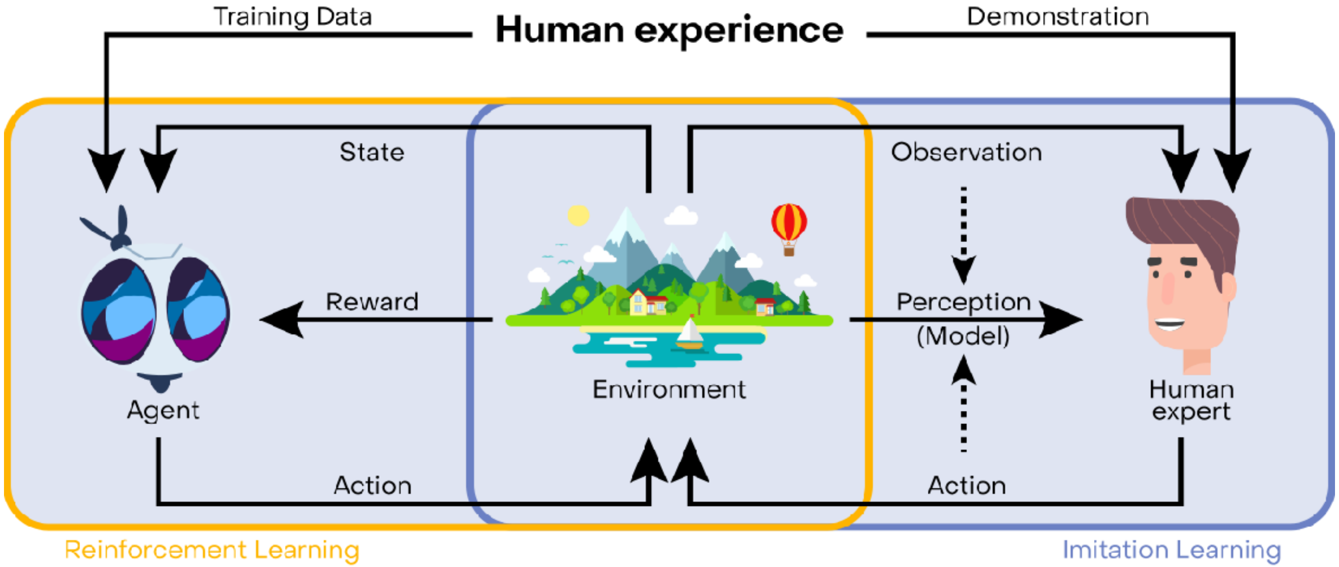
Robot Learning on the Job: Human-in-the-Loop Autonomy and Learning During Deployment

Presenter: Zhiyun Deng

10/10/2023

Motivation and Main Problem

Human feedback: interventions, preferences, rankings, scalar-valued feedback, and human gaze



Human-in-the-loop Learning [Source: Neda Navidi et al.]

Motivation and Main Problem

Scientific hypotheses: Policy learning benefits when human interventions inform:

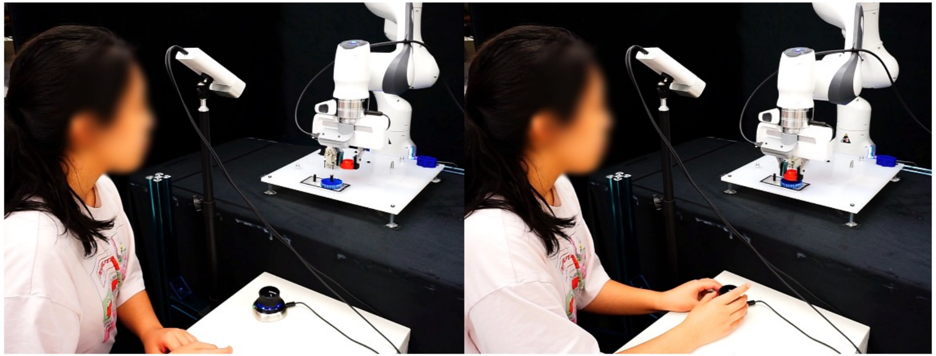
- **When** the human lacks trust in the robot
- **Where** the risk-sensitive task states are
- **How** to traverse these status

Challenges:

- How can we effectively and efficiently use the mixed-quality of data from human-robot collaborations for policy updates, especially when this data might be diverse and sub-optimal?
- How can we ensure the robot learns from positive behaviors (like human demonstrations) and reinforces them, while avoiding the replication of mistakes that could result in failures?

Problem Setting

Human-Robot Collaborative Manipulation System



Implicit Knowledge in Human-Robot Collaboration

- ❖ **When** the human lacks trust in the robot
- ❖ **Where** are the risk-sensitive task states
- ❖ **How** to traverse these states

Teleoperation Interface

(6-DoF SpaceMouse)

Operational Space

- ❖ **Position:** x-y-z
- ❖ **Orientation** yaw-pitch-roll
- ❖ **Gripper**: open-close command {1., -1.}

Problem Setting

Problem Formulation:

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P}, p_0, \gamma)$$

Intervention-based Learning Framework:

Binary indicator function of human interventions

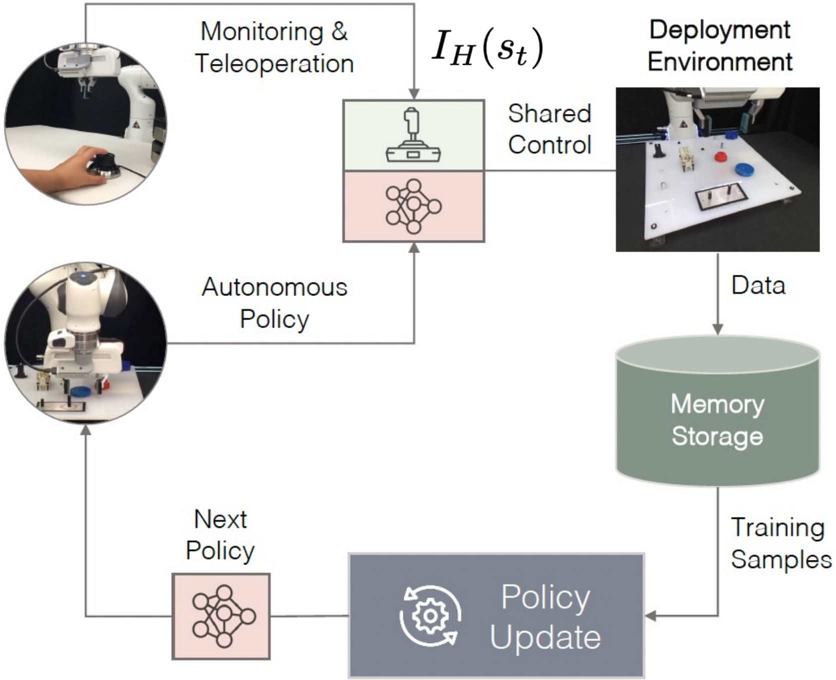
$$\pi(\cdot \mid s_t) = \underbrace{I_H(s_t)\pi_H(\cdot \mid s_t)}_{\text{Implicit human policy}} + \underbrace{(1 - I_H(s_t))\pi_R(\cdot \mid s_t)}_{\text{Robot policy}}$$

Learning Objective:

(obtain high-performance robot policy)

- Maximize $\mathbb{E}_{\pi_R}\left[\sum_{t=0}^{\infty} \gamma^t r\left(s_t, a_t, s_{t+1}\right)\right]$
- Minimize $\mathbb{E}_{\pi}\left[I_H(s_t)\right]$ (reduce human workload over time)

Human-in-the-loop Learning and Deployment Framework



Related Work

- **Human-in-the-loop Learning**: Human interventions have been incorporated in imitation learning or deep reinforcement learning; however, these method fail to incorporate human control feed back in deployment into the learning loop

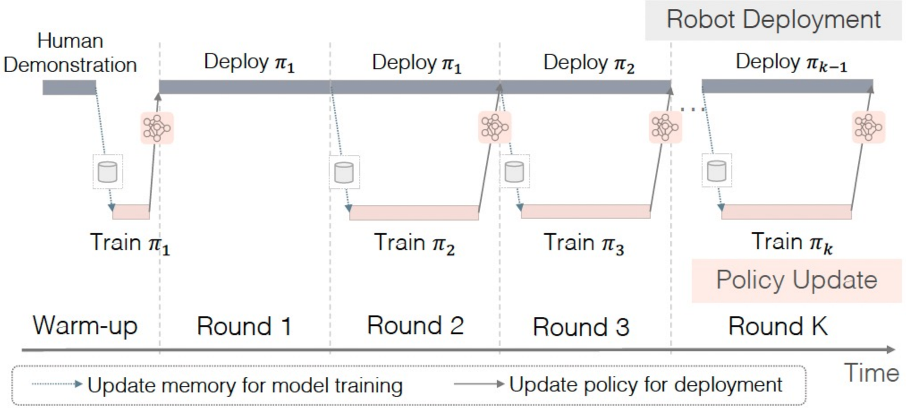
- **Shared Autonomy:** The existing literature focuses on efficient collaborative control from human intent prediction; however, they do not attempt to learn from human intervention feedback and there is no policy improvement

- **Learning from Offline Data:** Imitation learning and offline reinforcement learning can be used to learn from fixed robot datasets. The weighted behavior objective is used to learn the policy.

Proposed Approach: SIRIUS

Robot Deployment	The robot performs task executions with human monitoring
Policy Update	The system improves the policy with the deployment data for the next round of task execution

- Collect a small number of **human demonstrations**
  $\mathcal{D}^0 = \{\tau_j\} \quad \tau_j = \{s_t, a_t, r_t, c_t\} \quad c_t = \texttt{demo}$
- Train an **initial policy** using **BC** and deploy it
  π_1
- Collect a **new dataset of trajectories**
  $\mathcal{D}' = \{\tau_j\} \quad \tau_j = \{s_t, a_t, r_t, c_t\} \quad c_t = \begin{cases} \texttt{robot} \\ \texttt{intv} \end{cases}$
- Append this new data to the existing **memory buffer**
  $\mathcal{D}^1 \leftarrow \mathcal{D}^0 \cup \mathcal{D}'$ How to manage memory buffer?
- Train a **new policy** on this new dataset and deploy it
  π_2 How to learn from mixed-quality date?



Proposed Approach: Reweighting Scheme for BC

Classical BC | Weighted BC

BC Objective:

$$\theta^* = \arg\max_\theta \mathop{\mathbb{E}}_{(s,a)\sim\mathcal{D}} [\log \pi_\theta(a \mid s)]$$

$$= \arg\max_\theta \mathop{\mathbb{E}}_{P(c)} \mathop{\mathbb{E}}_{(s,a)\sim\mathcal{D}_c} [\log \pi_\theta(a \mid s)]$$

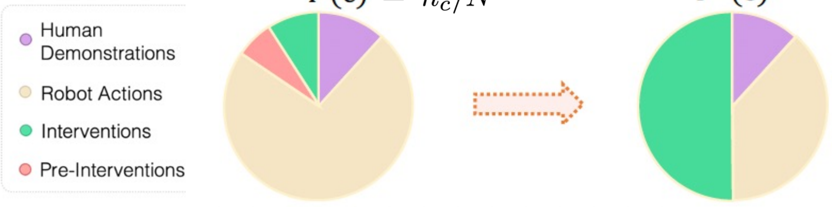
$$\theta^* = \arg\max_\theta \mathop{\mathbb{E}}_{(s,a)\sim\mathcal{D}} [w(s,a) \log \pi_\theta(a \mid s)]$$

Weighting Function

Intuition:

- We should upweight the state-action pairs of human intervention samples (↑)
- The samples before human intervention are less desirable and of low quality (↓)

Original Distribution $P(\text{c}) = n_c/N$ | After Reweighting $P^*(\text{c})$



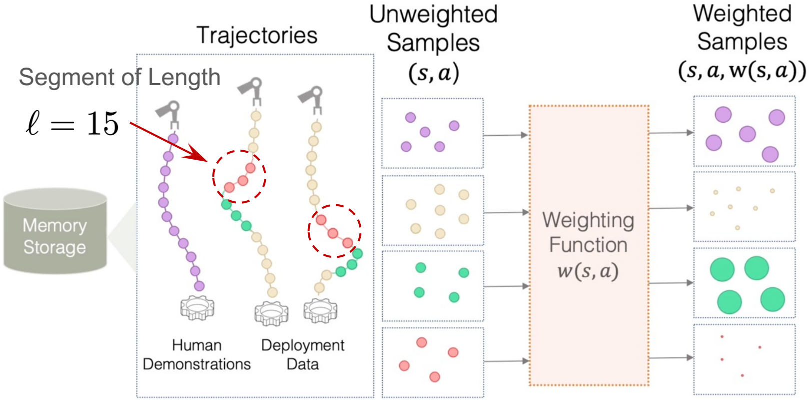
$P^*(\texttt{intv}) = \frac{1}{2}$

$P^*(\texttt{preintv}) = 0$

$P^*(\texttt{demo}) = P(\texttt{demo})$

$P^*(\texttt{robot})$

Weighting Function $w(s,a,c) = P^*(c)/P(c)$

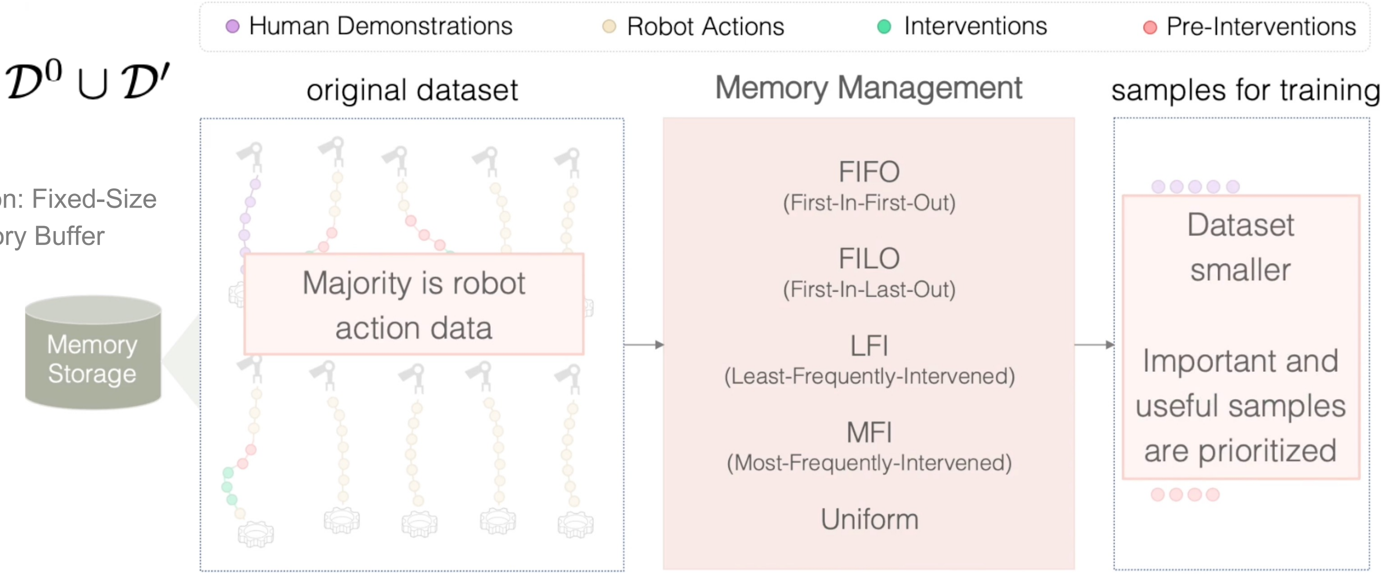


Proposed Approach: Memory Management

Research Question: How do we absorb the most useful data and preserve more valuable information for learning?

$$\mathcal{D}^1 \leftarrow \mathcal{D}^0 \cup \mathcal{D}'$$

Assumption: Fixed-Size Memory Buffer



Proposed Approach: Overall Workflow

Algorithm 1 Human-in-the-loop Learning at Deployment

Notations

L: memory buffer maximum fixed size
X: maximum deployment rounds
M: number of initial human demonstration trajectories
K: number of rollout episodes in each deployment round
b: batch size
n: number of gradient steps in each learning round
α: policy learning rate

▷ *warmstart phase*
Collect M human demonstrations $\tau_1, \ldots, \tau_M$
$\mathcal{D}^0 \leftarrow \{\tau_1, \ldots, \tau_M\}$
Initialize BC policy π_1^θ:
$\theta^* = \arg\max_\theta \mathbb{E}_{(s,a)\sim\mathcal{D}^0}\left[\log \pi_1^\theta(a \mid s)\right]$

Obtain Initial Policy

▷ *initial deployment data*
$\mathcal{D}^1 \leftarrow$ DEPLOYMENT(π_1^θ, $\mathcal{D}^0$)

▷ *deployment-learning loop*
for $i \leftarrow 1$ to X **do**
  Run in parallel:
    $\mathcal{D}^{i+1} \leftarrow$ DEPLOYMENT(π_i^θ, $\mathcal{D}^i$)
    $\pi_{i+1}^\theta \leftarrow$ LEARNING($\mathcal{D}^i$)

Robot Deployment

Policy Update

▷ *deployment thread*
function DEPLOYMENT(π_θ, $\mathcal{D}$)
  Collect rollout episodes $\tau_1, \ldots, \tau_K \sim p_{\pi_\theta}(\tau)$
  $\mathcal{D}^+ \leftarrow \mathcal{D} \cup \{\tau_1, \ldots, \tau_K\}$
  if $|\mathcal{D}^+| > L$ **then**
    Discard trajectories in $\mathcal{D}^+$ s.t. $|\mathcal{D}^+| \leq L$
      with a memory management strategy (in IV-C)
  return $\mathcal{D}^+$

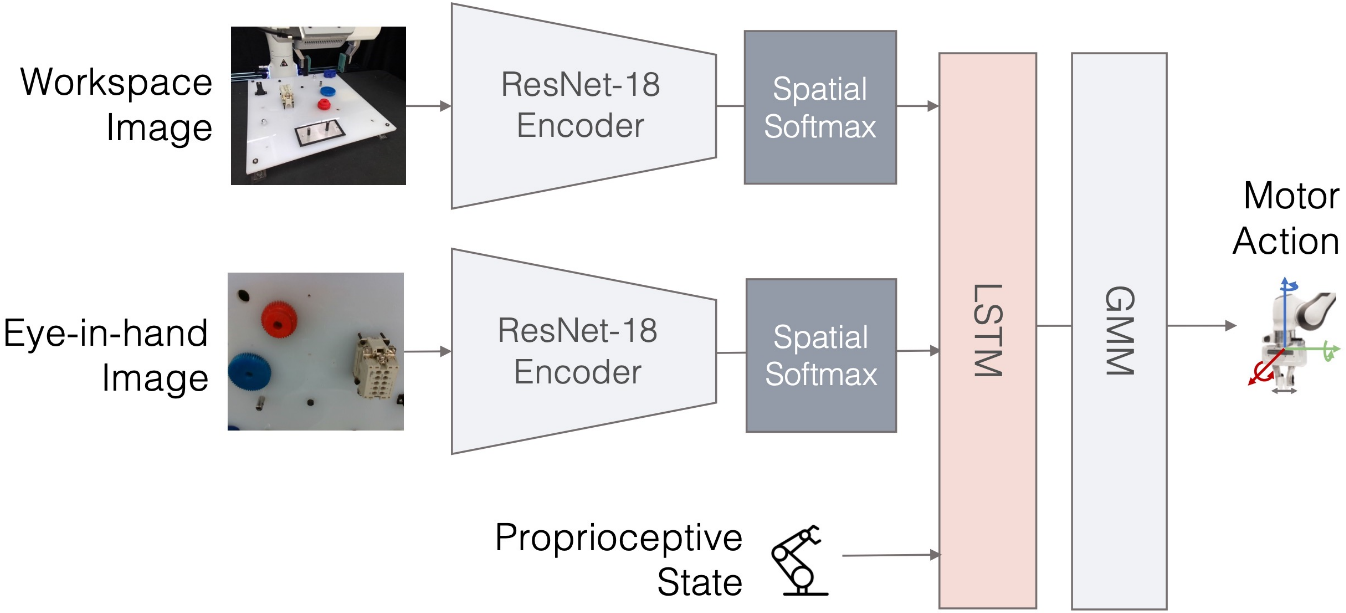
Memory Management

▷ *learning thread*
function LEARNING($\mathcal{D}$)
  Initialize π_θ
  for each class c **do**
    $\mathcal{D}_c \leftarrow \{(s, a, c') \in \mathcal{D} \mid c' = c\}$
    $P(c) \leftarrow |\mathcal{D}_c|/|\mathcal{D}|$
    Obtain $P^*(c)$ (see IV-D)
  for n gradient steps **do**
    Sample mini-batch $\left(s^i, a^i, c^i\right)_{i=1}^b \sim \mathcal{D}$
    Compute $w(s^i, a^i, c^i) \leftarrow \frac{P^*(c^i)}{P(c^i)}$ for the mini-batch
    $\mathcal{L}_\pi(\theta) = -\frac{1}{b}\sum_i \left[w(s^i, a^i, c^i) \cdot \log \pi_\theta(a^i \mid s^i)\right]$
    $\theta \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}_\pi(\theta)$
  return π_θ

Reweighting Scheme

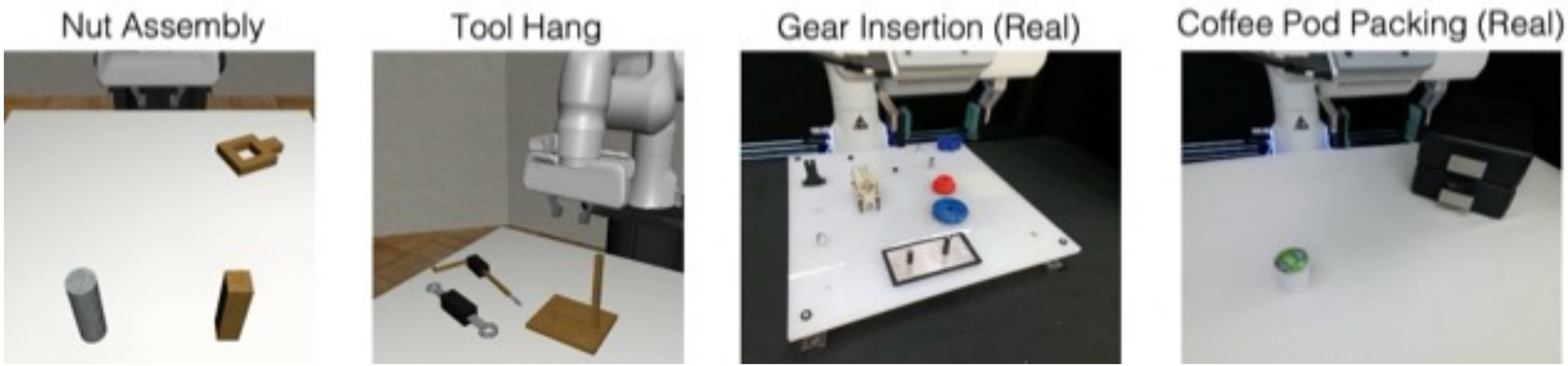
Implementation Details: Policy Architecture

Robot policy: BC-CNN



Experimental Setup

- **Robot hardware:** Franka Emika Panda robot arm equipped with a parallel jaw gripper
- **Simulation platform:** robosuite simulator
- **Human interface device:** spacemouse
- **Tasks:** Long-horizon and contact-rich manipulation tasks

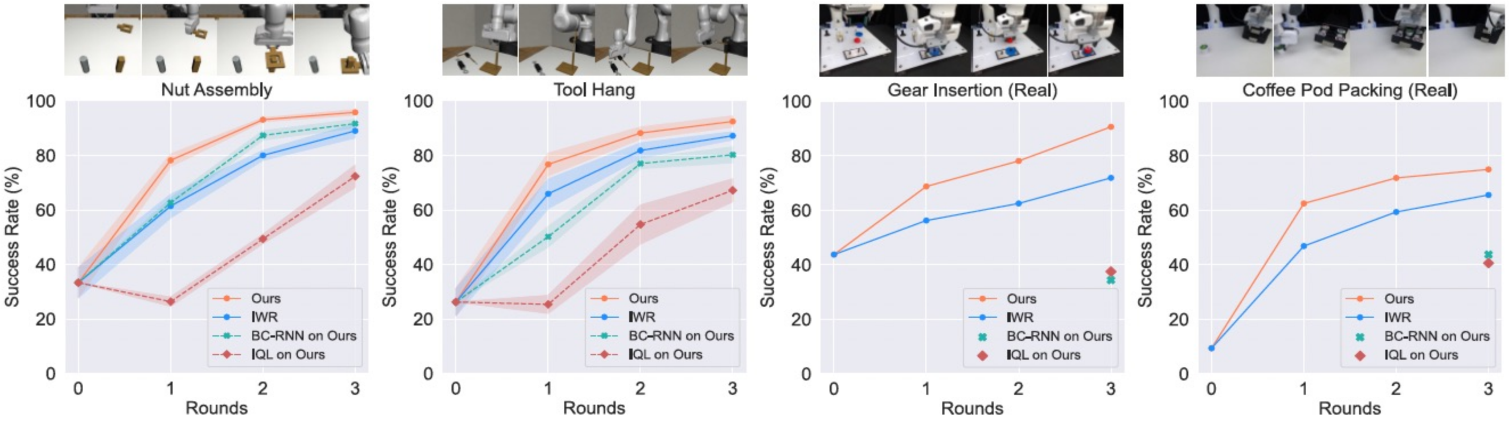


Experimental Results: Quantitative Evaluations

Baselines:

- Intervention Weighted Regression (**IWR**) → SOTA human-in-the-loop learning method for manipulation
- Behavioral Cloning with a policy network that's a RNN (**BC-RNN**) → SOTA Imitation learning algorithm
- Offline RL algorithm Implicit Q-Learning (**IQL**)

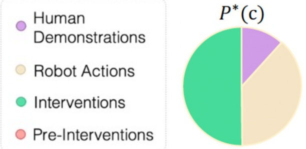
Success Rate of Autonomous Policy



Note: Human-robot team achieves a reliable task success of 100%

Experimental Results: Ablation Studies

Intervention Ratio Weight

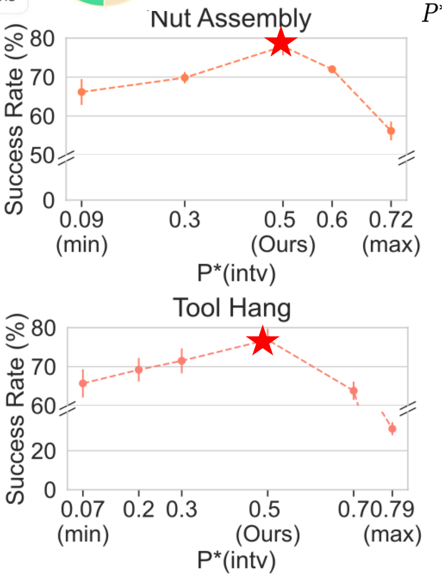


$P^*(\mathtt{intv}) = \frac{1}{2}$

$P^*(\mathtt{preintv}) = 0$

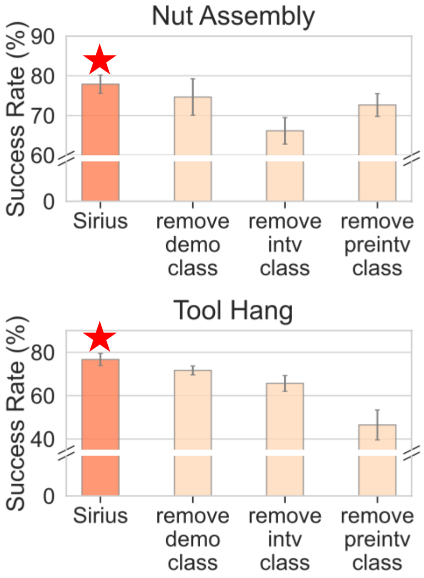
$P^*(\mathtt{demo}) = P(\mathtt{demo})$

$P^*(\mathtt{robot})$



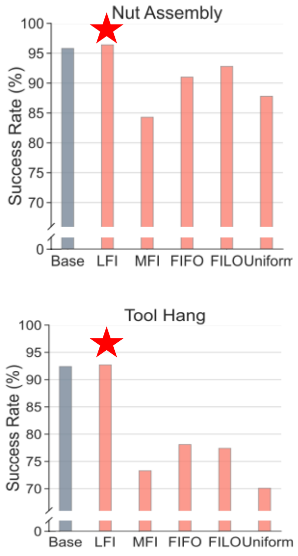
Weight Function Design

$$w(s, a, c) = P^*(c)/P(c)$$



Memory Management Strategies

$$\mathcal{D}^1 \leftarrow \mathcal{D}^0 \cup \mathcal{D}'$$

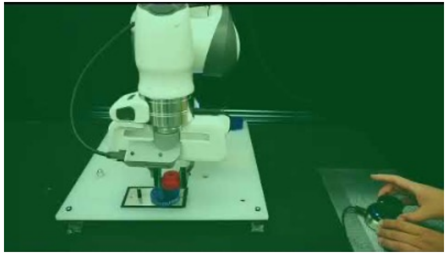


Experimental Results: Human Workload Reduction

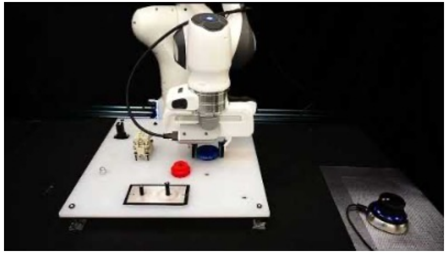
Minimize $\mathbb{E}_{\pi}[I_H(s_t)]$ (reduce human workload over time)

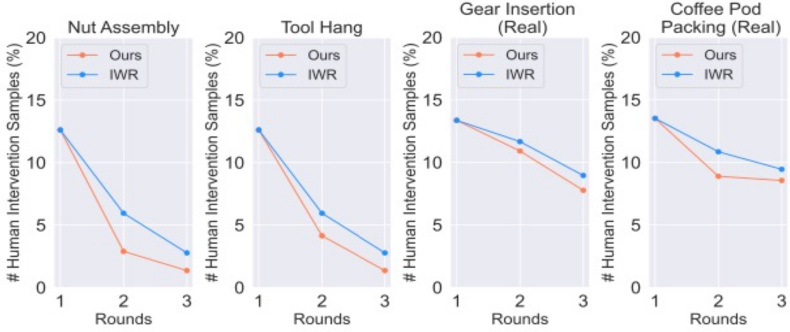
No-cut video of gear insertion deployment

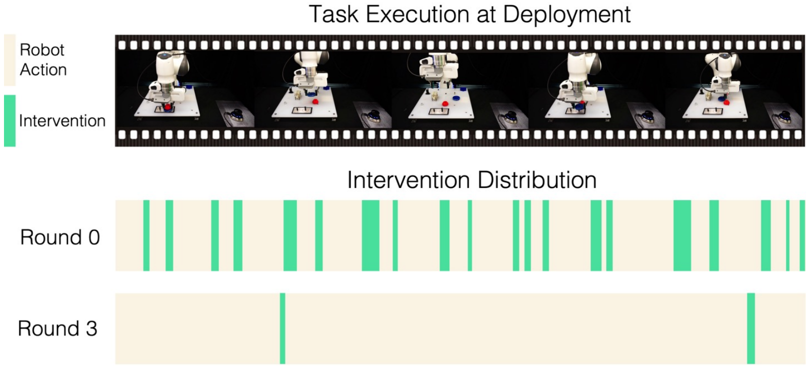
Round 0
(10 trials)



Round 3
(10 trials)







Critiques and Open Issues

- Policy Retraining and Computation Challenges
- Behavior Cloning and Negative Reinforcement
- Task-Specific Policy Networks vs. Lifelong Learning
- Success Rate of Autonomous Policy
- Integration of End-to-End and Hierarchical Approaches

Extended Readings

[Human-in-the-Loop Imitation Learning using Remote Teleoperation]

[Human-In-The-Loop Task and Motion Planning for Imitation Learning]

[Should I use Offline RL or Imitation Learning as the backbone for Human-in-the-loop Autonomy?]

Summary

Scientific hypotheses: Human interventions inform when the human lacks trust in the robot, where the risk-sensitive task states are, and how to traverse these status

Key insights:

- Introduce SIRIUS, a framework for human-in-the-loop robot manipulation and learning at deployment
- Develop an intervention-based weighted BC method for effectively using deployment data
- Design a practical system that trains and deploys new model continuously under memory constraints

Thank you!