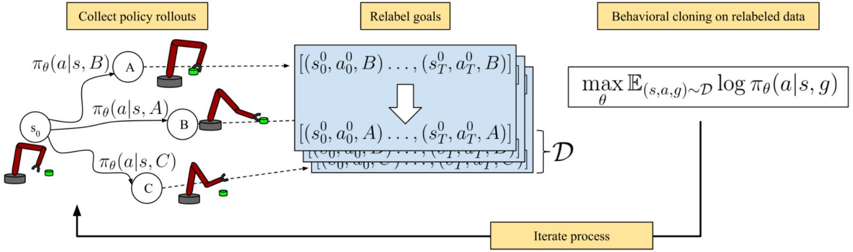
(Comparatively) easy to debug and validate

Success learning from large, precollected datasets

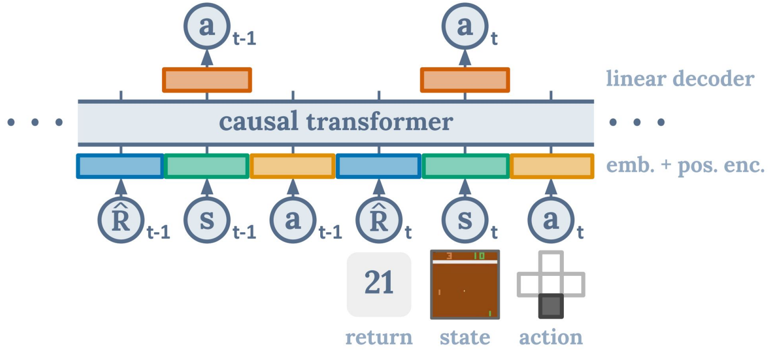
What Ingredients are Important? (Prior Work)

Reweight training data (RCP: Kumar *et al.*, 2019)

Iterative, online data collection (GCSL: Ghosh *et al.*, 2021)


Collect policy rollouts
Relabel goals
Behavioral cloning on relabeled data
$\pi_\theta(a|s, B)$
$\pi_\theta(a|s, A)$
$\pi_\theta(a|s, C)$
s_0
A
B
C
$[(s_0^0, a_0^0, B) \ldots, (s_T^0, a_T^0, B)]$
$[(s_0^0, a_0^0, A) \ldots, (s_T^0, a_T^0, A)]$
$\mathcal{D}$
$\max_\theta \mathbb{E}_{(s,a,g)\sim\mathcal{D}} \log \pi_\theta(a|s, g)$
Iterate process

Decision Transformer (DT: Chen *et al.*, 2021)



Key Questions

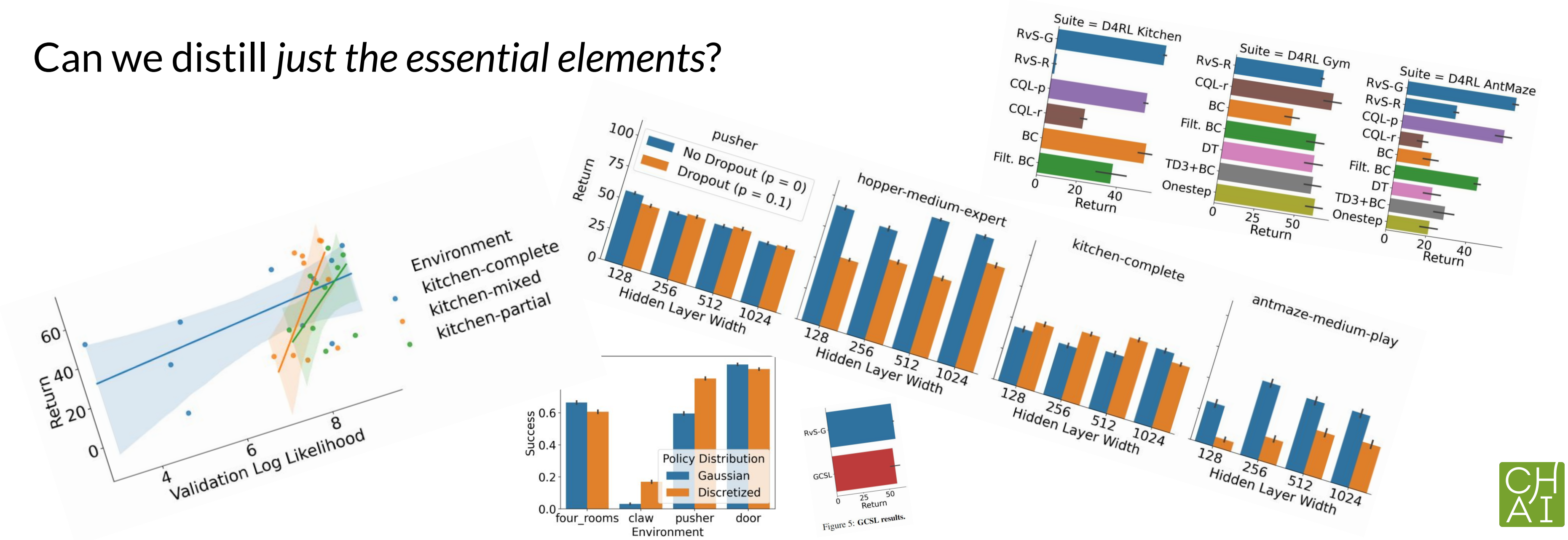
1. Which design decisions are critical for RL via supervised learning?
2. How well does it actually work?
3. What should we condition on? Does it matter?

Our Methodology

Experiments across 4 suites, 26 environments, and 8 algorithms

Vary model architecture, capacity, regularization, and conditioning space

Can we distill *just the essential elements*?



High-Performance Computing

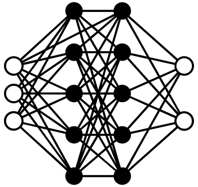
Experiments across 4 suites, 26 environments, and 8 algorithms

- 5 random seeds
- various policy architectures and distributions

Use Savio, the Berkeley Research Cluster!

- 470 nodes and 11,620 processor cores
- Nearly 450 peak teraFLOPS

CPUs

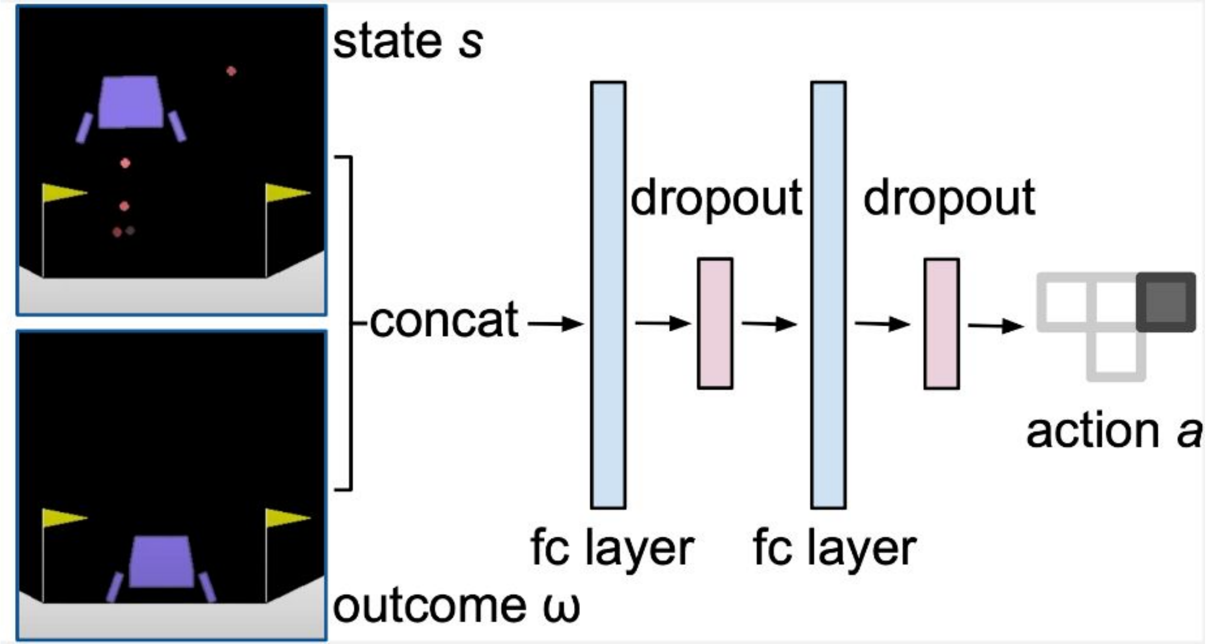


GPUs

BERKELEY LAB

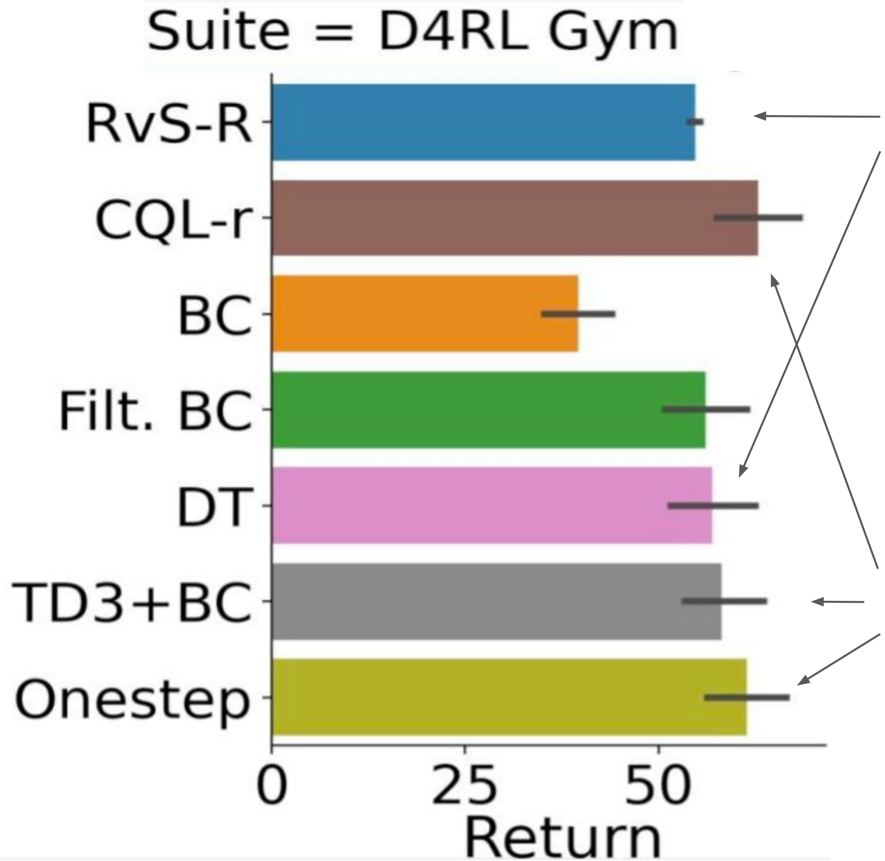
CHAI

Our Neural Network Architecture

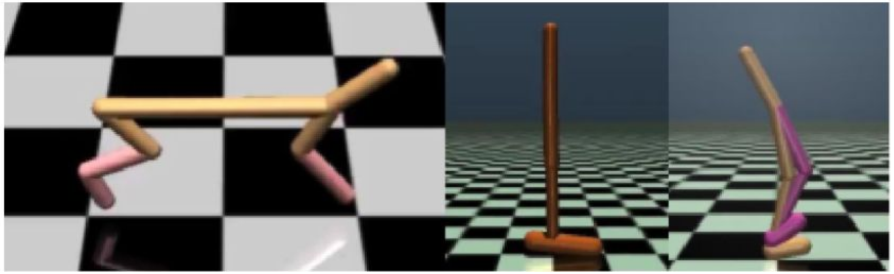


large capacity! (width 1024)
env-specific dropout

Overall Performance

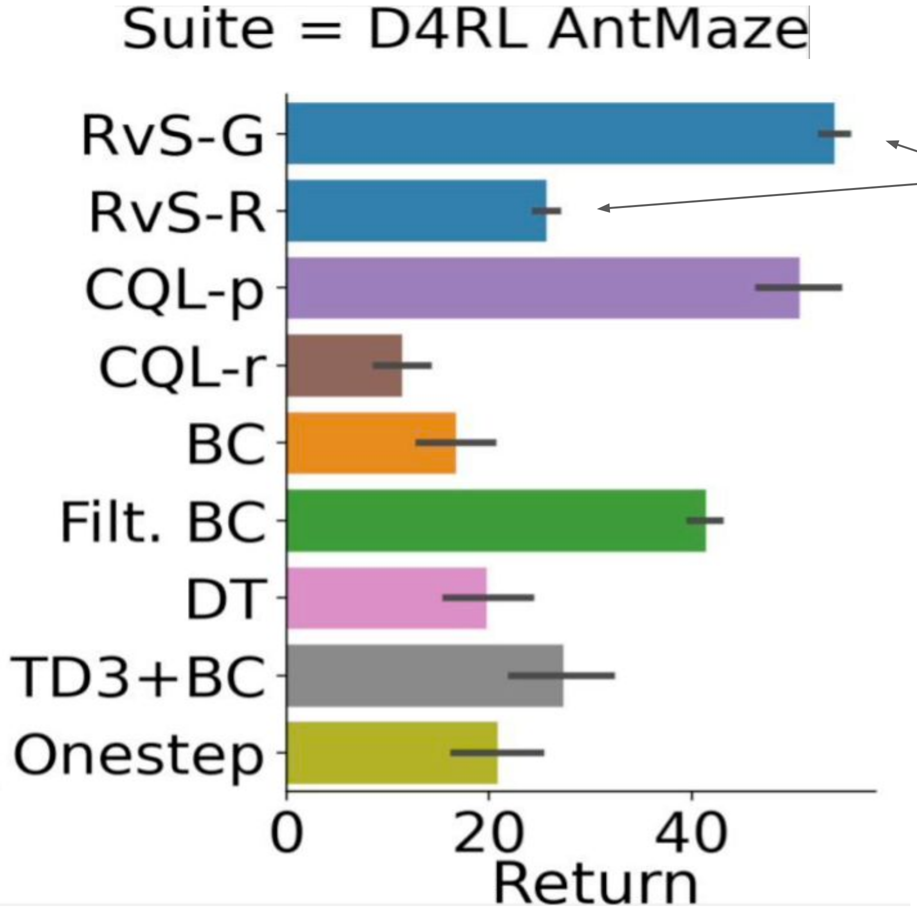


MLP matches performance of Transformer

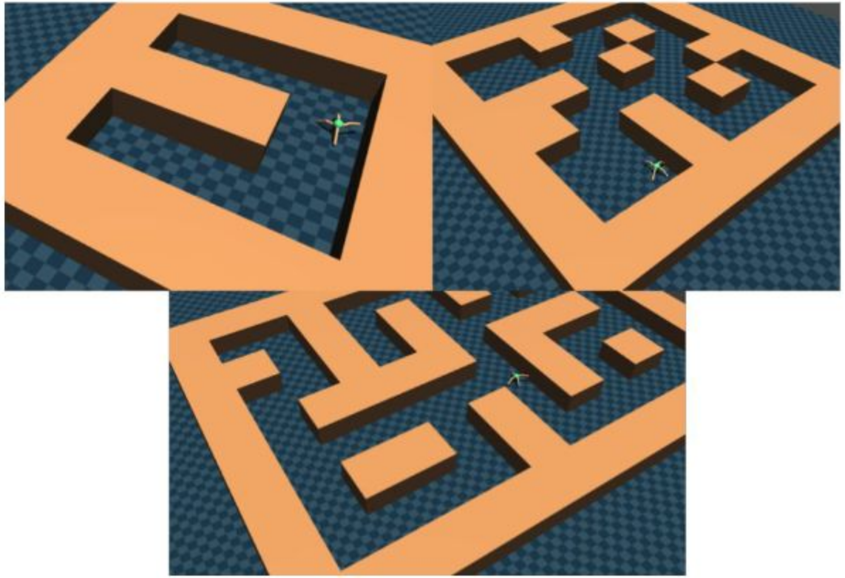


TD learning beats RvS on random data

Overall Performance



(x, y) coordinates >> reward-to-go!



Takeaways

You can do offline RL via pure supervised learning!

without reweighting data or Transformers

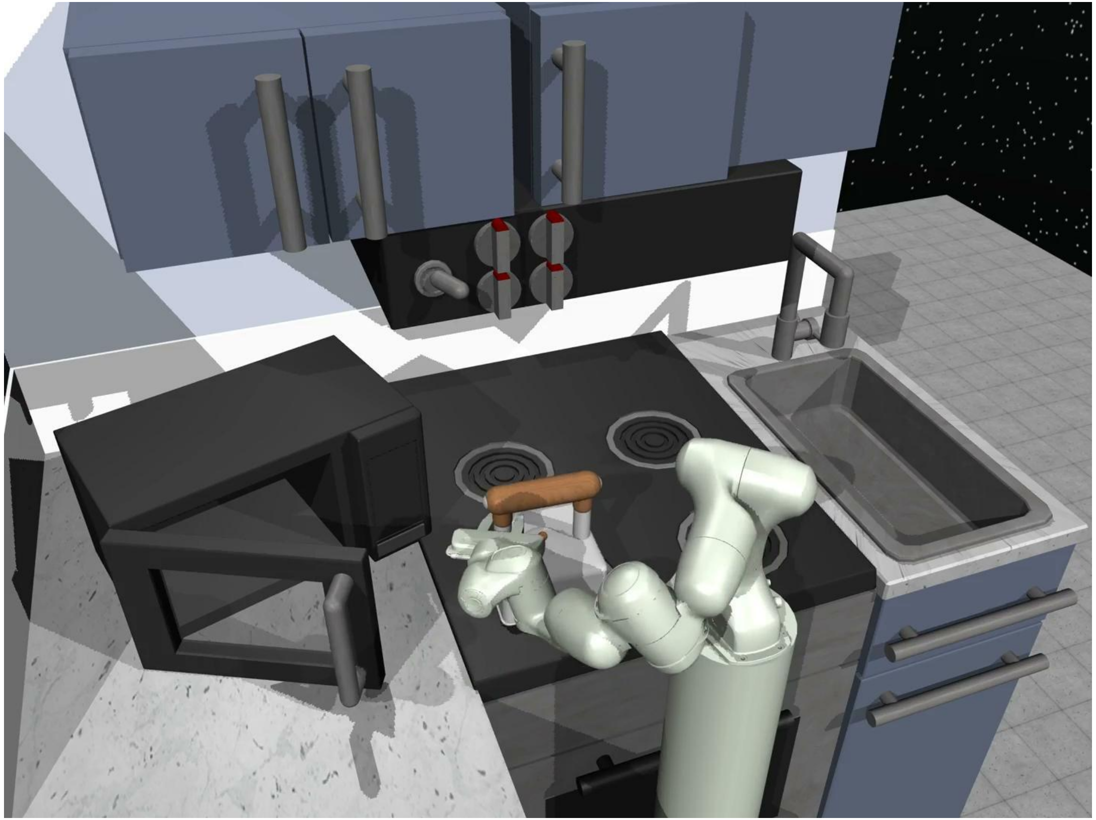
and achieve competitive results

across a wide variety of tasks

Model capacity, regularization, and the conditioning variable are key

Can we automate the choice of the conditioning variable?

RvS in D4RL Kitchen



(3x speed)