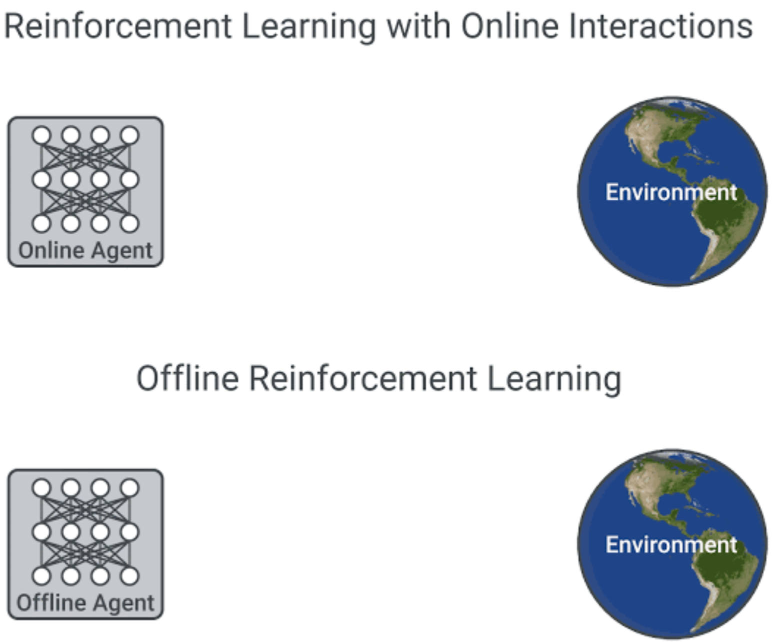
Versatile Offline Imitation from Observations and Examples via Regularized State Occupancy Matching

Jason Yecheng Ma, Andrew Shen, Dinesh Jayaraman, Osbert Bastani
University of Pennsylvania




Offline Imitation Learning from Observations



- (Few) expert observations: $\mathcal{D}^E = \{(s_0, ..., s_T)\}$
- Offline non-expert dataset: $\mathcal{D}^O = \{(s, a, s')\}$
- Objective: $\mathrm{D}_{\mathrm{KL}}(d^\pi(s) \| d^E(s))$
- Offline IL: $\pi = \mathcal{A}(\mathcal{D}^E, \mathcal{D}^O)$

How can we leverage small number of expert observations and large amount of unlabeled offline data to achieve offline imitation learning?

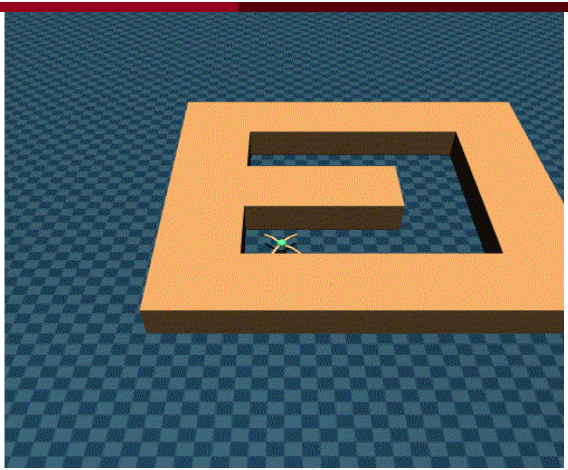
Penn Engineering

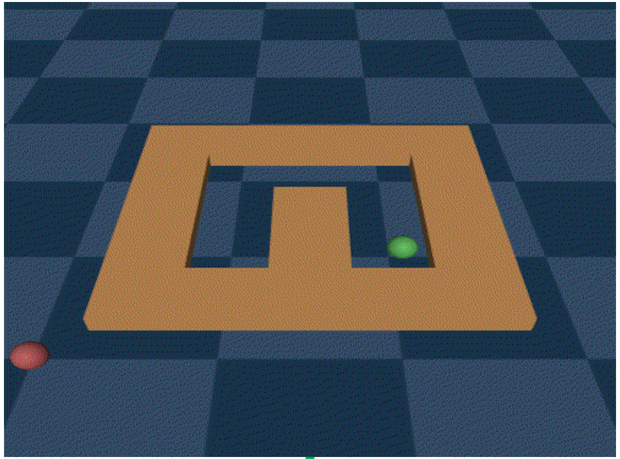
Objective: State Occupancy Matching

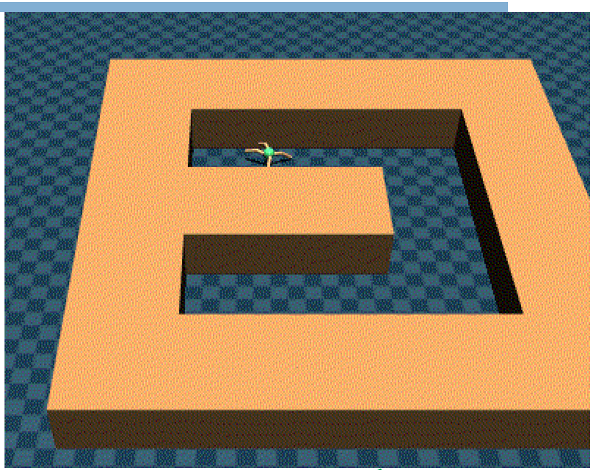
$$\min_{\pi} \mathrm{D}_{\mathrm{KL}}(d^{\pi}(s) \| d^{E}(s))$$

"distribution of task-relevant states the policy visits"

Versatility of State-Occupancy Matching



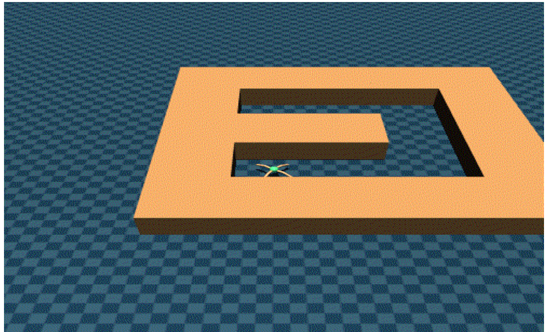




Observations

Mismatched Experts

Examples of Success

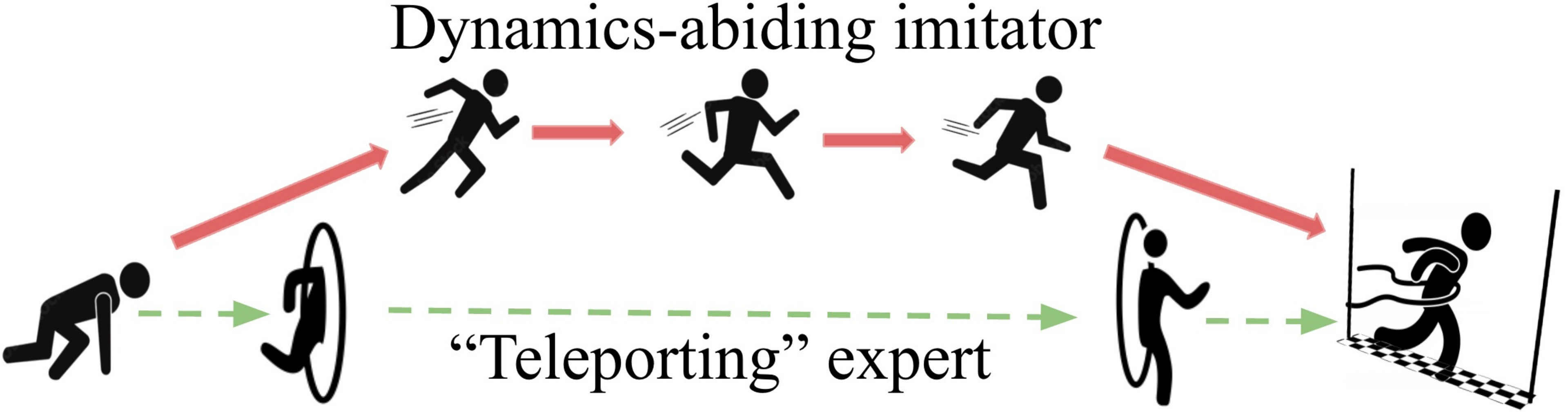


Penn Engineering

Imitation Learning from Examples

$$\min_{\pi} \mathrm{D}_{\mathrm{KL}}(d^{\pi}(s) \| d^{E}(s))$$

"state distribution of a teleporting expert"



Offline Imitation Learning from via State Matching

- Formulated as a state-matching problem:

$$\mathrm{D_{KL}}(d^\pi(s) \| d^E(s)) = \mathbb{E}_{s \sim \boxed{d^\pi}}\left[\log \frac{d^\pi(s)}{d^E(s)}\right]$$

- Key challenge is that this objective requires samples from the policy we are optimizing
- Difficult to do offline without access to the environment!

Regularized State-Occupancy Matching

Under some mild assumptions, for any f-divergence such that $\mathrm{D}_f \geq \mathrm{D}_{\mathrm{KL}}$,

$$\mathrm{D}_{\mathrm{KL}}(d^\pi(s) \| d^E(s)) \leq \mathbb{E}_{s \sim \boxed{d^\pi}}\left[\log\left(\frac{d^O(s)}{d^E(s)}\right)\right] + \mathrm{D}_f(d^\pi(s,a) \| d^O(s,a))$$

"reward function": encourages visiting expert states; learned using a discriminator!

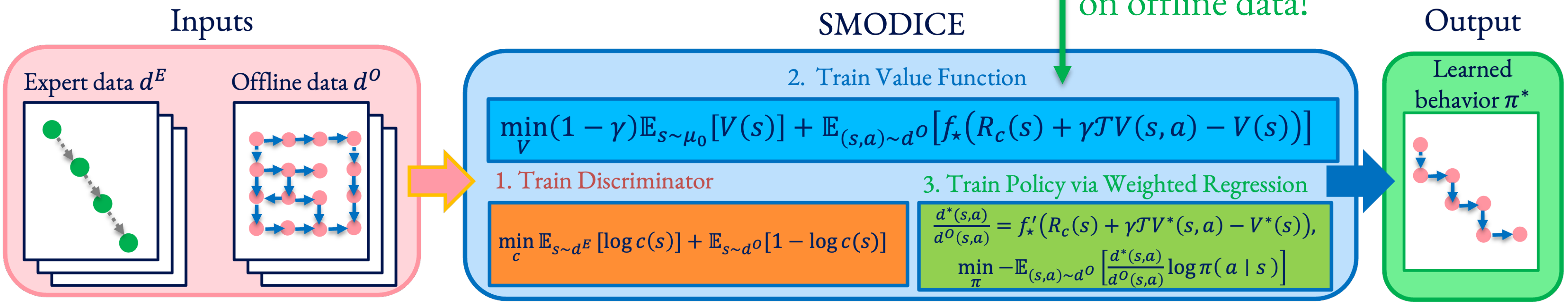
"constraint": encourages staying close to the offline dataset

Still requires on-policy samples!

State Matching Offline DIstribution Correction Estimation (SMODICE)

$$\mathrm{D}_{\mathrm{KL}}(d^\pi(s)\|d^E(s)) \leq \mathbb{E}_{s\sim d^\pi}\left[\log\left(\frac{d^O(s)}{d^E(s)}\right)\right] + \mathrm{D}_f(d^\pi(s,a)\|d^O(s,a))$$

Dual problem depends only on offline data!



☺ 3 Disjoint Optimization Steps

Penn Engineering

Experiments

Questions

1. Can SMODICE effectively learn from observations?
2. How robust is SMODICE to mismatched experts ?
3. Can SMODICE learn from examples of success outcomes?

1. Offline Imitation Learning from Observations

Outperforms state-of-art with privileged action information!

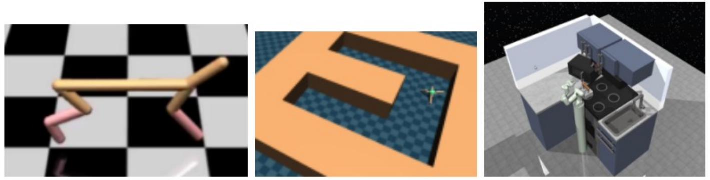
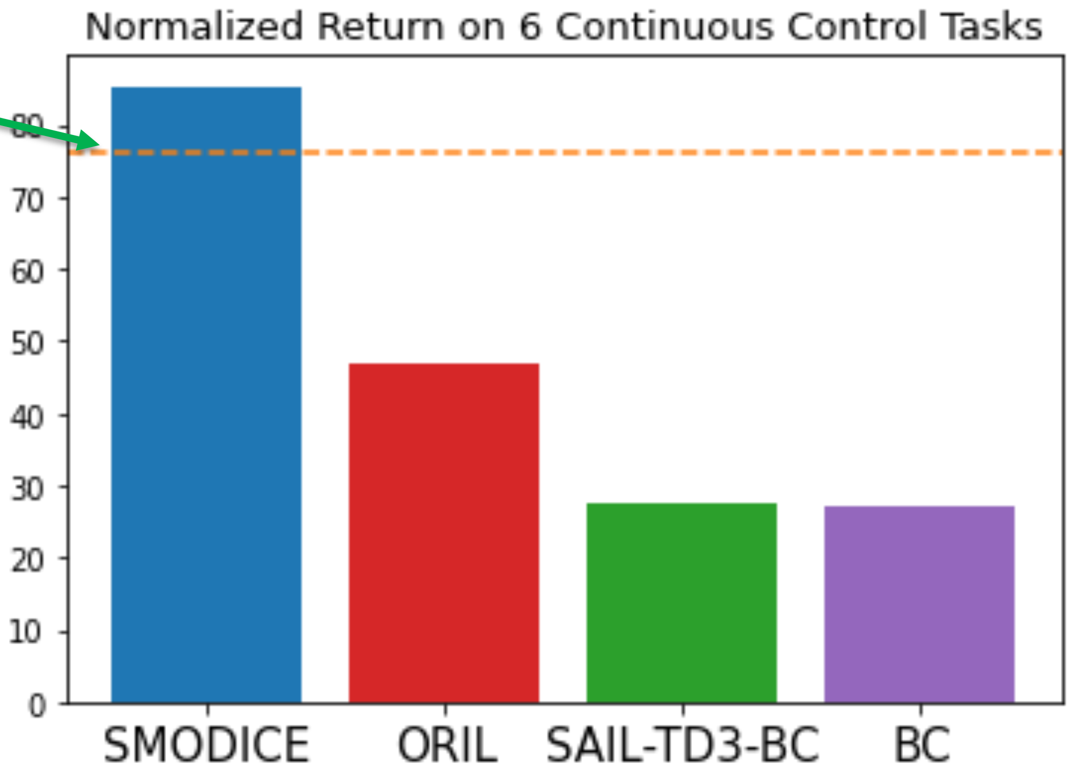
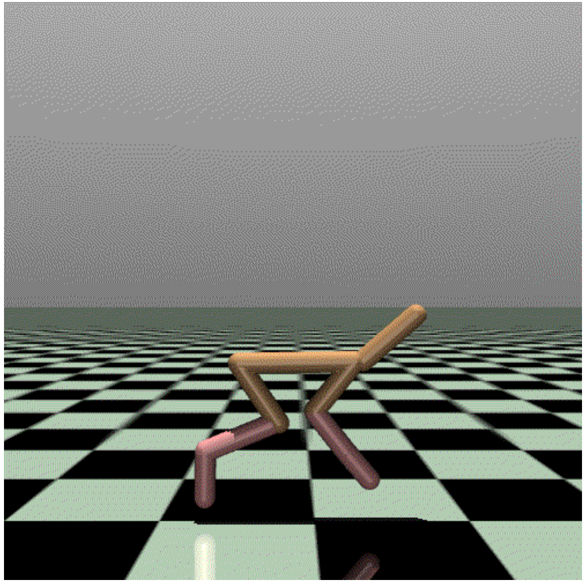
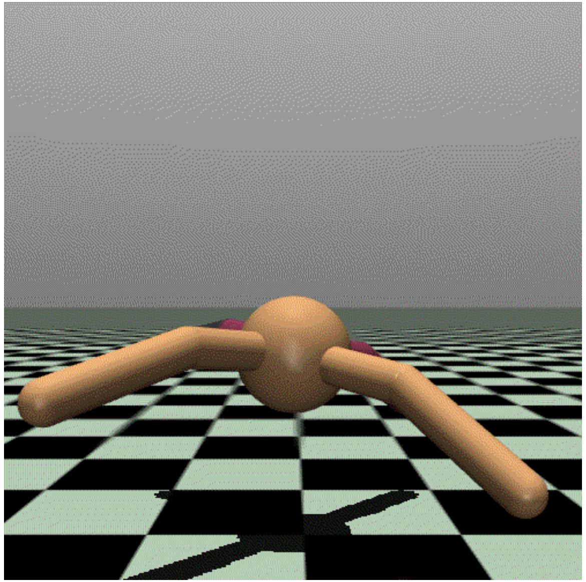
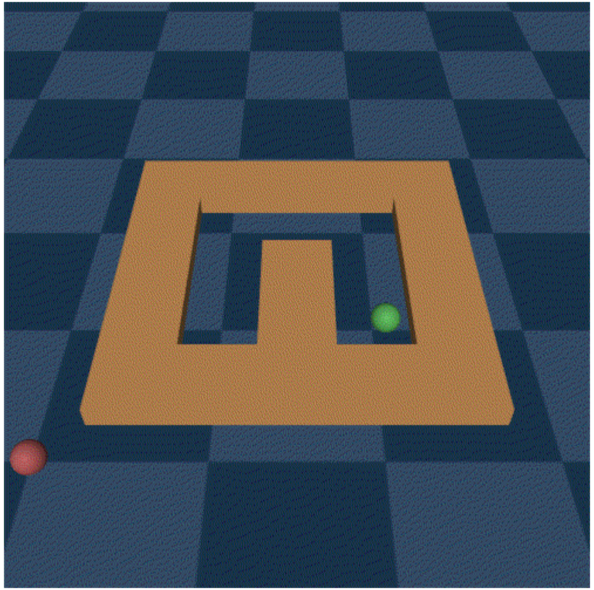


Figure 2. Illustrations of the evaluation environments.



2. Offline IL from Mismatched Experts

2. Offline IL from Mismatched Experts

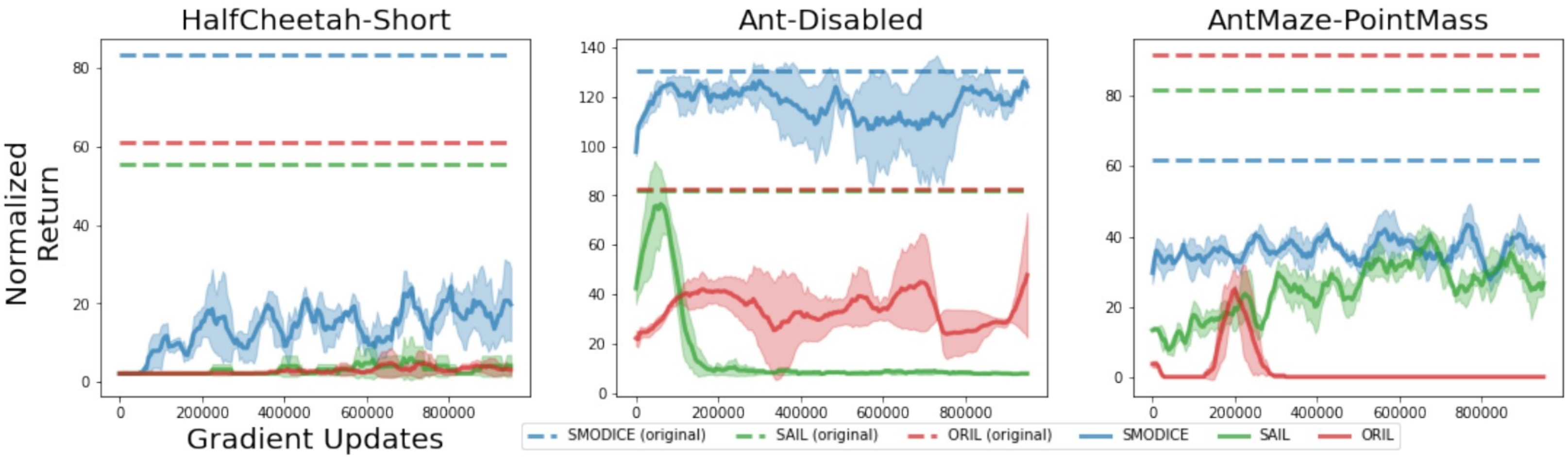


Figure 5. **Offline imitation learning from heterogeneous experts results.**

SMODICE is most robust to mismatched experts!

3. Offline IL from Examples

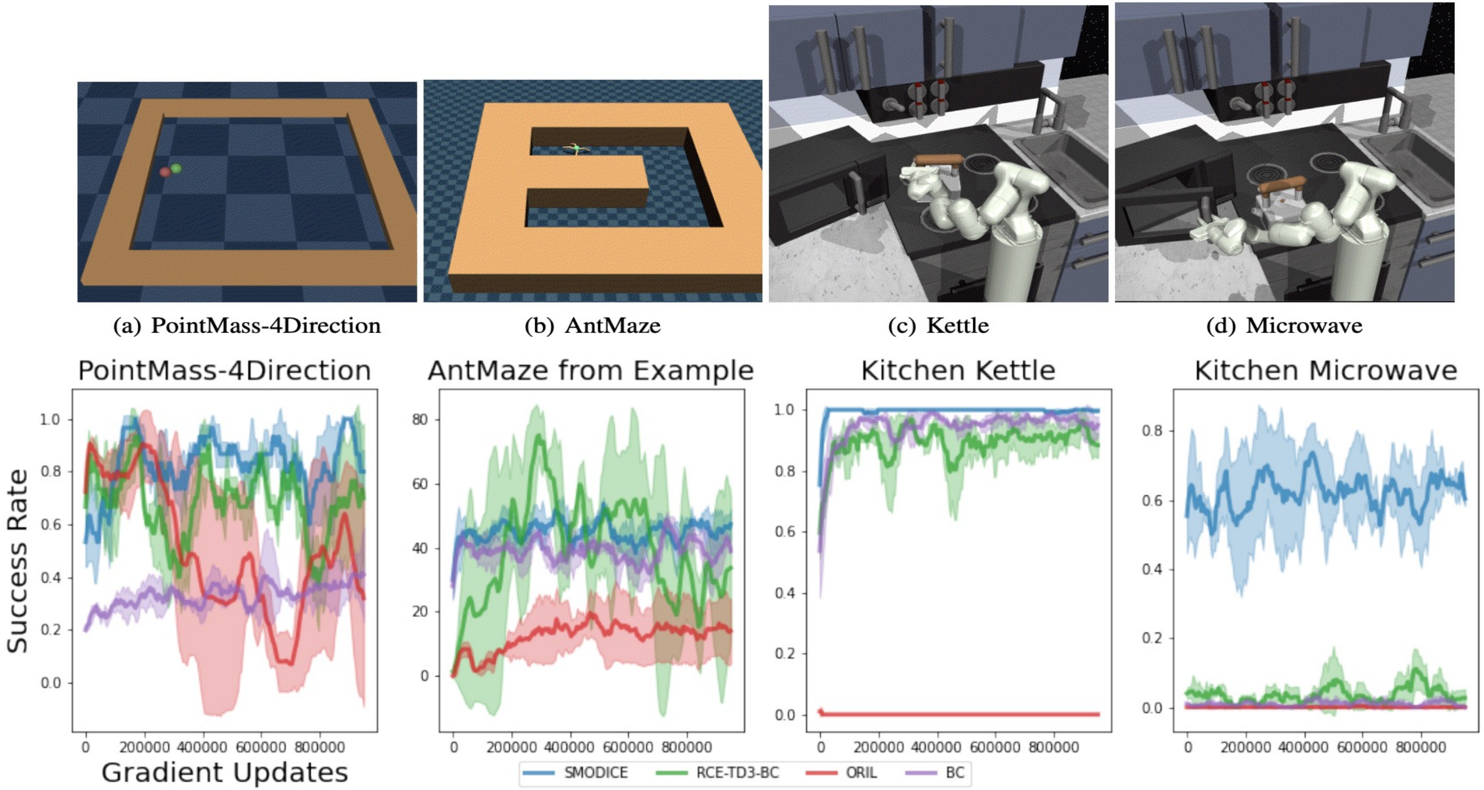
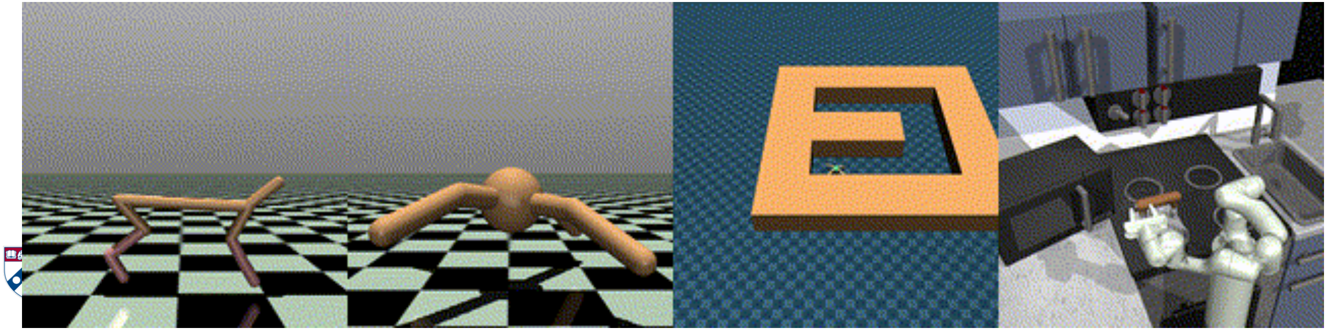


Figure 6. **Offline imitation learning from examples results.**

Conclusion

- State-matching as a framework for versatile offline IL
- SMODICE: A Regression-Based Offline IL Algorithm
- State-of-art results in all three settings without any hyperparameter tuning!



Project Website: https://sites.google.com/view/smodice/home

Jason Yecheng Ma, Andrew Shen, Dinesh Jayaraman, Osbert Bastani
University of Pennsylvania