IRIS: Implicit Reinforcement without Interaction at Scale for LEarning Control from Offline Robot Manipulation Data

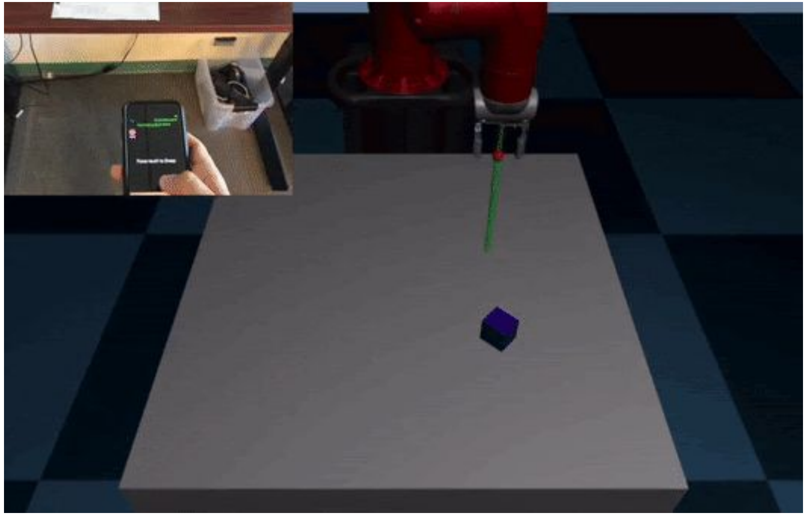
Liyan Chen

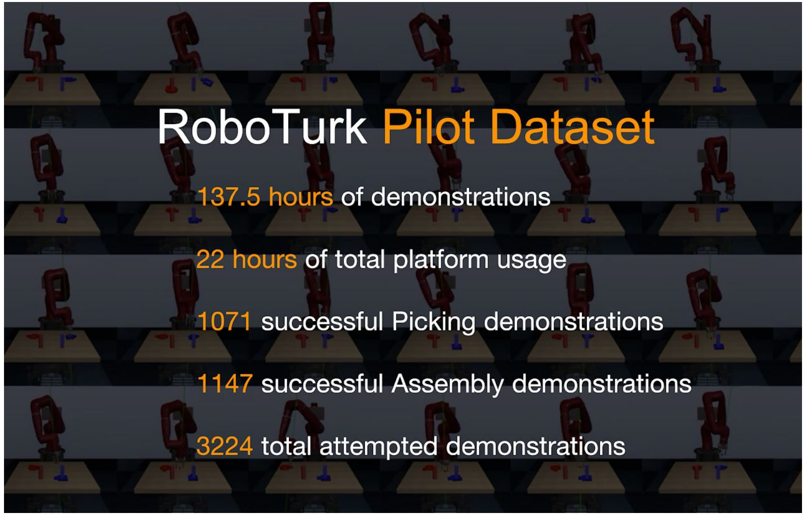
Oct 18, 2021

Motivation and Main Problem

- **Supervised learning has been successful through many areas except policy learning.**
- **Emerging large-scale dataset for task demonstration**

Existing Large-scale demonstration dataset





RoboTurk Pilot Dataset

137.5 hours of demonstrations

22 hours of total platform usage

1071 successful Picking demonstrations

1147 successful Assembly demonstrations

3224 total attempted demonstrations

Suboptimality

Fumbling the can

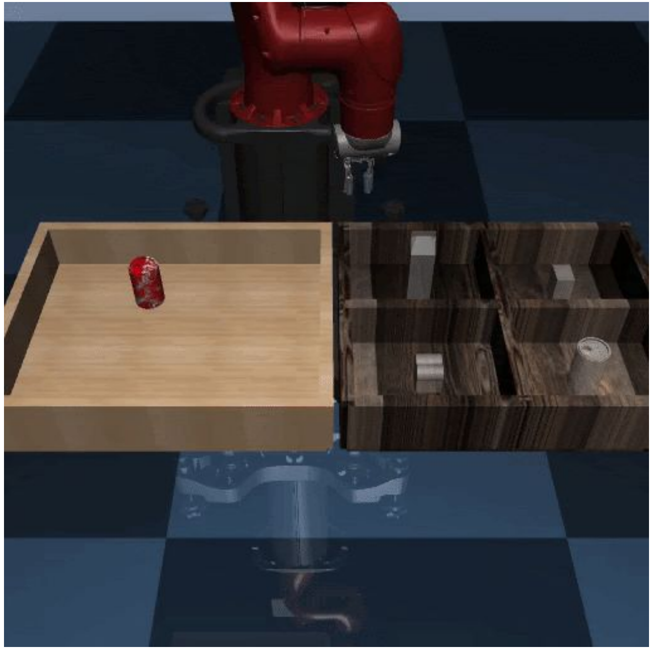
Suboptimality

Failed Sideways Grasp

Diversity

Straight Top-Down Grasp

Diversity



Tilt and Grab

Problem Setting

- MDP with absorbing goal states $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{T}, R, \gamma, \rho_0), \quad \mathcal{G} \subset \mathcal{S}$
- Task instantiation $s_0 \sim \rho_0(\cdot)$
- Maximize expected return $\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t, s_{t+1})].$
- Goal-Reaching Trajectories

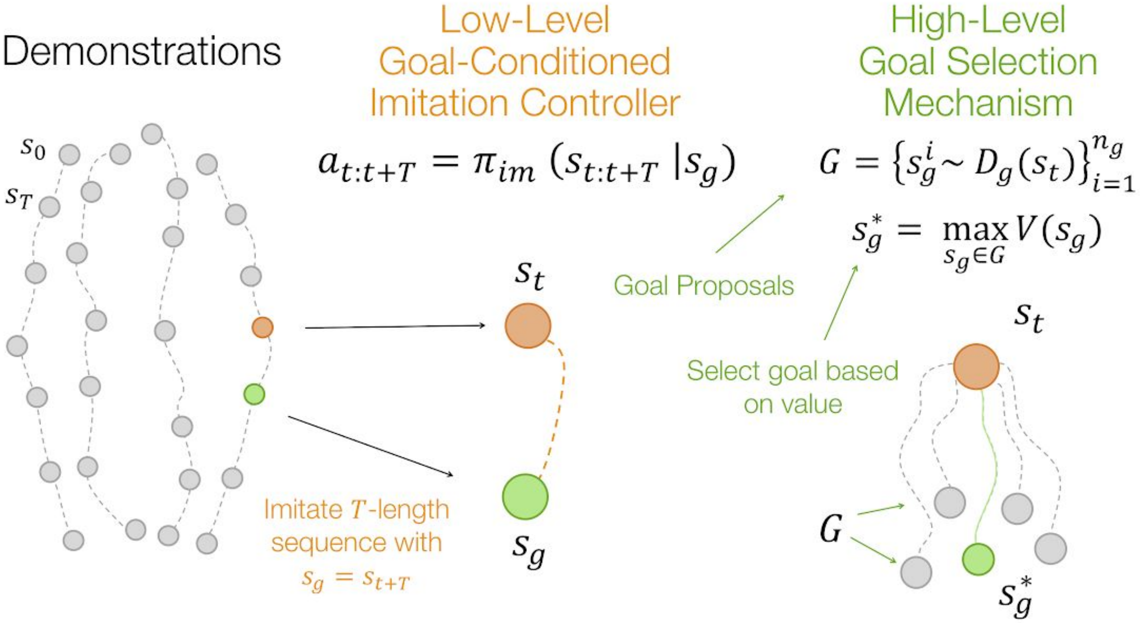
$$\tau = (s_0, a_0, r_0, s_1, \cdots, s_T), \text{s.t. } r_t = R(s_t, a_t, s_{t+1}), s_{t+1} \sim \mathcal{T}(\cdot | s_t, a_t)$$

Context / Related Work / Limitations of Prior Work

- **Imitation Learning**
  - **pros: reduce exploration cost**
  - **cons: task-specific, small scale**
- **Behavioral Cloning (BC)**
- **Batch-Constrained Q-Learning (BCQ)**

Proposed Approach: Goal learning + Goal proposals

IRIS: Implicit Reinforcement without Interaction at Scale



Demonstrations
s_0
s_T
Low-Level Goal-Conditioned Imitation Controller
Imitate T-length sequence with $s_g = s_{t+T}$
s_t
s_g
High-Level Goal Selection Mechanism
Goal Proposals
Select goal based on value
s_t
G
s_g^*

$$a_{t:t+T} = \pi_{im}\left(s_{t:t+T} \mid s_g\right)$$

$$G = \left\{s_g^i \sim D_g(s_t)\right\}_{i=1}^{n_g}$$

$$s_g^* = \max_{s_g \in G} V(s_g)$$

Proposed Approach: suboptimal demonstration

- Low-level controller learns short action sequences
- Goal generation chooses the most significant task goals

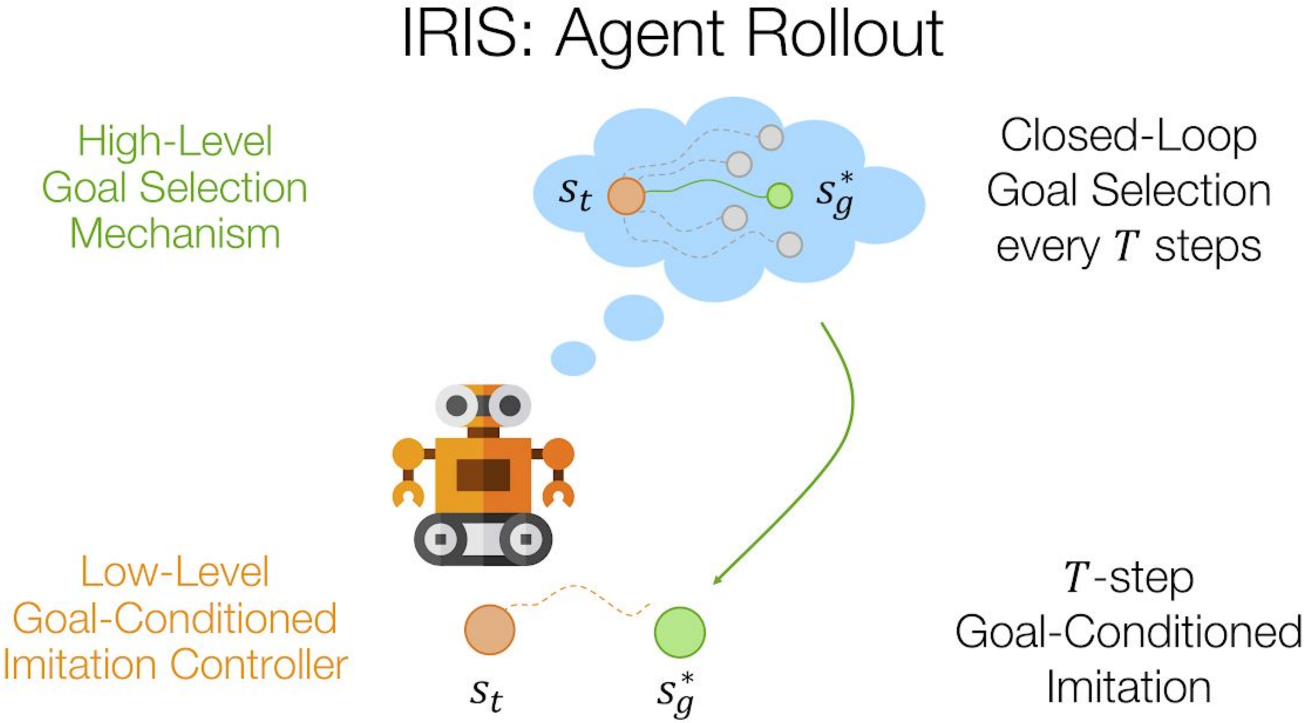
Proposed Approach: Diverse demonstration

- Low-level controller learns future goal observations at small temporal scale and produces unimodal action sequences as a result
- Goal generation proposes reachable goals from the current state, which induces diversity
- Decouples the problem into unimodal sequence learning and trajectory rollouts, achieving selective imitation.

Proposed Approach: Off-policy

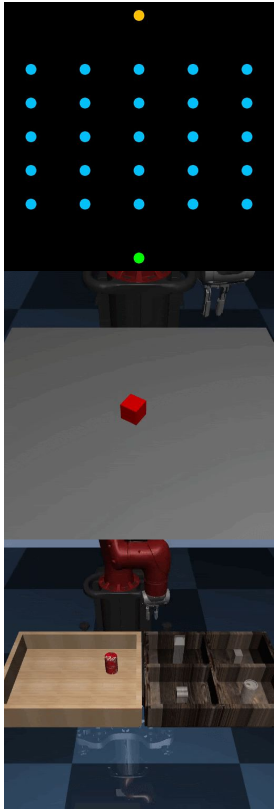
- Learning only in-distribution data
    - Goal generation proposes according to training data observations
    - Goal controller imitates training data sequences
- Goal selection avoids extrapolation error within the value learning part
    - Q-network is only queried on state-action pairs within the distribution

Proposed Approach: Inference



Experimental Setup

- Graph Reach
  - 2D Navigation, 5x5 grid
  - sampled random paths, demonstration of playing along the path
- Robosuite List
  - grasp and lift
  - demonstration of different approaches
- RoboTurk Can Pick and Place
  - pick and place
  - 225 fastest demonstrations, significant suboptimality and diversity



Experimental Results: Piecewise policy attains global integrity

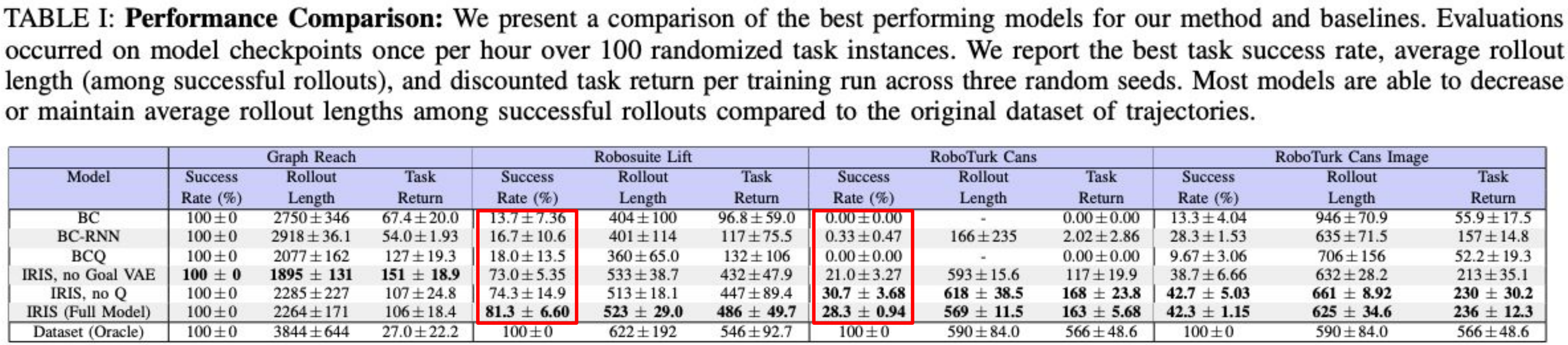


TABLE I: **Performance Comparison:** We present a comparison of the best performing models for our method and baselines. Evaluations occurred on model checkpoints once per hour over 100 randomized task instances. We report the best task success rate, average rollout length (among successful rollouts), and discounted task return per training run across three random seeds. Most models are able to decrease or maintain average rollout lengths among successful rollouts compared to the original dataset of trajectories.

	Graph Reach			Robosuite Lift			RoboTurk Cans			RoboTurk Cans Image		
Model	Success Rate (%)	Rollout Length	Task Return	Success Rate (%)	Rollout Length	Task Return	Success Rate (%)	Rollout Length	Task Return	Success Rate (%)	Rollout Length	Task Return
BC	100±0	2750±346	67.4±20.0	13.7±7.36	404±100	96.8±59.0	0.00±0.00	-	0.00±0.00	13.3±4.04	946±70.9	55.9±17.5
BC-RNN	100±0	2918±36.1	54.0±1.93	16.7±10.6	401±114	117±75.5	0.33±0.47	166±235	2.02±2.86	28.3±1.53	635±71.5	157±14.8
BCQ	100±0	2077±162	127±19.3	18.0±13.5	360±65.0	132±106	0.00±0.00	-	0.00±0.00	9.67±3.06	706±156	52.2±19.3
IRIS, no Goal VAE	**100 ± 0**	**1895 ± 131**	**151 ± 18.9**	73.0±5.35	533±38.7	432±47.9	21.0±3.27	593±15.6	117±19.9	38.7±6.66	632±28.2	213±35.1
IRIS, no Q	100±0	2285±227	107±24.8	74.3±14.9	513±18.1	447±89.4	**30.7 ± 3.68**	**618 ± 38.5**	**168 ± 23.8**	**42.7 ± 5.03**	**661 ± 8.92**	**230 ± 30.2**
IRIS (Full Model)	100±0	2264±171	106±18.4	**81.3 ± 6.60**	**523 ± 29.0**	**486 ± 49.7**	**28.3 ± 0.94**	**569 ± 11.5**	**163 ± 5.68**	**42.3 ± 1.15**	**625 ± 34.6**	**236 ± 12.3**
Dataset (Oracle)	100±0	3844±644	27.0±22.2	100±0	622±192	546±92.7	100±0	590±84.0	566±48.6	100±0	590±84.0	566±48.6

Experimental Results: Piecewise policy attains global integrity

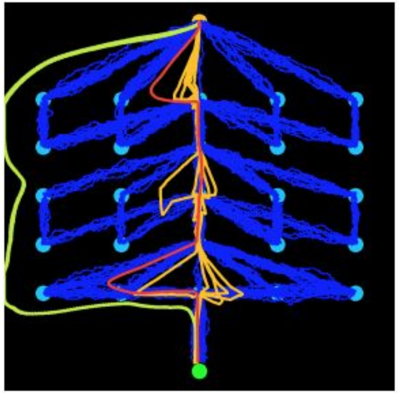


Fig. 4: **Qualitative Evaluation:** We visualize 5 trajectories taken by the best performing policies for the BC (red), BCQ (green), and IRIS (orange) models on the Graph Reach environment. A set of 50 trajectories from the dataset (blue) is also shown. Our model is both able to faithfully reconstruct demonstrated trajectories and leverage them to reach the goal quickly. By contrast, BCQ extrapolates an entirely new trajectory, while BC converges to a particular mode in the dataset that is slow to reach the goal. Unlike the other models, ours also exhibits variation in policy rollouts.

Empirical Analysis: Ablate Goal Proposal

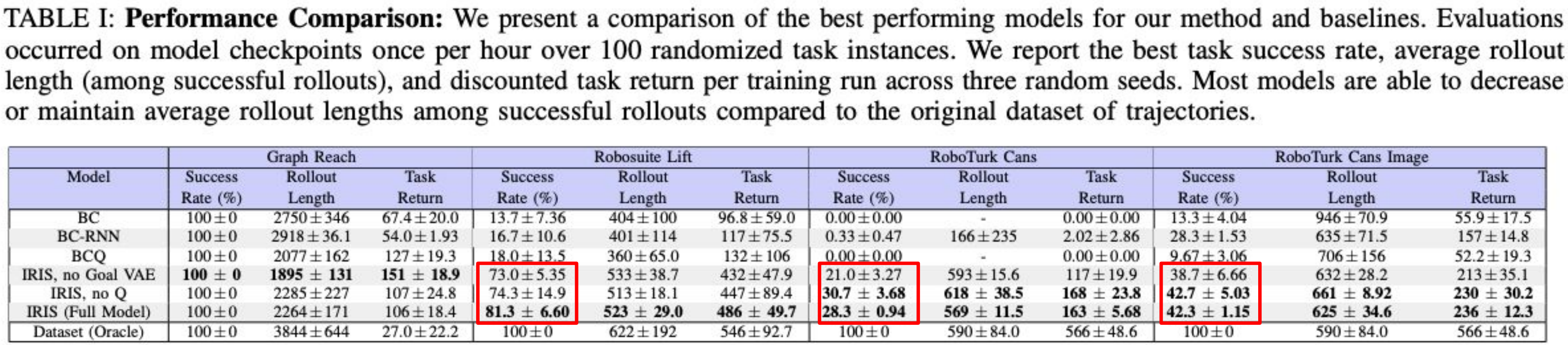


TABLE I: **Performance Comparison:** We present a comparison of the best performing models for our method and baselines. Evaluations occurred on model checkpoints once per hour over 100 randomized task instances. We report the best task success rate, average rollout length (among successful rollouts), and discounted task return per training run across three random seeds. Most models are able to decrease or maintain average rollout lengths among successful rollouts compared to the original dataset of trajectories.

	Graph Reach			Robosuite Lift			RoboTurk Cans			RoboTurk Cans Image		
Model	Success Rate (%)	Rollout Length	Task Return	Success Rate (%)	Rollout Length	Task Return	Success Rate (%)	Rollout Length	Task Return	Success Rate (%)	Rollout Length	Task Return
BC	100 ± 0	2750 ± 346	67.4 ± 20.0	13.7 ± 7.36	404 ± 100	96.8 ± 59.0	0.00 ± 0.00	-	0.00 ± 0.00	13.3 ± 4.04	946 ± 70.9	55.9 ± 17.5
BC-RNN	100 ± 0	2918 ± 36.1	54.0 ± 1.93	16.7 ± 10.6	401 ± 114	117 ± 75.5	0.33 ± 0.47	166 ± 235	2.02 ± 2.86	28.3 ± 1.53	635 ± 71.5	157 ± 14.8
BCQ	100 ± 0	2077 ± 162	127 ± 19.3	18.0 ± 13.5	360 ± 65.0	132 ± 106	0.00 ± 0.00	-	0.00 ± 0.00	9.67 ± 3.06	706 ± 156	52.2 ± 19.3
IRIS, no Goal VAE	**100 ± 0**	**1895 ± 131**	**151 ± 18.9**	73.0 ± 5.35	533 ± 38.7	432 ± 47.9	21.0 ± 3.27	593 ± 15.6	117 ± 19.9	38.7 ± 6.66	632 ± 28.2	213 ± 35.1
IRIS, no Q	100 ± 0	2285 ± 227	107 ± 24.8	74.3 ± 14.9	513 ± 18.1	447 ± 89.4	**30.7 ± 3.68**	**618 ± 38.5**	**168 ± 23.8**	**42.7 ± 5.03**	**661 ± 8.92**	**230 ± 30.2**
IRIS (Full Model)	100 ± 0	2264 ± 171	106 ± 18.4	**81.3 ± 6.60**	**523 ± 29.0**	**486 ± 49.7**	**28.3 ± 0.94**	**569 ± 11.5**	**163 ± 5.68**	**42.3 ± 1.15**	**625 ± 34.6**	**236 ± 12.3**
Dataset (Oracle)	100 ± 0	3844 ± 644	27.0 ± 22.2	100 ± 0	622 ± 192	546 ± 92.7	100 ± 0	590 ± 84.0	566 ± 48.6	100 ± 0	590 ± 84.0	566 ± 48.6

Empirical Analysis: Data efficiency

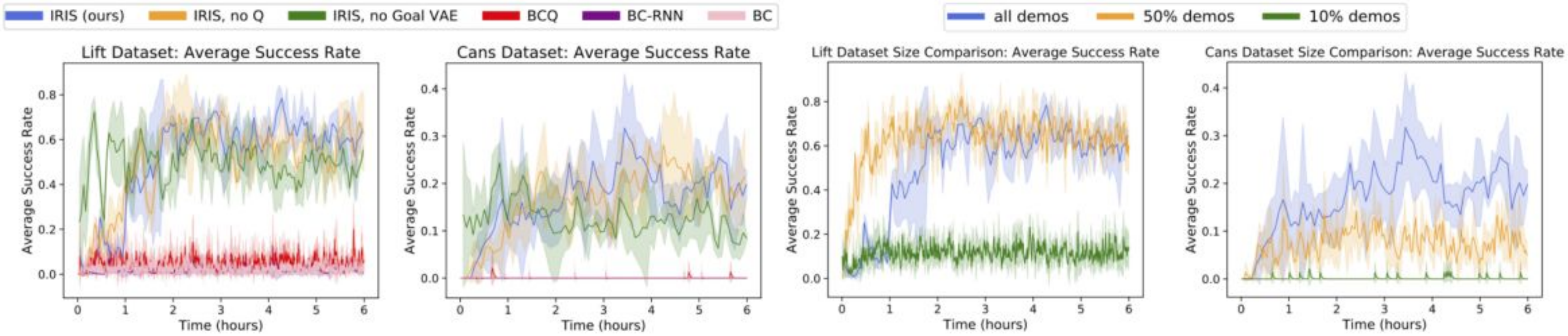
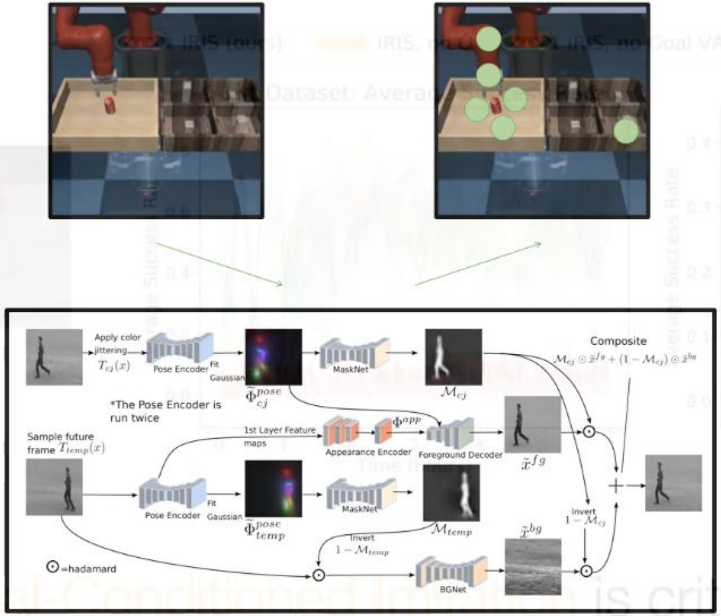


Fig. 3: **Manipulation Results:** We present a comparison of IRIS against several baselines on the Robosuite Lift and RoboTurk Cans datasets (left two plots). There is a stark contrast in performance between variants of IRIS and the baseline models, which suggests that *goal-conditioned imitation* is critical for good performance. We also perform a dataset size comparison (right two plots) to understand how the performance of IRIS is affected by different quantities of data.

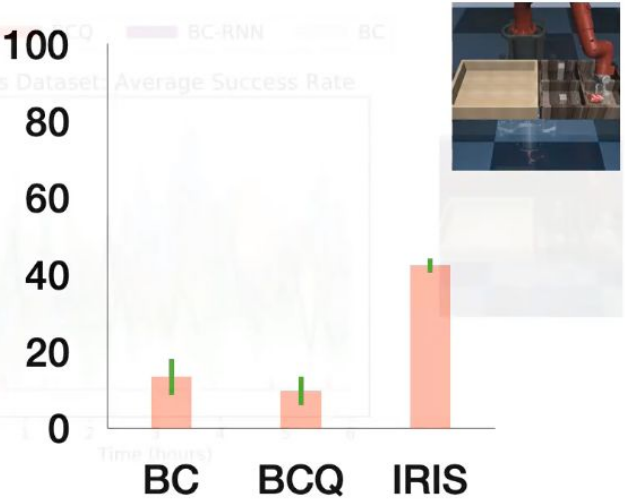
Empirical Analysis: Pixel Input

Unsupervised Representation Learning



Dundar et al. 2020

RoboTurk Cans Image



~40% Success Rate on Cans Task

Goal-Conditioned Imitation is critical for good performance

Limitations

- **Poor domain adaptation by construction**
- **Applicable only to fully observable systems (requires system states as supervision**
- **Reduced to hierarchical RL in some sense**

Future Work for Paper / Reading

- Domain Adaptation
- Generalize to larger-scale learning from weaker demonstrations, raw data in the wild, etc
- Extract stronger primitives (semantical/physical/hierarchical/etc) instead of vague goal proposals

Extended Readings

Fujimoto, Scott, David Meger, and Doina Precup. "**Off-policy deep reinforcement learning without exploration**." In *International Conference on Machine Learning*, pp. 2052-2062. PMLR, 2019.

Mandlekar, Ajay, Danfei Xu, Josiah Wong, Soroush Nasiriany, Chen Wang, Rohun Kulkarni, Li Fei-Fei, Silvio Savarese, Yuke Zhu, and Roberto Martín-Martín. "**What Matters in Learning from Offline Human Demonstrations for Robot Manipulation**." *arXiv preprint arXiv:2108.03298* (2021).

Summary

- Supervised learning for policy learning
- Suboptimality, diversity
- BC/BCQ suffers from poor extrapolation and separating suboptimality
- IRIS: Goal proposal + short trajectory policy
- IRIS attains global integrity, data-efficiency, rollout diversity
- Suffers from poor generalizability