Robotics: Science and Systems 2025
Los Angeles, CA, USA, June 21–25, 2025

Learning Interpretable Features from Interventions

Erin Hedlund-Botti*, Julianna Schalkwyk*, Nina Moorman, Sanne van Waveren, Lakshmi Seelam, Chuxuan Yang,
Russell Perkins, Paul Robinette, and Matthew Gombolay
Georgia Institute of Technology, University of Massachusettes Lowell
{erin.botti, julianna.schalkwyk, ninamoorman, swaveren, lseelam3, soyang}@gatech.edu,
russell_perkins@student.uml.edu, paul_robinette@uml.edu, matthew.gombolay@cc.gatech.edu

***Abstract*—The behavior of in-home robots must be adaptable to end-users to adequately address individual users' needs and preferences. Learning from Demonstration (LfD) is a common approach for customizing robot behavior, enabling non-expert users to teach robots how to perform tasks according to their preferences. While LfD allows users to teach robots tasks, it can be difficult for users to specify their individual needs a priori. Therefore, we propose Learning Interpretable Features from Interventions (LIFI), a user-friendly and streamlined method for personalizing robot behavior through interventions. This approach allows users to easily prompt the robot to adapt its behavior by intervening when the robot's behavior goes against user expectations. With LIFI, 1) the user intervenes to communicate that the robot is making a mistake, 2) the robot then learns an explanatory feature that describes the failure and 3) uses it to adjust its policy to correct the mistake, aligning with user-specific needs. In a between-subjects evaluation experiment with 48 participants, where the robot attempts household manipulation tasks, we demonstrate that adding features via LIFI improves objective performance and subjective measures, i.e., perceived workload, usability, and trust, compared to a no-feature baseline.**

***Index Terms*—robot learning, collaborative robots, human-robot interaction**

I. Introduction

Due to an aging population, there is a shortage of caregivers [24, 40]. Robots have the potential to offer assistance and enable older adults to age in place [42]. Assistive robots must be adaptable in order to meet users' diverse environments and changing needs as they age. Furthermore, useful assistive robots will need to adapt to new situations and recover from any failures or miss-steps that will inevitably happen when operating in the real world [43]. However, many current personalization approaches are costly or not scalable [22, 38]. Training a generalized policy that can adapt to different scenarios a priori requires a large data set for the policy to function well in a variety settings [44]. On the other hand, a custom-built control policy for each scenario requires significant time and effort from an expert roboticist [6]. Instead, users need an intuitive and sample-efficient method to correct a robot's behavior. Therefore, we introduce Learning Interpretable Features from Interventions (LIFI) (Figure 1), a novel framework that enables a robot to learn interpretable features from user interventions and utilizes these features to correct its mistakes.

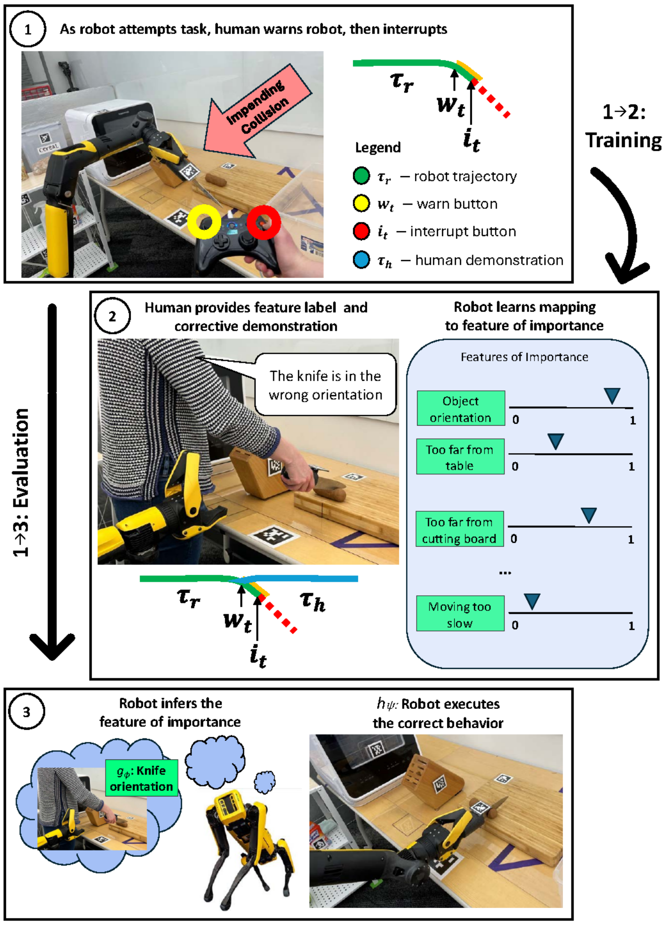

1
As robot attempts task, human warns robot, then interrupts
Impending Collision
τ_r
w_t
i_t
Legend
● τ_r – robot trajectory
● w_t – warn button
● i_t – interrupt button
● τ_h – human demonstration
1→2: Training
1→3: Evaluation
2
Human provides feature label and corrective demonstration
The knife is in the wrong orientation
τ_r
w_t
i_t
τ_h
Robot learns mapping to feature of importance
Features of Importance
Object orientation 0 1
Too far from table 0 1
Too far from cutting board 0 1
...
Moving too slow 0 1
3
Robot infers the feature of importance
g_ψ: Knife orientation
h_ψ: Robot executes the correct behavior

Fig. 1: (1) With LIFI, the human warns when the robot may begin to fail and interrupts when the robot is concretely failing. (2) To train LIFI, we collect a dataset of participant interruptions, feature labels, and corrective demonstrations. The demonstrations start from the warning to show the robot how to avoid the failure. (3) With a new user, the robot infers their reason for interruption and improves its behavior.

Our LIFI approach builds from Learning from Demonstration (LfD) [30]. In LfD, the robot learns from a recording of the human demonstrating the task. The simplest form of LfD, Behavioral Cloning (BC), mimics the human; however, BC is susceptible to covariate shift [31]. Inverse Reinforcement Learning (IRL) techniques attempt to understand context by

*Co-First Authors.

learning a reward function for the human's goal [26]; yet, IRL methods are not sample-efficient [3, 16]. Furthermore, most LfD techniques are not interpretable due to their reliance on black-box neural networks [32].

We want to enable people to teach a robot their preferences with a method that is sample-efficient and provides semantically meaningful feedback. Therefore, we learn interpretable features in the LIFI framework to teach the robot to correct its mistakes. We define interpretable here as the robot outputting semantically meaningful features in human-worded language. With LIFI, participants intervene once during robot task execution when the robot's behavior does not match the human's preference. Ideally, a user would communicate which features are important to them, and the robot would learn to prioritize the appropriate parts of the task, resulting in higher performance and trust in the robot [35]. However, it can be difficult for people to identify and communicate their preferences and key task features a priori [13]. A more effective approach may be for people to critique while observing a robot's attempt to perform a desired task. Prior work has investigated learning features from physical interventions [5, 7], but these methods do not use interpretable features and users were told a priori which features to pay attention to.

In our approach, we show that when people intervene during a robot failure, there is semantic information, environmental and temporal context clues, from which the robot can learn. For example, if someone stops the robot when the robot is putting bleach on a shelf the user deems "wrong", the robot could observe its distance relative to objects in the environment and learn that the important feature is to put the bleach with the other cleaning products instead of next to the food.

In this paper, we first introduce the LIFI framework. Then, we experimentally validate LIFI through multiple user studies, in which the robot attempts a series of household tasks and participants intervene when the robot fails or does not behave as desired. Initially, we conduct a pilot study to elicit a list of features from users. Then we collect a dataset of interventions, feature labels, and corrective demonstrations to train LIFI. Lastly, we evaluate the LIFI framework compared to baselines (i.e. without features, BC, and ablations of the method).

In this paper, we contribute the following:

1) We develop a new, user-friendly, low cognitive demand interface for feature specification through interventions.
2) We contribute the LIFI framework, an interpretable, feature-based approach that enables LfD policies to learn what users want and adapt to individual user's needs.
3) We demonstrate that LIFI's feature prediction generalizes to novel users and LIFI outperforms a no feature baseline ($p < .05$) for performance, workload, usability, and trust.

II. Related Works

LfD seeks to enable humans to teach robots new skills via human task demonstrations without requiring users to have programming experience [30]. Behavioral Cloning (BC) learns to mimic human actions, but inherently does not understand the human's intent behind the demonstrations [30]. Inverse Reinforcement Learning (IRL) methods attempt to learn a reward function for the human's intent, but these methods are not sample efficient [3]. LIFI fills both gaps by adding a feature vector to a BC policy that attempts to understand human intent, while being sample-efficient. Unlike IRL methods, LIFI does not need hours of training time at evaluation.

LIFI also addresses the need to know which features a user cares about when specifying a demonstration. Without this a priori feature knowledge, a feature mismatch may cause the robot to fail to learn the skill, which can result in trust degradation [20, 34, 17]. Researchers have considered various approaches for learning from robot failures with LfD. Our work is inspired by Kelly et al. [21] and Spencer et al. [39] who proposed having people take over task execution when the robot deviates from the desired behavior. However, prior work that learns from human interventions [21, 39, 5] do not learn interpretable features or provide explanations for the updated policy. LIFI improves upon state-of-the-art by learning interpretable and semantically meaningful features. Additionally, these methods require demonstrations from all users, whereas LIFI only requires the intervention, not a demonstration, from users at test time.

Prior work has also explored learning what feature prompted an instance of user feedback [7]. Bajcsy et al. show that learning one feature per intervention compared to all at once improves objective and subjective results [5]. In contrast, Levine et al. investigated learning which features to consider by constructing features from components using logical conjugations [23]. While this work shows how both features and a policy can be learned at the same time, it does not consider whether the features selected by this method are deemed important by the demonstrator. This could lead to a "correct" policy that does not adhere to the demonstrator's preferences.

We endeavor for our features to be based on human preference, readable, human-worded, and understandable [45]. Learning-based approaches that use neural networks are not yet guaranteed to be interpretable to a non-expert user [32]. Additionally, explainability creates more trust in a robot via transparency about the reasoning behind its actions [33]. Das et al. expand on learning interpretable features, showing that presenting both the context of the failure and preceding robots actions to be helpful [10]. As prior work has found that querying people affords interpretable, relevant features, we obtain our features directly from participants by soliciting feedback as they observe the robot attempt the task [9].

III. Technical Approach

In this section, we introduce the LIFI framework.

A. Preliminaries

LfD can be formulated as a Markov Decision Process without the reward function (MDP\R) which is represented by the 4-tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{T}, \gamma \rangle$. $\mathcal{S}$ is the set of states and $\mathcal{A}$ is the set of actions. In a model-based paradigm, the transition function $\mathcal{T} : \mathcal{S} \times \mathcal{A} \times \mathcal{S}' \rightarrow [0, 1]$ represents the probability of transitioning from state s to state, s', via action, a. γ is

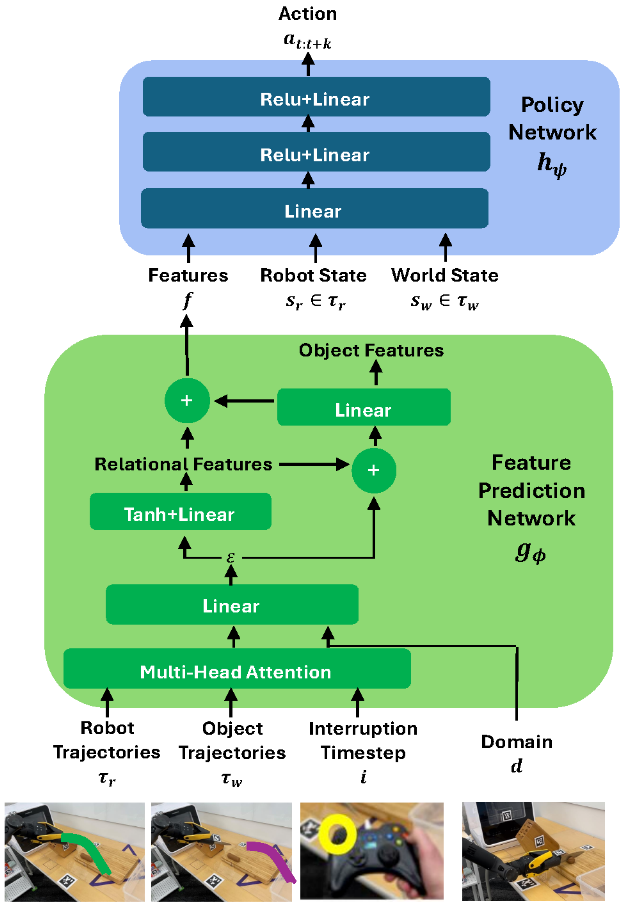


Fig. 2: The LIFI framework predicts a feature vector, f, (the reason that the user interrupted) with the Feature Prediction Network, g_ϕ, (green) and an improved robot policy, h_ψ, (blue). g_ϕ takes in the robot trajectory, τ_r, the objects in the environment (purple), τ_w, and the time of the interruption, i, to predict f. h_ψ predicts the robot's next action, a, from f, the state of the robot, s_r, and the environment objects, s_w.

the discount factor for future rewards. In an LfD paradigm, the agent learns a policy, $\pi : \mathcal{S} \rightarrow \mathcal{A}$, mapping states to actions, from a demonstrator provided set of trajectories $\{(s_t, a_t), \forall t \in \{1, 2, ...T\}\}$. Our approach is model-free, meaning that we do not know the transition function a priori.

B. LIFI Framework

The LIFI model architecture (Figure 2, Algorithm 1) consists of two parts. First, the Feature Prediction Network, g_ϕ, takes in the robot state trajectory, τ_r, the states of the objects in the world, τ_w, and the time of interruption, i, and outputs a feature vector, f. The time of interruption i is used to crop τ_r so that only the robot trajectory up to the time of interruption is given as input. Thus, g_ϕ uses the world state up to the time of interruption to predict the feature. Second, the Policy Network, h_ψ, takes the current robot state, s_r, world object state, s_w, and f and outputs the next robot action, a. In the LIFI framework, the set of features, comprising f, (detailed in Section IV-C) is split into two types: 1) relational features (e.g., too high, too close, etc.), which inform the robot of its behavior with respect to itself or objects in the environment and 2) the environment objects (e.g. table, dishwasher, etc.). For example, if the person interrupts because the robot is cutting the table instead of the food, the robot may learn the feature "too close to the table." To determine the features, we interview users in the Pilot Study (Section IV-C). We synthesize their responses to ascertain a list of features that are generalizable across manipulation tasks.

1) Feature Prediction Network: To learn the feature vector or reason for interruption, f, the g_ϕ network feeds τ_r, τ_w, and i through a multi-head self-attention layer. The reasoning for the attention layer is for the features to index on important aspects of the trajectories. The trajectories are the Cartesian pose of the robot end-effector and objects (position and quaternion). The attention layer output is a classification embedding, which is concatenated with a one-hot encoding of the domain and fed through a linear layer. This output is an encoding, ε, and is fed through a $\tanh$ activation and a linear layer, to predict the relational features. The relational features are then concatenated with ε and fed through a linear layer to predict the object features.

We train this model using the labeled features from the data collection study (Section V-A) and cross-entropy loss

$$L(\phi) = -\sum_{j}^{F} f \log \frac{e^{\hat{f}_j}}{\sum_{k}^{F} e^{\hat{f}_k}} \tag{1}$$

The list of features are known a priori and are determined to be common across manipulation tasks, from the Pilot Study (Section IV-C). Using softmax, the robot can then predict the most likely feature from the feature vector, f, and utilize templated language, from the labels, to communicate the feature to the user.

Algorithm 1: Training LIFI Framework

```
1. for each trial do
    if get interruption then
        collect interruptions $(i, \tau_r, \tau_w)$, feature labels
          $(f)$ and corrective demonstrations
          $(D = s_r, s_w)$.
    end
end
2. Initialize $\phi, \psi$.
3. for $(i, \tau_r, \tau_w, D)$ do
    Obtain predicted feature: $\hat{f} = g_\phi(i, \tau_r, \tau_w)$
    for $s_r, s_w, a \in D$ do
        Obtain predicted action: $\hat{a} = h_\psi(s_r, s_w, \hat{f})$
        Take one step of gradient descent on $\psi$ with
          $\hat{f}, s_r, s_w$ with Equation 2
    end
    Take one step of gradient descent on $\phi$ with
      $i, \tau_r, \tau_w, f$ via Equation 1
end
```

2) Policy Network: The policy network learns an improved robot policy based on user interruptions and demonstrations.

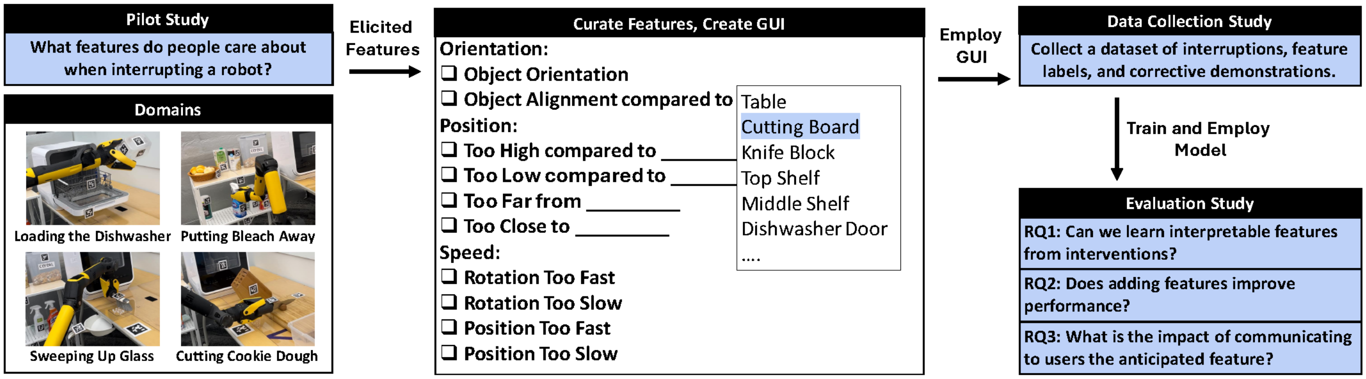

Pilot Study
What features do people care about when interrupting a robot?
Domains
Loading the Dishwasher
Putting Bleach Away
Sweeping Up Glass
Cutting Cookie Dough
Elicited Features
Curate Features, Create GUI
Orientation:
❑ Object Orientation
❑ Object Alignment compared to
Position:
❑ Too High compared to ______
❑ Too Low compared to ______
❑ Too Far from ______
❑ Too Close to ______
Speed:
❑ Rotation Too Fast
❑ Rotation Too Slow
❑ Position Too Fast
❑ Position Too Slow
Table
Cutting Board
Knife Block
Top Shelf
Middle Shelf
Dishwasher Door
....
Employ GUI
Data Collection Study
Collect a dataset of interruptions, feature labels, and corrective demonstrations.
Train and Employ Model
Evaluation Study
RQ1: Can we learn interpretable features from interventions?
RQ2: Does adding features improve performance?
RQ3: What is the impact of communicating to users the anticipated feature?

Fig. 3: We elicit features of interest from a Pilot Study and develop a user interface. Then, in the Data Collection Study, we collect a dataset of human interruptions, feature labels, and corrective demonstrations to train the LIFI model. Lastly, we validate the LIFI framework in the Evaluation Study. In all studies, the robot attempts household manipulation tasks.

The base policy is behavioral cloning [14], where the robot state, s_r, and world object state, s_w, are mapped to the robot's action, a. In the LIFI framework, the feature vector, f, is an additional input to the Policy Network, h_ψ, resulting in a feature-conditioned policy that adapts to user feedback. h_ψ employs action chunking, predicting the next k actions, to prevent the robot from stopping mid-trajectory [14], and utilizes three linear layers with ReLU activations. The correct action labels are derived from the combination of the robot trajectories prior to interruption concatenated with their respective corrective demonstrations. The policy network learns via an mean-squared error loss of the predicted actions $\hat{a}$ compared to the corrective actions a,

$$L(\psi) = \frac{1}{N}\sum_{j}^{N} ||a_j - \hat{a}_j||^2 \tag{2}$$

Further details (e.g., layer sizes) are in Appendix A.

Assumptions: The model assumes that the reason for interruption, in a given state, is homogeneous across people. Therefore, if two people interrupt in the same state and all previous states were identical, the model will predict the same feature. However, we assume that people will interrupt in different states based on their preferences. In practice, this is a reasonable assumption as the state space is continuous so two identical interruptions are highly unlikely.

IV. Experimental Design

We conduct three human-subjects experiments; an overview is depicted in Figure 3. We first conduct a pilot study to elicit a list of relevant features from users. We then conduct a data collection study where users interrupt the robot, label features, and provide corrective demonstrations to train the LIFI model. Third, we evaluate LIFI compared to baselines. This section details the research questions, experimental setup, and Institutional-Review-Board approved study procedures.

A. Research Questions

RQ1: *Can we learn interpretable features from interventions?* We investigate if we can accurately predict features of interest from user interventions, and if these predictions generalize to novel users.

RQ2: *Does adding features improve performance over a baseline without features?* We evaluate if learning the relevant features improves objective and subjective robot performance.

RQ3: *How does communicating to users the anticipated feature of interest impact the users' perceptions of the robot?* We investigate how communicating the predicted feature of interest when a user interrupts the robot changes the user's perceptions of the robot.

B. Experiment Setup

We employ the Spot robot [2], a ZED camera [1], and AprilTags [29], for object localization.

1) Domains: Prior work has shown that chore tasks are relevant for assistive robots [37, 36]. We design four household tasks as the domains (Figure 3):

1) **Loading the Dishwasher -** the goal is to place the plastic dish in the tabletop dishwasher.
2) **Putting Bleach Away -** the goal is to place the bleach bottle on the pantry shelf.
3) **Sweeping Up Glass -** with a hand-held broom, the goal is to sweep the "broken glass" into a dustpan.
4) **Cutting Food -** the robot must use a knife to cut the Play-Doh cookie roll.

Additionally, each domain affords human preferences (e.g., where to place a dish in the dishwasher, not putting cleaning items next to food).

2) Wizard-of-Oz Trajectories: To show the participants a consistent set of robot behaviors, we pre-specify five trajectories in each of the four domains (20 total). To compare across a variety of behaviors, each domain includes: one success, two objective failures, and two subjective failures. For the failures, we focus on system errors, in which the robot does not act as intended, and design errors, in which the robot acts as intended but should not have acted in that way [41]. An objective failure occurs when the robot fails to complete the task goal (e.g., colliding with the dishwasher). A subjective failure is when the robot achieves the goal without satisfying a user's preferences (e.g., placing the dish in the wrong orientation). To investigate preferences, we introduce failures that are subjective. These trajectories are used as the initial robot policy for all studies.

3) Domain and Trial Ordering: In all three studies (pilot, data collection, and evaluation), participants watch the robot complete the Wizard-of-Oz (WoZ) trajectories and interrupt if the robot is making a mistake or not behaving as desired. The frequency and timing of robot errors impact user behavior and perception of the robot [11]. As such, the order of successes and failures are randomized and then held constant across participants. In the data collection phase, participants experience one success, two objective failures, and two subjective failures per domain. Appendix Table II lists the ordering of trial outcomes for each domain that each participant experiences in the data collection phase. During the evaluation phase, participants observe three trials per domain (one success, one objective failure, and one subjective failure). Appendix Table III lists the ordering for the evaluation phase. Which of the two subjective and objective failures is randomized and counterbalanced across participants.

Participants also experience multiple domains in all three studies. Appendix Table I lists four domain orders, obtained via a Latin square. Each participant experiences one domain ordering condition. The domain orders are randomly assigned and counterbalanced across participants.

C. Pilot Study

The pilot study has two goals: 1) to elicit a list of features from participants and 2) to assess our study design. Since the robot behavior is WoZ, we need to determine whether participants perceive each trajectory as intended (i.e., successful trajectories are perceived as successes, objective failures as failures, and subjective failures are *sometimes* perceived as failures). Past experiments have shown that participants do not intervene, even if the robot is colliding with objects [27]. As such, we evaluate and improve the instructions to ensure participants intervene when observing objective robot failures.

1) Pilot Study Procedure: Participants observe each trajectory in each domain (as specified in Section IV-B3). After each trial, participants are asked to rate the success of the trajectory on a scale of 1 (unsuccessful) to 10 (successful). Participants also answer why they did or did not interrupt the robot. We transcribe the interviews and conduct a thematic content analysis with two reviewers [4] to determine a core set of important features across domains. This yields a dataset of features of importance from the population rather than using an experimenter-defined dataset. Based on our findings, we design a Graphic User Interface (GUI) for users to choose which feature reflects their reason for interruption.

2) Pilot Study Results: The pilot study consisted of 13 participants with a mean age of 23.8 and standard deviation (SD) of 3.58 (30.8% Female, 69.2% Male). On average, participants rated successes with a score of 7.8 out of 10, subjective failures with 5.8, and objective failures with 4.0. On the dishwasher task, participants rated successes lower than expected: 5.9, due to the robot releasing the dish from too high. On the bleach task, the objective failures were rated higher than expected: 5.6, due to not all participants rating collisions negatively. Therefore, for the data collection study, we redesigned the dishwasher task and told participants that the robot should complete the tasks without colliding.

Additionally, many participants did not interrupt until after the robot failed irrecoverably (e.g., wiped all the glass on the floor). We included a warn button that does not stop the robot so participants can indicate when the robot might be about to make a mistake. We later use this pre-failure moment to let people provide corrective demonstrations that show how to avoid the failure. We found that warning the robot allowed participants to signal to the robot, while still satisfying their curiosity about what the robot would do next.

After each trial, we asked participants why they interrupted the robot. The most common reasons were the orientation of the object that the robot was holding, the position of the object compared to other objects, and the speed of the robot. Participants also mentioned the force applied by the robot, specifically in the cutting task. However, we chose not to include force due to not being able to measure force from video demonstrations. The list of features is included in Figure 3. Representative quotes from participants are in Appendix D.

D. Data Collection Study

The goal of the data collection study is to collect a dataset from participants to train the LIFI framework.

1) Data Collection Procedure: We conduct a within-subjects ($n = 44$) data collection study where participants observe the robot complete trials in each of the domains, akin to the pilot study. In the data collection phase, the participant observes the robot execute a WoZ policy for five trials per domain, resulting in one success, two objective failures and two subjective failures. The domains are randomized and counterbalanced via Latin Squares (see Section IV-B3).

To interrupt the robot, the participants are instructed to press a yellow warn button before the robot makes a mistake, and the red stop button after the robot has made a mistake or the participant thinks the robot will not recover. After interrupting, participants indicate on our GUI which feature is their reason for interruption. Participants can choose more than one feature but must indicate which feature is most important. Next, we show participants a picture of when they first pressed the warn button. Using this image, participants then provide a demonstration via motion-capture, starting from the position where they first pressed the warn button, to show the robot how it should have performed the task differently. We use this data to train the LIFI model (Section III-B).

2) Manipulation Check: The data collection study included 44 participants with a mean age of 23.2 (SD=4.01, 29.5% Female, 70.5% Male). As with the pilot study, we first evaluated how participants perceived and interrupted the different types of trials. We employed Friedman's tests with Nemenyi-Wilcoxon-Wilcox post-hoc tests. Out of 10, participants rated successful trials with a mean of 8.26 (SD=.984), subjective failures with a mean of 5.81 (SD=1.21), and objective failures

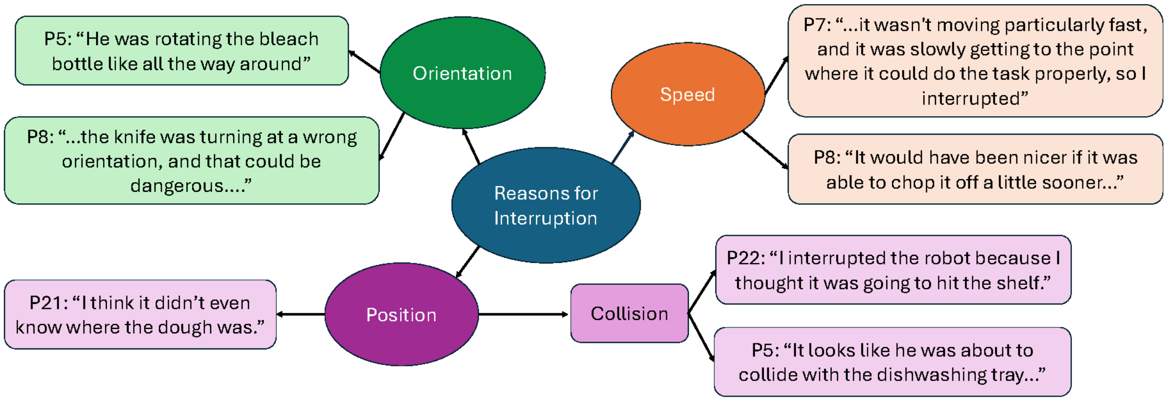

P5: “He was rotating the bleach bottle like all the way around”
Orientation
P8: “...the knife was turning at a wrong orientation, and that could be dangerous....”
Speed
P7: “...it wasn’t moving particularly fast, and it was slowly getting to the point where it could do the task properly, so I interrupted”
P8: “It would have been nicer if it was able to chop it off a little sooner...”
Reasons for Interruption
P22: “I interrupted the robot because I thought it was going to hit the shelf.”
P21: “I think it didn’t even know where the dough was.”
Position
Collision
P5: “It looks like he was about to collide with the dishwashing tray...”

Fig. 4: We solicited reasons for interruption in the Pilot Study ($n = 13$). We found that the main reasons people interrupted the robot were the position of the item, the orientation of the item, and the speed at which the robot was moving. Within position, people interrupted both because the position was generally incorrect and because of an impending collision with something else in the environment.

with a mean of 3.28 (SD=.969). There was a significant main effect between the trial types ($\chi^2(2) = 84.0$, $p < .001$). Successful trials were rated significantly higher than subjective failures ($p < .001$) and objective failures ($p < .001$), and subjective failures were rated significantly higher than objective failures ($p < .001$). Similarly, there was a significant main effect for interruption rate between the trial types ($\chi^2(s) = 80.0$, $p < .001$). Participants interrupted successful trials at a rate of 18.8% (SD=18.1), subjective failures at a rate of 51.1% (SD=23.2), and objective failures at a rate of 97.7% (SD=5.6). Participants interrupted successful trials significantly less than subjective ($p < .001$) and objective failures ($p < .001$). Lastly, participants interrupted subjective failures significantly less than objective failures ($p < .001$). Overall, the types of trials were designed and perceived as expected.

E. Evaluation Study

1) Conditions: In our third experiment, after our pilot study (Section IV-C) and data collection study (Section V-A), we manipulate the independent variable, *Feature Condition,* with the following conditions:

LIFI – Learned Feature (Ours): The robot infers the highest probability feature and then attempts the task using the learned feature (Section III-B). Then the robot communicates the feature (using templated language, e.g. “I think you interrupted because of X.”) as an explanation to the user.

MI-LIFI – Mixed-Initiative (Ours): The robot infers the best feature and then communicates the feature to the user. If the user feels that the robot should have chosen a different feature, the user can input the correct feature using the GUI. The robot then utilizes the learned, or corrected feature, as an input to the policy network, and attempts the task again.

BC – No Feature: The robot learns how to accomplish the task via BC, does not predict a feature, and does not communicate with the user.

Adv-LIFI – Adversarial Feature: This condition is the same as LIFI, except the robot infers the worst feature (minimum probability). Adv-LIFI accounts for bias when working with interactive systems.

2) Evaluation Study Procedure: We conduct a 4x4 between-subjects experiment ($n = 12$ per condition), where each participant experiences one Feature Condition (Section IV-E1). Participants experience three trials for each of the four domains. The domains are randomized and counterbalanced via Latin Squares (see Section IV-B3). In the evaluation phase, the number of trials was reduced because the BC policy performed similarly across trials, due to BC not personalizing to individual participant’s interruptions. It may have become obvious to participants that the robot was not learning from their interruptions if the robot performed similarly across too many trials. Therefore, the participant observes three trials per domain: one success, one objective failure, and one subjective failure. The objective and subjective failure trajectory chosen, from the two in the data collection, are randomized and counterbalanced across participants.

Participants first fill out pre-study surveys. Then participants observe the robot attempt a series of household tasks, and participants are instructed to interrupt using the same procedure as the data collection study. If participants interrupt a trial, the robot attempts the task again and communicates the feature of importance, dependent on the Feature Condition. After each trial, participants complete the post-trial surveys. Lastly, participants complete the post-study surveys.

F. Metrics

We now describe the metrics employed to evaluate our framework. More details on metrics are in Appendix C.

1) Framework Training Metrics: We first validate the model offline with a holdout set from the data collection data.

Feature Accuracy: We compare the model predicted features to the participant labeled features using an all-or-nothing and a partial-credit metric. For both metrics, if the participant’s top feature matched the model’s predicted feature, we scored that as a 1. For the all-or-nothing metric, if the model’s feature differed at all from the top feature given by the participant, we gave it a score of 0. For the partial-credit metric, if the predicted feature was a feature given by the participant but was not the top feature, we gave it a score of 0.75. If the model guessed a feature or object that the participant input, we gave a score of 0.5.

Policy Error: We define Policy Error as the difference between the policy generated trajectories and the successful WoZ trajectories to measure error. Additionally, to quantify personalization from the features, we compare the generated trajectories to the participant demonstrations. We calculate error by aligning trajectories using dynamic time warping [25] and measuring the absolute pose error (APE) [15]. To gain further insight into the policy error, we also separate out the error into orientation and position error.

2) Pre-Study Metrics: At the beginning of the evaluation study, we collect the following metrics.

(a) **Demographics:** We collect participant age and gender.
(b) **Personality:** We employ the Mini-IPIP [12].
(c) **Negative Attitudes Towards Robots (NARS):** We measure the three NARS subscales [28].

3) Post-Trial Metrics: After each trial, we measure the accuracy of the features and robot policy. All scales from 1 to 10 are from 1 (not successful) to 10 (successful).

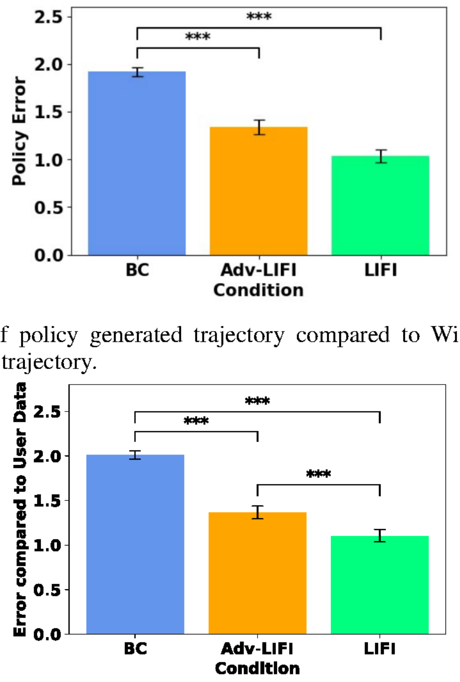
(a) **Feature Ratings:** Participants rate the robot's predicted feature on a scale from 1 to 10.
(b) **Feature Accuracy:** We compare participant-provided feature(s) to the model-predicted feature as described in the Framework Training Metrics.
(c) **Perceived Policy Accuracy:** Users rate the robot's performance on a scale from 1 to 10.
(d) **Policy Error:** We compare the robot's trajectory to a successful trajectory using the same method as the framework training metric.
(e) **Qualitative Interview:** We interview participants to understand why they did or did not interrupt.

4) Post-Study Metrics: After the evaluation study, we collect the following metrics:

(a) **Interruption Rate:** We measure how often participants intervened across trials.
(b) **Usability:** We employ the System Usability Scale [8].
(c) **Workload:** We measure workload via the NASA Task Load Index [18].
(d) **Trust:** We employ the trust scale by Jian et al. [19].
(e) **Qualitative Interview:** We interview participants about the process of working with the robot.

V. Results and Discussion

In the results, for each statistical test, we compared conditions using an Analysis of Variance (ANOVA). To check for confounding factors on subjective metrics, we systematically added the demographic variables as covariates to each model (i.e., age, gender, personality, and attitudes towards robots), only keeping the covariate if adding it lowered the model's AICc (Akaike Information Criterion for small sample sizes). Additionally, each parametric model was tested for normality and homoscedasticity. If assumptions failed, a non-parametric version of the test was employed. Further details on models and tests for assumptions are in the Appendix Table VI.



(a) APE of policy generated trajectory compared to Wizard-of-Oz successful trajectory.

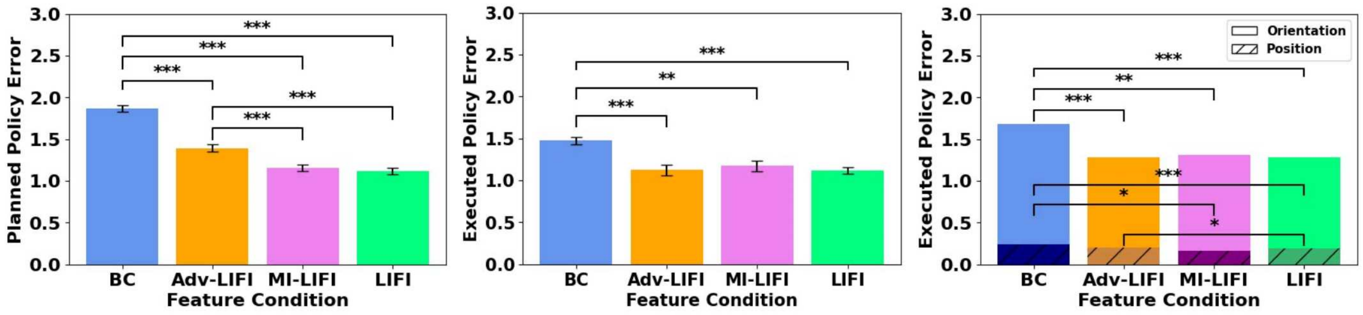
(b) APE of policy generated trajectory compared to user demonstrated trajectory.

Fig. 5: LIFI outperforms baselines in offline model validation. In Figure 5a, LIFI is significantly closer to a successful trajectory compared to the trajectories generated by BC and Adv-LIFI. In Figure 5b, LIFI is significantly closer to the participants' demonstrated trajectories compared to the trajectories generated by BC and Adv-LIFI. Also, Adv-LIFI is significantly closer to participant trajectories than BC. The LIFI framework better personalizes to user preferences for the held-out test set.

A. Data Collection Study: Model Validation

We first conduct an offline validation of the model on the data collection dataset. We split the dataset into an 80/20 training and test set with 5-fold cross-validation. After training the LIFI framework on the training set, we evaluated the model on the test set using feature accuracy and policy error metrics.

1) Feature Accuracy: We compared the predicted features to the labeled features in the test dataset. The model predicted the user-specified most important feature with 67% accuracy. While this accuracy may seem low, there is ambiguity in the feature labels and participants often chose more than one label. One participant might say the "knife was too close to the table," while another would say the "knife was too low compared to the table," and a third may choose both options. After inspecting where the model guessed "wrong," many incorrect guesses were due to ambiguity (e.g., choosing object orientation instead of object alignment). Therefore, we determine the utility of our feature predictions in the evaluation study, and discuss positive results on the accuracy of our fea-



(a) Planned policy error.

(b) Executed policy error.

(c) Orientation and position policy error.

Fig. 6: (a): LIFI plans a more accurate policy than BC and Adv-LIFI. (b) LIFI outperforms BC during policy execution. (c) After separating orientation and position error of the executed policy, LIFI outperforms BC and Adv-LIFI on position error.

ture prediction (Section V-B). **Takeaway: Combining results from the data collection and evaluation studies, LIFI learns an informative relationship between user interruptions and the semantic concept of that failure (RQ1).**

2) Policy Error: We generated trajectories for the LIFI, Adv-LIFI, and BC conditions, and compared using Friedman's tests with Nemenyi-Wilcoxon-Wilcox post-hocs. In Figure 5a, we found a significant difference in the error between the generated trajectories and WoZ successful policies across conditions ($\chi^2(2) = 69.3$, $p < .001$). LIFI had significantly lower error than BC ($p < .001$), as did Adv-LIFI ($p < .001$).

We also calculate the error between the generated trajectories and the participants' demonstrations. There was a significant main effect for the policy error across the conditions ($\chi^2(2) = 89.1$, $p < .001$). As shown in Figure 5b, LIFI had significantly lower error than BC ($p < .001$) and Adv-LIFI ($p < .001$), showing that adding correct features better personalized to users. Also, Adv-LIFI had significantly lower error than BC ($p < .001$), meaning that providing the incorrect feature still personalized better than no features. **Takeaway: On the hold-out set, adding features improved performance over a baseline without features (RQ2).**

B. Evaluation Study

The evaluation study included 48 participants with a mean age of 24.2 and an SD of 3.24 (33.3% Female, 64.6% Male, and 2.1% Other). We conducted ANOVAs with Tukey post-hocs (or Kruksal-Wallis tests with Dunn's post-hoc, for non-parametric tests). We report effect sizes from Tukey post-hoc as TD and effect sizes from Dunn's post hoc as r.

1) Manipulation Check: We checked interruption rates to ensure that the WoZ trajectories were not perceived differently across conditions. There was no significant difference in interruption rate between conditions ($\chi^2(3) = 1.18, p = .758$).

We next analyzed how participants perceived the features inferred by LIFI and MI-LIFI compared to Adv-LIFI. We found a significant main effect for perceived success between feature conditions ($F(2, 33) = 23.0$, $p < .001$). Participants perceived features inferred by LIFI ($p < .001, TD = 2.27$) and MI-LIFI ($p < .001, TD = 2.83$) as significantly more correct than features from Adv-LIFI.

When comparing between LIFI's guess and the features labeled by participants, LIFI and MI-LIFI combined had an average all-or-nothing score of 0.43 (SD=.50) and a partial-credit score of 0.64 (SD=.36). Comparatively, Adv-LIFI had an all-or-nothing score of 0.17 (SD=.38) and a partial-credit score of 0.23 (SD=.39). We found a statistically significant difference in score between LIFI, MI-LIFI, and Adv-LIFI for the all-or-nothing metric ($\chi^2 = 15.11, p < .001$) and the partial-credit metric ($\chi^2 = 41.24, p < .001$). For the all-or-nothing metric, LIFI outperformed Adv-LIFI ($p = .04, r = 0.21$) and MI-LIFI outperformed Adv-LIFI ($p < .001, r = 33$). In the partial credit metric, LIFI outperformed Adv-LIFI ($p < .001, r = 0.45$) as did MI-LIFI ($p < 0.001, r = 0.51$). **Takeaway: LIFI and MI-LIFI were better at picking a feature that matched participant expectations than Adv-LIFI. The LIFI framework can adequately predict correct and incorrect features intentionally, providing further evidence for RQ1.**

2) Objective Metrics (Policy Error): We compare the trajectories generated by the policy for each condition to the WoZ successful trajectory (Figure 6a). We found a significant main effect for *planned policy error* between conditions ($F(3, 44) = 70.1$, $p < .001$). The planned policy error for BC is significantly higher than Adv-LIFI ($p < .001, TD = 0.47$), LIFI ($p < .001, TD = 0.75$), and MI-LIFI ($p < .001, TD = 0.71$). Furthermore, error for Adv-LIFI is significantly higher than LIFI ($p < .001, TD = 0.27$) and MI-LIFI ($p < .001, TD = 0.24$). The trajectories for LIFI and MI-LIFI are significantly closer to success compared to BC and Adv-LIFI, and Adv-LIFI is significantly closer to success than BC.

The planned trajectory was sent to the robot, however, the robot may have been unable to complete the trajectory due to infeasible kinematics or the experimenter interrupting the trajectory early due to a collision. Therefore, we also calculate error between the robot's executed trajectory and the WoZ successful trajectory across conditions. We found a main effect across feature conditions ($F(3, 44) = 19.1, p < .001$). BC has significantly higher *executed policy error* compared to Adv-LIFI ($p < .001, TD = 0.35$), LIFI ($p < .001, TD = 0.36$), and MI-LIFI ($p = .001, TD = 0.30$). Figure 6b shows that LIFI, MI-LIFI, and Adv-LIFI are all significantly closer to success than BC. **Takeaway: LIFI outperforms BC, even**

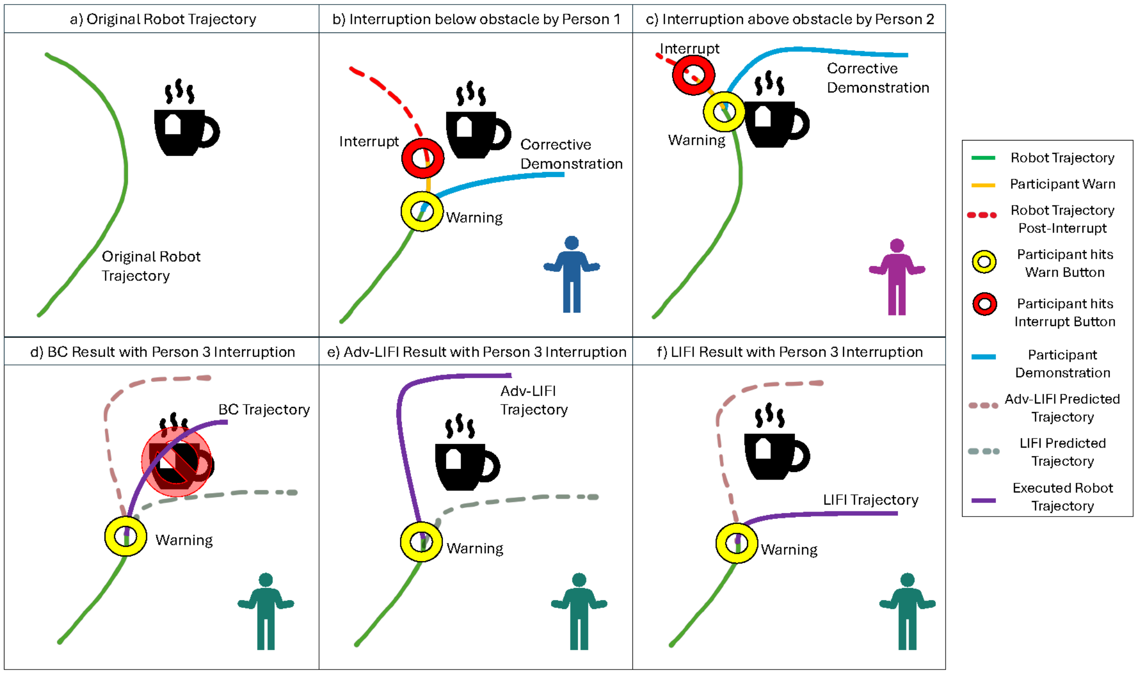

a) Original Robot Trajectory
Original Robot Trajectory
b) Interruption below obstacle by Person 1
Interrupt
Corrective Demonstration
Warning
c) Interruption above obstacle by Person 2
Interrupt
Corrective Demonstration
Warning
d) BC Result with Person 3 Interruption
BC Trajectory
Warning
e) Adv-LIFI Result with Person 3 Interruption
Adv-LIFI Trajectory
Warning
f) LIFI Result with Person 3 Interruption
LIFI Trajectory
Warning
Robot Trajectory
Participant Warn
Robot Trajectory Post-Interrupt
Participant hits Warn Button
Participant hits Interrupt Button
Participant Demonstration
Adv-LIFI Predicted Trajectory
LIFI Predicted Trajectory
Executed Robot Trajectory

Fig. 7: This figure shows the differences between the BC, Adv-LIFI, and LIFI results. An original robot trajectory in panel (a) is interrupted twice, once below the obstacle in panel (b) and once next to the obstacle in panel (c). As shown in panels (d), (e), and (f), the three algorithms have differing results, even when interrupted at the same point. In panel (d), the BC baseline averages the two demonstrated trajectories in (b) and (c), producing a trajectory that collides with the obstacle. In panel (e), Adv-LIFI follows the path from the interruption above the obstacle instead of the one below the obstacle (green dashed line). It does not collide with the obstacle but may not have been what the user intended, as Person 1, who interrupted at a similar point, gave a different demonstration than what the robot produced. In panel (f), LIFI correctly replicates the demonstration of what to do when interrupted below the obstacle as shown in panel (b).

when using the wrong feature.

Due to the unexpected similarities between the Adv-LIFI, MI-LIFI, and LIFI errors, we further analyzed the executed policy error by comparing the position and orientation error separately (Figure 6c). There was a significant main effect for position error across conditions ($\chi^2(3) = 22.0$, $p < .001$). LIFI had significantly less position error than BC ($p < .001, r = -0.94$) and Adv-LIFI ($p = .029, r = -0.55$). MI-LIFI also had significantly less position error than BC ($p = .026, r = -0.57$). For orientation error, we found a significant main effect between conditions ($\chi^2(3) = 19.1$, $p < .001$). The orientation error followed the same trends as the overall executed policy error. BC had significantly higher orientation error compared to Adv-LIFI ($p < .001, r = 0.72$), LIFI ($p < .001, r = 0.79$), and MI-LIFI ($p = .001, r = 0.64$).

Remark: While it may seem counterintuitive that BC would perform worse than Adv-LIFI, this difference in performance could be attributed to a mode collapse that is present in BC but not in Adv-LIFI. Participants' demonstrations varied widely, yielding multiple "modes." As shown in Figure 7, BC combines all corrective demonstrations, ignoring distinct "modes" or interruption points, resulting in an averaged trajectory that performs poorly as it tries to address all reasons for interruption at once. Adv-LIFI, as with all LIFI models, still learns to distinguish modes and, even though Adv-LIFI chooses the wrong mode at test time, the result is better than BC suffering from mode collapse.

While the overall executed policy error did not show significant differences between LIFI and Adv-LIFI, LIFI outperformed Adv-LIFI with respect to position error. Furthermore, from observations, the robot performed distinct behaviors (example trajectories are depicted in Appendix F). For example, LIFI successfully places the bleach on the shelf, however, the robot first stops mid-trajectory and spins the bottle in place before succeeding. BC and Adv-LIFI both drop the bleach bottle on the ground after spinning. This demonstrates how LIFI could have lower position error than Adv-LIFI, but not significantly different orientation error. **Takeaway: LIFI outperformed both Adv-LIFI and BC in terms of position error (RQ2).**

3) Subjective Metrics: We compared participants' perceived success of the learned trajectories and perceived workload, usability, and trust ratings across feature conditions, to answer **RQ3**. Participant quotes are included in Appendix G.

a) Perceived Improvement: Due to the between-subjects experiment design, some participants' baseline ratings were higher than others. We subtracted the success ratings of the trial that participants interrupted from the learned trajectory to determine which conditions improved perceived performance. In Figure 8, LIFI and MI-LIFI have ratings above 0, meaning

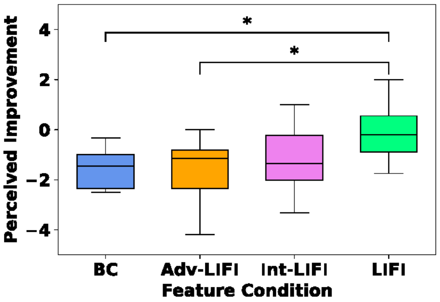


Fig. 8: Participants perceive that LIFI improves after interruption better than BC and Adv-LIFI.

that performance improved, whereas BC and Adv-LIFI usually degraded performance. We found a main effect for perceived improvement between conditions ($F(3, 42) = 4.40, p = .009$). LIFI had significantly better perceived improvement compared to BC ($p = .027, TD = 1.43$) and Adv-LIFI ($p = .027, TD = 1.43$). **Takeaway: LIFI and MI-LIFI show improvement, while BC and Adv-LIFI usually worsen. Participants rated LIFI with significantly more improvement than BC and Adv-LIFI.**

b) Workload: We found a significant main effect for perceived workload between conditions ($F(3, 42) = 4.89$, $p = .005$). Participants perceived BC to be significantly more workload than LIFI ($p = .012, TD = 21.7$) and MI-LIFI ($p = .008, TD = 22.8$). Additionally, the NARS subscale for social influence of robots positively impacted perceived workload ($F(1, 42) = 5.02$, $p = .030$). People who feel more negatively towards the social influence of robots perceived interacting with the robot as more workload. **Takeaway: Participants perceived conditions with correct feature explanations (LIFI and MI-LIFI) as significantly less workload than BC.**

c) Usability: There was a significant main effect for usability across feature conditions ($F(3, 41) = 3.53$, $p = .023$). LIFI had significantly higher usability scores compared to BC ($p = .040, TD = 16.7$). The NARS subscale for social influence of robots negatively impacted usability scores ($F(1, 41) = 14.1, p < .001$). Participants who were more wary of robot social influence thought the robot was less usable. **Takeaway: Participants perceived LIFI as significantly more usable than BC.**

d) Trust: We found a significant main effect for trust across feature conditions ($F(3, 41) = 4.68$, $p = .007$). LIFI had significantly higher trust scores compared to BC ($p = .004, TD = 9.41$). Trust and usability can be correlated with performance [17], therefore, we posit that MI-LIFI was not perceived as highly due to lower performance than LIFI. The personality trait agreeableness positively impacted trust ($F(1, 41) = 4.41$, $p = .042$). Furthermore, the NARS subscale for social influence negatively impacted trust scores ($F(1, 41) = 17.0$, $p < .001$). Participants who are more agreeable were more likely to trust the robot, whereas participants who are more wary of robot social influence were less likely to trust the robot. **Takeaway: Participants trusted LIFI significantly more than BC.**

VI. Limitations and Future Work

This work presents a promising framework for learning from interventions. We would like to now acknowledge areas for growth in future work. Our experiments included relatively small sample sizes ($n = 44, 48$). Therefore, the data collection phase did not adequately cover the distribution of possible responses, resulting in some out-of-distribution features having lower performance in the evaluation study. In the future, we will recruit a more diverse population, including older adults and users of assistive robots, to improve our model.

Another limitation is that participants selected from a predefined set of features. When participants felt the options were lacking, they produced a broader variety of responses, resulting in lower feature success ratings. There is some ambiguity or overlap between some of the features. the robot attempts to place the bleach on the shelf, the options “too low compared to middle shelf” and “too close to middle shelf” have different meanings. However, in the case where the robot cuts the table instead of the cookie dough on the cutting board, the labels “too close to table” and “too low compared to table” are similar. During the evaluation, the robot often predicted “too far from cutting board.” Some participants agreed with the robot, some gave the robot partial credit, and others said the robot should have said a feature involving the table. The LIFI framework assumes that if two participants interrupt at the same point in the trajectory, it is for the same reason; however, the way participants describe the problem is not always homogeneous. During data collection, participants did not provide homogeneous labels, which added noise to the dataset. Additionally, during evaluation, participants were not homogeneous in their interpretation of the features. Incorporating these multiple types of heterogeneity is an exciting avenue for future work. We plan to incorporate natural language reasons for interruption instead of choosing from a list to cover a wider range of features and heterogeneity of responses.

Furthermore, the pre-defined errors in the WoZ trajectories may have biased the dataset. Therefore, we plan to collect data using a wider variety of initial policies as well as with a wider variety of tasks to further demonstrate how LIFI can generalize. We also plan to improve the framework for generalization to new domains.

Additionally, our framework currently only uses one interruption. A more natural flow could include more of a conversational interaction, with multiple interactions as needed to correct the robot’s behavior if the first correction is not successful or does not fully correct the behavior. In future work, we plan to incorporate these multiple interactions.

VII. Conclusion

We introduced LIFI, a novel framework for learning interpretable features from user interventions. We conducted a pilot experiment to obtain relevant features for our tasks, a data collection experiment to collect corrective demonstrations that we used to train our model, and an evaluation experiment to investigate the efficacy of the LIFI framework. The model learns features, from these corrective interruptions, that improve a robot's policy. The LIFI framework outperformed a no feature baseline on objective metrics including executed policy error as well as on subjective metrics such as perceived improvement, workload, usability, and trust.

Acknowledgments

Funding for this work was graciously provided by the National Science Foundation under grant IIS-2340177 and grant IIS-2112633.

References

[1] Stereolabs zed 2 stereo camera, 2024. URL https://www.stereolabs.com/products/zed-2.

[2] February 2024. URL https://bostondynamics.com/products/spot/. Publication Title: Boston Dynamics.

[3] Pieter Abbeel and Andrew Y. Ng. Apprenticeship Learning via Inverse Reinforcement Learning. In *Proceedings of the Twenty-First International Conference on Machine Learning*, ICML '04, page 1, New York, NY, USA, 2004. Association for Computing Machinery. ISBN 1-58113-838-5. doi: 10.1145/1015330.1015430. URL https://doi.org/10.1145/1015330.1015430. event-place: Banff, Alberta, Canada.

[4] Rosemarie Anderson. Thematic content analysis (TCA). *Descriptive presentation of qualitative data*, 3:1–4, 2007.

[5] Andrea Bajcsy, Dylan P. Losey, Marcia K. O'Malley, and Anca D. Dragan. Learning from Physical Human Corrections, One Feature at a Time. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '18, pages 141–149, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 978-1-4503-4953-6. doi: 10.1145/3171221.3171267. URL https://doi.org/10.1145/3171221.3171267. event-place: Chicago, IL, USA.

[6] Pietro Bilancia, Juliana Schmidt, Roberto Raffaeli, Margherita Peruzzini, and Marcello Pellicciari. An overview of industrial robots control and programming approaches. *Applied Sciences*, 13(4):2582, 2023.

[7] Andreea Bobu, Marius Wiggert, Claire Tomlin, and Anca D Dragan. Feature expansive reward learning: Rethinking human input. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, pages 216–224, 2021.

[8] John Brooke. SUS-A quick and dirty usability scale. *Usability evaluation in industry*, 189(194):4–7, 1996. Publisher: London, England.

[9] Justin Cheng and Michael S Bernstein. Flock: Hybrid crowd-machine learning classifiers. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*, pages 600–611, 2015.

[10] Devleena Das, Siddhartha Banerjee, and Sonia Chernova. Explainable AI for Robot Failures: Generating Explanations That Improve User Assistance in Fault Recovery. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '21, pages 351–360, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 978-1-4503-8289-2. doi: 10.1145/3434073.3444657. URL https://doi.org/10.1145/3434073.3444657. event-place: Boulder, CO, USA.

[11] M. Desai, P. Kaniarasu, M. Medvedev, A. Steinfeld, and H. Yanco. Impact of robot failures and feedback on real-time trust. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 251–258, 2013. doi: 10.1109/HRI.2013.6483596.

[12] M Donnellan, Frederick Oswald, Brendan Baird, and Richard Lucas. The Mini-IPIP Scales: Tiny-yet-Effective Measures of the Big Five Factors of Personality. *Psychological assessment*, 18:192–203, July 2006. doi: 10.1037/1040-3590.18.2.192.

[13] Boi Faltings, Pearl Pu, and Paolo Viappiani. Preference-based Search using Example-Critiquing with Suggestions. *The journal of artificial intelligence research*, 27: 465–503, December 2006. doi: 10.1613/jair.2075.

[14] Zipeng Fu, Tony Z. Zhao, and Chelsea Finn. Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation, 2024.

[15] Michael Grupp. evo: Python package for the evaluation of odometry and SLAM., 2017. URL https://github.com/MichaelGrupp/evo.

[16] Shuai Han, Mehdi Dastani, and Shihan Wang. Sample Efficient Reinforcement Learning by Automatically Learning to Compose Subtasks, 2024. _eprint: 2401.14226.

[17] Peter A. Hancock, Deborah R. Billings, Kristin E. Schaefer, Jessie Y.C. Chen, Ewart J. De Visser, and Raja Parasuraman. A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction. *Human Factors*, 53 (5):517–527, 2011. ISSN 00187208. doi: 10.1177/0018720811417254. ISBN: 0018720811417.

[18] S. G. Hart and Lowell E. Staveland. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. 1988.

[19] Jiun-Yin Jian, Ann Bisantz, and Colin Drury. Foundations for an Empirically Determined Scale of Trust in Automated Systems. *International Journal of Cognitive Ergonomics*, 4:53–71, 2000. doi: 10.1207/S15327566IJCE0401_04.

[20] Jason Johnson. *Type of Automation Failure: The Effects on Trust and Reliance in Automation*. PhD Thesis, 2004. Issue: December.

[21] Michael Kelly, Chelsea Sidrane, Katherine Driggs-Campbell, and Mykel J. Kochenderfer. *HG-DAgger: Interactive Imitation Learning with Human Experts*. 2018. _eprint: 1810.02890.

[22] Timothée Lesort, Vincenzo Lomonaco, Andrei Stoian, Davide Maltoni, David Filliat, and Natalia Díaz-Rodríguez. Continual learning for robotics: Definition, framework, learning strategies, opportunities and challenges. *Information fusion*, 58:52–68, 2020.

[23] Sergey Levine, Zoran Popovic, and Vladlen Koltun.

Feature construction for inverse reinforcement learning. *Advances in neural information processing systems*, 23, 2010.

[24] Meredith Mealer, Ellen L. Burnham, Colleen J. Goode, Barbara Rothbaum, and Marc Moss. The prevalence and impact of post traumatic stress disorder and burnout syndrome in nurses. *Depression and anxiety*, 26(12): 1118–1126, 2009. ISSN 1520-6394 1091-4269. doi: 10.1002/da.20631. Place: United States.

[25] Wannes Meert, Kilian Hendrickx, Toon Van Craenendonck, Pieter Robberechts, Hendrik Blockeel, and Jesse Davis. Dtaidistance, 2020. URL https://github.com/wannesm/dtaidistance.

[26] Smitha Milli, Dylan Hadfield-Menell, Anca Dragan, and Stuart Russell. Should Robots be Obedient?, 2017. _eprint: 1705.09990.

[27] Nina Moorman, Erin Hedlund-Botti, Mariah Schrum, Manisha Natarajan, and Matthew C. Gombolay. Impacts of Robot Learning on User Attitude and Behavior. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '23, pages 534–543, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 978-1-4503-9964-7. doi: 10.1145/3568162.3576996. URL https://doi.org/10.1145/3568162.3576996. event-place: Stockholm, Sweden.

[28] Tatsuya Nomura, Tomohiro Suzuki, Takayuki Kanda, and Kensuke Kato. Measurement of negative attitudes toward robots. *Interaction Studies. Social Behaviour and Communication in Biological and Artificial Systems*, 7(3): 437–454, 2006. Publisher: John Benjamins Publishing Company Amsterdam/Philadephia.

[29] Edwin Olson. AprilTag: A robust and flexible visual fiducial system. *Proceedings - IEEE International Conference on Robotics and Automation*, pages 3400–3407, 2011. ISSN 10504729. doi: 10.1109/ICRA.2011.5979561. ISBN: 9781612843865.

[30] Harish Ravichandar, Athanasios S. Polydoros, Sonia Chernova, and Aude Billard. Recent Advances in Robot Learning from Demonstration. *Annual Review of Control, Robotics, and Autonomous Systems*, 3(1): 297–330, May 2020. ISSN 2573-5144. doi: 10.1146/annurev-control-100819-063206. URL https://doi.org/10.1146/annurev-control-100819-063206. Publisher: Annual Reviews.

[31] Stéphane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. No-Regret Reductions for Imitation Learning and Structured Prediction. In *14th Internation Conference on Artificial Intelligence and Statistics*, Fort Lauderdale, FL, 2011. URL https://arxiv.org/abs/1011.0686v3.

[32] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, May 2019. ISSN 2522-5839. doi: 10.1038/s42256-019-0048-x. URL https://doi.org/10.1038/s42256-019-0048-x.

[33] Fatai Sado, Chu Kiong Loo, Wei Shiung Liew, Matthias Kerzel, and Stefan Wermter. Explainable goal-driven agents and robots-a comprehensive review. *ACM Computing Surveys*, 55(10):1–41, 2023.

[34] Maha Salem, Gabriella Lakatos, Farshid Amirabdollahian, and Kerstin Dautenhahn. Would You Trust a (Faulty) Robot?: Effects of Error, Task Type and Personality on Human-Robot Cooperation and Trust. *ACM/IEEE International Conference on Human-Robot Interaction*, 2015-March:141–148, 2015. ISSN 21672148. doi: 10.1145/2696454.2696497. ISBN: 9781450328821 Publisher: ACM.

[35] Sebastian Schneider and Franz Kummert. Comparing robot and human guided personalization: Adaptive exercise robots are perceived as more competent and trustworthy. *International Journal of Social Robotics*, 13, 04 2021. doi: 10.1007/s12369-020-00629-w.

[36] Lakshmi Seelam, Erin Hedlund-Botti, Chuxuan Yang, and Matthew Gombolay. Interface Design for Learning from Demonstration with Older Adults. In *Association for the Advancement of Artificial Intelligence Fall Symposium Series*, 2023.

[37] Cory-Ann Smarr, Tracy L. Mitzner, Jenay M. Beer, Akanksha Prakash, Tiffany L. Chen, Charles C. Kemp, and Wendy A. Rogers. Domestic Robots for Older Adults: Attitudes, Preferences, and Potential. *International journal of social robotics*, 6(2):229–247, April 2014. ISSN 1875-4791 1875-4805. doi: 10.1007/s12369-013-0220-0.

[38] Arturo Daniel Sosa-Ceron, Hugo Gustavo Gonzalez-Hernandez, and Jorge Antonio Reyes-Avendaño. Learning from demonstrations in human–robot collaborative scenarios: A survey. *Robotics*, 11(6):126, 2022.

[39] Jonathan Spencer, Sanjiban Choudhury, Matt Barnes, Matthew Schmittle, Mung Chiang, Peter Ramadge, and Siddhartha Srinivasa. Learning from Interventions: Human-robot interaction as both explicit and implicit feedback. 2020. doi: 10.15607/RSS.2020.XVI.055.

[40] Adriana Tapus, Maja J. Mataric, and Brian Scassellati. Socially assistive robotics [Grand Challenges of Robotics]. *IEEE Robotics & Automation Magazine*, 14 (1):35–42, 2007. doi: 10.1109/MRA.2007.339605.

[41] Suzanne Tolmeijer, Astrid Weiss, Marc Hanheide, Felix Lindner, Thomas M Powers, Clare Dixon, and Myrthe L Tielman. Taxonomy of trust-relevant failures and mitigation strategies. In *Proceedings of the 2020 acm/ieee international conference on human-robot interaction*, pages 3–12, 2020.

[42] Garrett Wilson, Christopher Pereyda, Nisha Raghunath, Gabriel de la Cruz, Shivam Goel, Sepehr Nesaei, Bryan Minor, Maureen Schmitter-Edgecombe, Matthew E. Taylor, and Diane J. Cook. Robot-enabled support of daily activities in smart home environments. *Cognitive Systems Research*, 54:258–272, 2019. ISSN 1389-0417. doi: https://doi.org/10.1016/j.cogsys.2018.10.032.

[43] James Wright. *Robots won't save Japan: An ethnography of eldercare automation.* Cornell University Press, 2023.

[44] Tianyu Wu, Shizhu He, Jingping Liu, Siqi Sun, Kang Liu, Qing-Long Han, and Yang Tang. A brief overview of chatgpt: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica*, 10(5):1122–1136, 2023.

[45] Alexandra Zytek, Ignacio Arnaldo, Dongyu Liu, Laure Berti-Equille, and Kalyan Veeramachaneni. The need for interpretable features: motivation and taxonomy. *ACM SIGKDD Explorations Newsletter*, 24(1):1–13, 2022. Publisher: ACM New York, NY, USA.