Predicting Human Impressions of Robot Performance During Navigation Tasks

QIPING ZHANG*, Yale University, USA
NATHAN TSOI*, Yale University, USA
MOFEED NAGIB, Yale University, USA
BOOYEON CHOI, Yale University, USA
JIE TAN, Google DeepMind, USA
HAO-TIEN LEWIS CHIANG, Google DeepMind, USA
MARYNEL VÁZQUEZ, Yale University, USA

Human impressions of robot performance are often measured through surveys. As a more scalable and cost-effective alternative, we investigate the possibility of predicting people's impressions of robot behavior using non-verbal behavioral cues and machine learning techniques. To this end, we first contribute the SEAN TOGETHER Dataset consisting of observations of an interaction between a person and a mobile robot in a Virtual Reality simulation, together with impressions of robot performance provided by users on a 5-point scale. Second, we contribute analyses of how well humans and supervised learning techniques can predict perceived robot performance based on different observation types (like facial expression features, and features that describe the navigation behavior of the robot and pedestrians). Our results suggest that facial expressions alone provide useful information about human impressions of robot performance; but in the navigation scenarios that we considered, reasoning about spatial features in context is critical for the prediction task. Also, supervised learning techniques showed promise because they outperformed humans' predictions of robot performance in most cases. Further, when predicting robot performance as a binary classification task on unseen users' data, the F_1-Score of machine learning models more than doubled in comparison to predicting performance on a 5-point scale. This suggested that the models can have good generalization capabilities, although they are better at telling the directionality of robot performance than predicting exact performance ratings. Based on our findings in simulation, we conducted a real-world demonstration in which a mobile robot uses a machine learning model to predict how a human that follows it perceives it in a university campus. Finally, we discuss the implications of our results for implementing such supervised learning models in real-world navigation scenarios.

CCS Concepts: • **Computing methodologies → Learning from implicit feedback**; **Interactive simulation**; • **Human-centered computing → Social navigation**.

Additional Key Words and Phrases: implicit human feedback, human-robot interaction, social robot navigation, virtual reality

*Equally contributing authors

Authors' Contact Information: Qiping Zhang, Yale University, New Haven, USA, qiping.zhang@yale.edu; Nathan Tsoi, Yale University, New Haven, USA, nathan.tsoi@yale.edu; Mofeed Nagib, Yale University, New Haven, USA, mofeed.nagib@yale.edu; Booyeon Choi, Yale University, New Haven, USA, brian.choi@yale.edu; Jie Tan, Google DeepMind, Mountain View, USA, jietan@google.com; Hao-Tien Lewis Chiang, Google DeepMind, Boulder, USA, lewispro@google.com; Marynel Vázquez, Yale University, New Haven, USA, marynel.vazquez@yale.edu.

ACM XXXX-XXXX/2024/10-ART
https://doi.org/10.1145/nnnnnnn.nnnnnnn

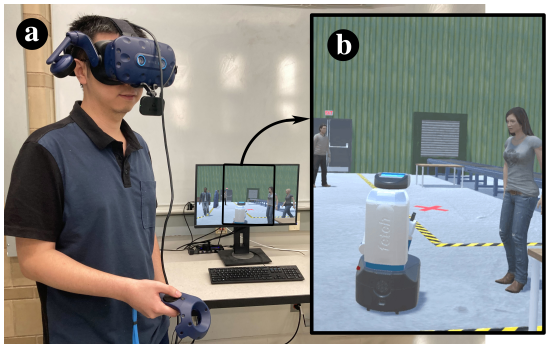

Fig. 1. Data collection. Humans controlled an avatar in the simulation with VR (a) while they were guided by a Fetch robot (b). The screen on the desk shows what the user saw.

1 INTRODUCTION

As a scalable alternative to measuring subjective impressions of robot performance through surveys, recent work in Human-Robot Interaction (HRI) has explored using *implicit* human feedback to predict these impressions [3, 22, 64, 80]. These are communicative signals that are unintentionally exhibited by people [41]. They can be reflected in human actions that change the world's physical state [60] or can be nonverbal cues, such as facial expressions [22, 64] and gaze [3, 54], displayed during social interactions. Implicit feedback serves as a burden-free information channel that sometimes persists even when people don't intend to communicate [40].

We expand the existing line of research on predicting impressions of robot performance from nonverbal human behavior to dynamic scenarios involving robot navigation. Prior work has often considered stationary tasks, like physical assembly at a desk [65] or robot photography [80], in laboratory environments. We instead explore the potential of using observations of the body motion, gaze, and facial expressions of a person to predict their impressions of a robot's performance while a robot guides them to a destination in a crowded environment. These impressions – which we also refer to as human perceptions in this paper – correspond to subjective opinions of how well a robot is performing the navigation task. Predicting them in crowded navigation scenarios is more challenging than in stationary settings because human nonverbal behavior can be a result of not only robot behavior, but also other interactants in the environment. Further, because of motion, nonverbal responses to the robot may change as a function of the environment. For example, imagine that the person that follows the robot looks downwards. This could reflect paying attention to the robot, or be a result of the person inspecting their nearby physical space, which varies during navigation.

To study implicit feedback during navigation tasks, we performed a systematic data collection using the Social Environment for Autonomous Navigation (SEAN) 2.0 [74] with Virtual Reality (VR) [81].[1] Humans took part in the simulations through an avatar, which was controlled using a VR headset, as in Fig. 1. The headset enabled immersion and allowed us to capture implicit feedback features like gaze. Also, it facilitated querying the human about robot performance as navigation tasks took place. We considered robot performance as a multi-dimensional construct, similar to [80], because humans may care about many aspects of a robot's navigation behavior, as discussed in the social robot navigation literature [25, 26, 51].

Then, we studied fundamental questions about the value of implicit feedback signals in predicting subjective impressions of robot performance using the VR data. First, we investigated to what extent humans can predict a person's impression of the robot's performance (along the dimensions of perceived competence, surprise, and intention) based on visualizations of observations of the human-robot interaction, as recorded in our VR

[1]Dataset available at: https://sean-together.interactive-machines.com/.

navigation dataset. Second, we investigated how well various supervised learning models do this type of inference in comparison to humans. Third, we studied the generalization capabilities of supervised learning methods to users unseen at training time.

Our analyses bring understanding to the complexity of predicting humans' impressions of robot performance in navigation tasks and enabled us to finally conduct a real-world demonstration in which a robot uses a machine learning model to predict how a human perceives it in a university campus. We conclude this paper by discussing the implications of our results for implementing autonomous systems that infer human perceptions of robot performance using implicit feedback in real-world navigation scenarios. We hope that our recommendations facilitate future efforts to make robots more aware of their failures during navigation [70], as well as facilitate aligning robot behavior to human preferences based on implicit feedback [18, 22, 52].

2 RELATED WORK

This section discusses prior work in relation to our contributions. First, we discuss human impressions of robot performance, especially in regards to robot motion. Then, we distinguish between explicit and implicit human feedback, the latter being the focus of our work. Finally, we briefly review data collection methodologies in HRI.

2.1 Impressions of Robot Performance

Understanding human impressions of robot performance is important. They can be used to evaluate robot policies [46, 56, 68] and to create better robot behavior [8, 21, 54, 69], increasing the likelihood of robot adoption. In this work, we focus on inferring three robot performance dimensions relevant to navigation [26]: robot competence, surprising behavior, and clear intent. Robot competence is a popular performance metric [16], especially in robot navigation [2, 50, 73]. Surprising behavior violates expectations. It is often considered undesired [4, 25] and may require explanations by the robot [12]. Meanwhile, showing clear intent means that the robot enables an observer to infer the goal of its motion [24]. Prior work suggests that if humans fail to anticipate the motion of a robot because it acts surprisingly or its intent is unclear, they will likely have trouble coordinating their own behavior with it [23, 61].

2.2 Implicit Human Feedback

We distinguish between explicit and implicit human feedback about robot performance. Explicit feedback corresponds to purposeful or deliberate information conveyed by humans to robots, e.g., through preferences [10, 67] or survey instruments [5, 50]. Meanwhile, implicit feedback are cues and signals that people exhibit without intending to communicate some specific information about robot performance, yet they can be used to infer such perceptions. Inferring performance from implicit feedback can reduce the chances of excessively querying users for explicit feedback in robot learning scenarios [29, 59], thereby minimizing the risk of feedback fatigue [45]. Learning from implicit feedback is not without challenges, however, as it can be difficult to interpret [22, 64]. For example, this can happen due to inter-person variability in facial expressions [30], similar signals being produced for different reasons [14], or signals changing over time as interactions progress [15].

Our work considers a variety of nonverbal implicit signals, including gaze, body motion, and facial expressions, which have long been studied in social signal processing [77]. While in some cases these signals are treated as explicit feedback (e.g., to interrupt an agent [79]), we consider them implicit feedback because we do not prime humans to react in specific ways to a robot. As such, our work is closer to [14, 22, 52, 63, 78], which used these signals to identify critical states during robot operation, detect robot errors, and adjust robot behavior.

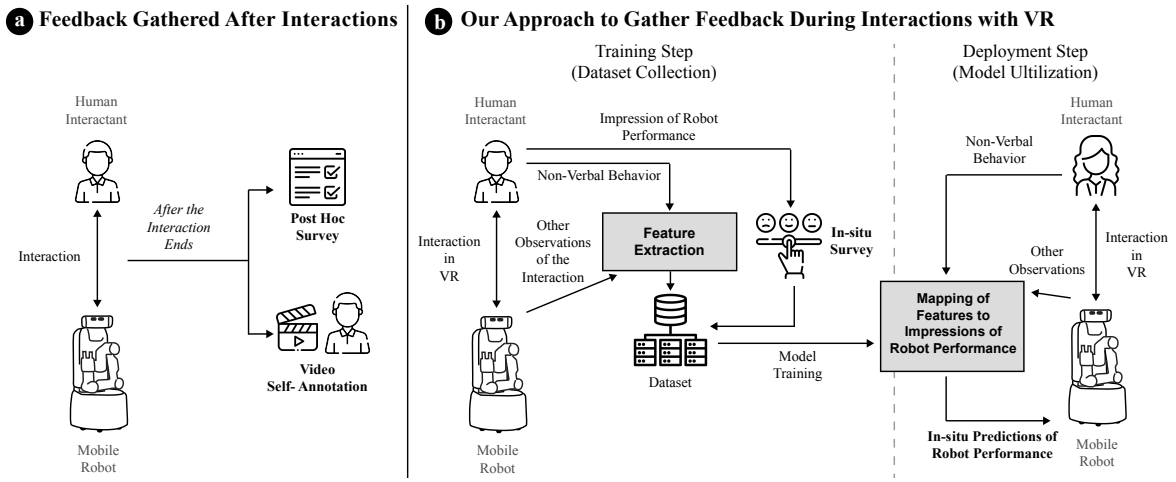

Fig. 2. a) It is typical to gather explicit human feedback about robot performance using surveys after human-robot interactions conclude because interruptions by the experimenters can easily bias human-robot social encounters. Unfortunately, the feedback from surveys tends to be very limited, making it difficult to understand robot performance at a granular level. Alternatively, participants may complete video annotations of their experiences [81], but this can be time consuming and taxing, especially in continuous navigation tasks. b) In this work, we first collect a dataset of human impressions of a robot's performance by prompting participants *during* interactions using VR (Training Step in the diagram). Then, we use this explicit feedback to train models that infer human impressions of robot performance based on observations of the interactions, especially including observations of human implicit feedback. The value of such a model is that once it is trained, it can be reused to estimate robot performance during new interactions (Deployment Step), without having to ask humans for explicit feedback as in the training step.

2.3 Data Collection in HRI: VR and Other Methodologies

Different kinds of HRI research methods have been used in the literature to gather interaction data, such as in-person user studies (e.g., [27, 50, 72]), observational public data collections (e.g., [39, 49]), crowdsourcing studies (e.g., [13, 37, 71]), etc. See [6] for an introduction to these methods.

We considered different ways of conducting our data collection, but ultimately opted for gathering data with simulated human-robot interactions in VR for several reasons. First, in contrast to real-world data collection, simulation facilitated querying humans about their impressions of robot performance during interactions and resulted in fewer negative consequences for interrupting the navigation task. This is illustrated in Fig. 2. In lab studies, for instance, surveys that gather general impressions of a robot are typically administered at the end of interactions to avoid interrupting the natural flow of events [80], which can cause unintended effects on collaborative tasks and interactants. In VR simulations, however, we can gather feedback in-situ. We can freeze time during human-robot interactions, query a participant about their impressions of robot performance through the VR display, and then resume the simulation as if the interruption had not occurred.

Second, we started our research by utilizing VR because, simulations made interactions safer in contrast to those in the real-world. The reason is that we wanted to expose participants not only to good robot navigation behavior, but also bad behavior. This was key for inducing a wide range of impressions about robot performance during data collection and, thus, capturing varied implicit feedback. Prior work has used simulations in HRI for safety reasons as well [35, 53].

Third, in contrast to crowdsourcing data collection procedures, our in-person data collection reduced unrelated participant distractions [11] and minimized potential issues with participant's internet speed [36, 73]. Early in our research, we considered using interactive surveys [73] for our data collection while capturing implicit feedback signals through the webcams of remote participants (e.g., as in [14]). However, after testing both this setup and VR, we thought that the increased level of immersion afforded by VR was important to gather naturalistic feedback.

While we opted for using simulations in our work, they are not without limitations. In particular, simulations can result in a sim-to-real gap, as discussed before in HRI and other robotics areas (e.g., [1, 9, 19, 20, 34, 44]). This gap can emerge in HRI because of differences in physics between simulation and the real-world as well as the human-robot interactions in simulation not reflecting the real-world experience [34]. Indeed, prior work suggests that virtual robots may be perceived as more discomforting than real robots [44]. Thus, towards the end of this paper, we explored applying the insights from our work with VR data to a real-world demonstration, paving the path towards predicting impressions of robot performance in real application scenarios.

3 PROBLEM STATEMENT & RESEARCH QUESTIONS

We study if a person's impression of a robot's performance can be predicted using observations of their interaction in dynamic tasks involving navigation. Specifically, we aim to learn a mapping from a sequence of observations to an individual's reported impressions at the end of the sequence (as in Fig. 2b). We consider multiple robot performance dimensions on a 5-point scale, as detailed later in Sec. 4.

Consider a dataset of observations and performance labels, $\mathcal{D} = \{(\mathbf{o}^i_{1:T}, y^i)\}$, where $\mathbf{o}_{1:T}$ is an observation sequence of length T, y is a performance rating given by a robot user at the end of the sequence, and i identifies a given data sample. We place emphasis on predicting a person's impression of a robot by considering observations of their implicit feedback. Thus, the observations $\mathbf{o}^i_t$ include features that describe the person's non-verbal behavior, such as their motion, gaze and facial expressions. Also, the observations include features that describe the spatial behavior of all the agents in the environment, the navigation task, and the space occupied by static objects. Given this data, we investigate three fundamental research questions:

(1) ***How well can human observers predict a user's impression of robot performance?*** By answering this question, we obtain a human baseline for learning a function $f : O_{1:T} \rightarrow \mathcal{Y}$, where O is the observation space and $\mathcal{Y}$ is performance. Also, through this question, we study the impact of two types of observations in the prediction task: observations that describe fine-grained facial expressions for a robot user; and other observations about the user, the robot and their environment. As mentioned earlier, observations of fine-grained expressions have gained popularity in recent work to infer human perceptions of an agent's behavior [14, 22, 65, 80]. Other observations (e.g. body motion and nearby static obstacles) can be more easily computed in real-world navigation tasks, but their usefulness on a robot's ability to infer users' impression of their performance is less understood.

(2) ***Can machine learning methods predict impressions of robot performance as well as humans?*** Ultimately, we are interested in bringing us forward to a future where machine learning models facilitate evaluating robot performance at scale, without having to necessarily ask users all the time for explicit feedback (as in the Deployment Step of Fig. 2b). Thus, we evaluate various machine learning models to approximate the function f, as defined for the prior question.

(3) ***How well can machine learning models generalize to unseen users?*** In future robot deployments, a robot may interact with completely new users. Thus, we analyze the performance of various machine learning models in predicting impressions of robot performance according to users for whom the model had no data at training time.

We study the above questions using data from SEAN-VR [81], as described in the next two sections. Later, in Sec. 6, we leverage our findings in VR to create a real-world demonstration through which we investigate predicting human impressions of robot performance in two university environments.

4 DATA COLLECTION WITH SEAN AND VR

For our VR data collection, we leveraged the SEAN 2.0 simulator [74]. SEAN 2.0 integrates with the Robot Operating System (ROS) [58] and supports Virtual Reality [81]. Participants used a Vive Pro Eye VR device to control an avatar in a warehouse (as in Fig. 1(a)). The VR headset captured implicit signals from the participants, like eye and lip movements.

During data collection, the participants had to follow a Fetch robot that guided them to a destination that was unknown to them a priori but was marked by a red cross on the ground. Fig. 1(b) shows a first-person view of the simulation during robot-guided navigation. The Fetch robot was controlled with ROS in SEAN. The environment contained other algorithmically controlled pedestrians and warehouse obstacles provided by SEAN 2.0.

The participants provided ratings of robot performance through the simulation's VR interface. The frame rate of the rendering of the virtual environment in the participants' first-person view in VR was over 30 frames per second. Our data collection protocol, described below, was approved by our local Institutional Review Board and refined via pilots.

4.1 Participants

We recruited 60 participants using flyers and by word of mouth. They were at least 18 years old, fluent in English, and had normal or corrected-to-normal vision. Overall, 19 participants identified as female, 40 as male, and 1 as non-binary or third gender. Most of them were university students, and ages ranged from 18 to 43 years old. Participants were somewhat familiar with robots, as indicated by a mean rating of M = 4.20 (with standard error SE = 0.18) on a 7-point Likert responding format (1 being lowest). Yet, they were somewhat unfamiliar with VR (M = 3.72, SE = 0.20). No participant had prior experience with SEAN or social robot navigation in VR.

4.2 Data Collection Procedure

Protocol: A data collection session took place as follows. First, the participant provided demographics data. Second, the experimenter introduced the robot, explained the navigation task in which the participant was to follow the robot, and demonstrated how to use the VR device to control their avatar in SEAN and label robot performance. Third, the participant experienced four navigation tasks with the robot, each with a particular starting position and destination. For consistency, the pedestrians were controlled using the same behavior graph controller provided in SEAN 2.0 [74] and the robot used the same navigation logic across the tasks.

In each task, the robot guided the participant to the destination and repeatedly changed its behavior (as further detailed below). Importantly, the interaction was paused before and after each behavior change took place, at which point the participant was asked to evaluate the robot's most recent navigation performance. A typical data collection session was completed in 45 min to 1 hour. Participants were compensated US$15 for their time.

Robot Behaviors: During a navigation task, the robot switched between one of these three types of behavior:

1. Nav-Stack. The robot navigated efficiently to the destination based on the path planned by the ROS Navigation Stack with social costs [47]. The planned paths generally minimized navigation time while avoiding collisions and invading personal space. This behavior lasted 40 seconds.

2. Spinning. The robot rotated at its current position, which we expected to be perceived as if the robot was confused. This behavior lasted 20 seconds. It was implemented by sending angular velocity commands to the robot's motion controller.

3. Wrong-Way. The robot moved in the wrong direction, away from the task's destination, effectively making a mistake during navigation. This behavior lasted 20 seconds and was implemented using the Navigation Stack with social costs as well, but with an incorrect navigation goal.

Unbeknownst to the participants, the robot switched to *Nav-Stack* behavior after *Spinning* or *Wrong-Way* during navigation. It randomly switched to *Spinning* or *Wrong-Way* after finishing *Nav-Stack*. The design was intended to maintain a consistent rate of sub-optimal behavior and avoid user boredom or significant confusion, which can be caused by more stochastic behavior patterns that are hard for participants to reason about. We expected the behaviors to elicit both positive and negative views of the robot, leading to a large variety of non-verbal reactions and impressions of robot performance.

Impressions of Robot Performance: During a navigation task, we paused the interaction at 4 seconds *before*, and at 8 seconds *after* the robot switched between behaviors. The elapsed time for the latter pause was longer in order to give people enough time to experience the latest robot behavior.

As shown in the supplementary video, impressions of robot performance were provided through an interface embedded in the simulation. The interface asked the participants to indicate their impression about the robot's most recent performance in regard to: 1) *"how competent was the robot at navigating,"* 2) *"how surprising was the robot's navigation behavior,"* and 3) *"how clear were the robot's intentions during navigation."* Participants provided ratings for these three dimensions of robot performance on a 5-point Likert responding format, e.g., with 1 being "incompetent", 2 being "somewhat incompetent", 3 being "neither competent nor incompetent", 4 being "somewhat competent", and 5 being "competent".

4.3 Observations

We organized observations of human-robot interactions, as recorded in SEAN-VR [81], into the features described below. More details about these features are provided in the Appendix.

Participants' Facial Expression Features: We captured the participants' eye and lip movements, as well as their gaze through the VR headset using the VIVE Eye and Facial Tracking (SRanipal) SDK. The eye and lip movements corresponded to 73 features that described the geometry of the face through blend shapes. The gaze was a 3D vector providing the direction of gaze of the person relative to their face.

Spatial Behavior Features: During navigation, we captured the poses of the robot, the participant, and the other automatically-controlled avatars on the ground plane of the scene. Then, we computed the poses of the avatars relative to the robot, considering only those within a 7.2m radius, as this region is typically considered a robot's public space [31, 38, 62]. Each of the features were (x, y, θ) tuples with x, y being the position and θ the body orientation (yaw angle) relative to a coordinate frame attached to the robot.

Goal Features: A navigation task had an associated destination or goal that the robot had to reach. We converted the goal pose in a global frame in the warehouse to a pose in a coordinate frame attached to the robot. This pose described the robot's proximity and relative orientation to its destination.

Occupancy Features: During navigation, the robot localized [28] against a 2-Dimensional (2D) map of the warehouse. We used a cropped section of the map around the robot (of 7.2m $\times$ 7.2m) to describe the occupancy of nearby space by static objects.

4.4 Perceived Robot Performance

Impressions of robot performance were as expected: ratings for competence and clear intention were generally higher for *Nav-Stack* than for *Spinning* and *Wrong-Way*, while the latter two tended to be more surprising than the former. Pairs of performance dimensions were significantly correlated with absolute Pearson r-values greater

than 0.6. An exploratory factor analysis suggested that the dimensions could be combined into one performance factor (which explained 77% of the variance).

Using the features described before and the impressions of robot performance provided by the participants, we created a dataset of paired observation sequences and target performance values. We further refer to this data as the SEAN virTual rObot GuidE with impliciT Human fEedback and peRformance Dataset (SEAN TOGETHER Dataset). As described below, we used this dataset to investigate the research questions in Sec. 3.

5 FINDINGS

5.1 How Well Can Human Observers Predict a User's Impression of Robot Performance?

To better understand the complexity of inferring impressions of robot performance, we evaluated how well human annotators could solve the prediction problem. To this end, we administered an online survey through `www.prolific.co`, a platform for human data collection and online research studies. In our survey, human annotators observed visualizations of observations in our SEAN TOGETHER Dataset. Then, they tried to predict performance ratings provided by the people who followed the robot.

Method: For the survey, we randomly selected 2 data samples from each of the 60 participants in our data collection, with one gathered before and the other gathered after the robot's behavior changed. The observations in each sample corresponded to an 8-second 5-hz window of features right before the corresponding performance label was provided.

As shown in Fig. 3, data samples were visualized in two ways:

1. Facial Rendering. We created a human face rendering in Unity by replaying the facial expression features on an SRanipal compatible avatar, as shown in Fig. 3 (right). This visualization was motivated by the use of facial expressions in prior work on implicit feedback (e.g., [22]).

2. Navigation Rendering. We created a plot of features that described the navigation behavior of the robot and the avatars in the simulation. The plot showed features that, using existing perception techniques, may be easier to estimate than facial features in real-world deployments. These features are the spatial behavior features, the robot's goal location, the occupied space near the robot, and the gaze direction of the participant – the last of which could be approximated using an estimate of the person's head orientation [55]. Because prior work suggests that it is easier to make sense of implicit human feedback in context [14], the plot was always centered on the robot, making its surroundings always visible as in Fig. 3 (left).

We used the visualizations to create three annotation conditions that helped understand the value of different features: 1) ***Nav.-Only***: annotators only saw the navigation rendering (e.g., as in the left image of Fig. 4); 2) ***Facial-Only***: for a given data sample, annotators only saw the facial rendering (e.g., as in the right image of Fig. 4); and 3) ***Nav.+Facial***: annotators saw the navigation rendering first, then the facial rendering and, finally, saw a video with both visualizations next to each other (as in Fig. 3).

Each of the data samples was annotated by 10 unique people in each condition. The annotators were instructed to predict how the participant who controlled the avatar that followed the robot perceived the robot's performance. The samples they annotated were presented in random order. Each annotator was paid US$7.5 for approximately 30 min of annotation time. To encourage high-quality annotations, we also gave them a bonus of US$0.125 for each correct prediction that they made.

Annotators: We recruited a total of 100 annotators. Thirty-five of them identified as female, 60 as male, and 5 as non-binary or third gender. Ages ranged from 18 to 75 years old. Annotators indicated similar familiarity with robots (M = 4.12, SE = 0.14) as the data collection participants, though the annotators were slightly more familiar with VR (M = 4.50, SE = 0.16). See the Appendix for details on annotator reliability.

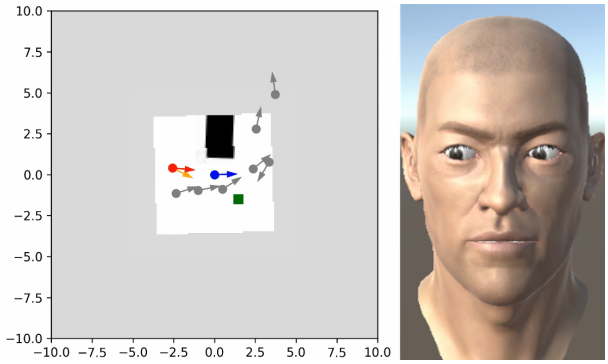

Fig. 3. A data sample from the *Nav.+Facial* condition. The **left** plot shows gaze, spatial behavior, goal, and occupancy features: ●→ is the robot's pose; ●→ is the pose of the participant following the robot during the VR interaction; → indicates the gaze of the participant; ●→ are the poses of algorithmically controlled avatars; ■ is the destination position that the robot navigated towards; and occupancy in the environment is indicated by black pixels (occupied) and white pixels (unoccupied). The **right** visualization shows a rendering of the facial expression features of the participant.

Results: We used linear mixed models estimated with REstricted Maximum Likelihood (REML) [32, 66] to analyze errors in the predictions for each performance dimension. Our independent variables were Before/After Robot Behavior Change (*Before*, *After*) and Annotation Condition (*Facial-Only*, *Nav.-Only*, *Nav.+Facial*). Also, we considered Annotator ID as a random effect because annotators provided predictions for multiple data samples. Our dependent variables were the absolute error between an annotator's prediction and the performance rating in our SEAN TOGETHER Dataset.

We found that the Annotation Condition had a significant effect on the absolute error for Competence, Surprise, and Intention ($p < 0.0001$ in all cases). As in Fig. 5 (left), Tukey HSD post-hoc tests showed that for Competence and Surprise, the errors for *Nav.+Facial* and *Nav.-Only* were significantly lower than *Facial-Only*, yet the difference between the former two conditions was not significant. For Intention, all conditions led to significantly different errors. *Nav.+Facial* resulted in the lowest error, followed by *Nav.-Only* and then *Facial-Only*. These results suggest that facial expressions provide information about impressions of robot performance though, more generally, the features used to create the Navigation Renderings seemed to be the most critical for these predictions.

Before/After Robot Behavior Change had a significant effect on the prediction errors for Competence and Intention ($p < 0.0001$ in both cases). As in Fig. 5 (right), the error was significantly lower for samples *Before* a behavior change than for samples *After* a change for these performance dimensions. We suspect this was because the robot sometimes demonstrated two behaviors in the samples collected *After* a behavior change, but in the case of *Before* behavior change, the robot only showed one behavior making these data samples more consistent and easier to reason about.

Table 1 shows the F_1-Scores for the annotator predictions (see HA rows). The low F_1 scores suggest that correctly predicting impressions of robot performance on a 5-point responding format was difficult for humans. Despite this, we suspected that humans could do a more reasonable job at distinguishing impressions of poor robot performance from other impressions. If this was the case, then this could open up doors in the future to using this binary signal (instead of the more fine-grained feedback) as a reward signal to adapt robot behavior in navigation tasks, e.g., in line with [42, 48]. Thus, we transformed the ground truth ratings from our data collection to binary values, one corresponding to low performance (e.g., 1-2 ratings for competence) and another to medium-to-high performance (3-5 ratings for competence). Also, we transformed the annotators' predictions similarly. This led to F_1 scores of 0.69 for Competence, 0.64 for Surprise, and 0.69 for Intention. As expected,

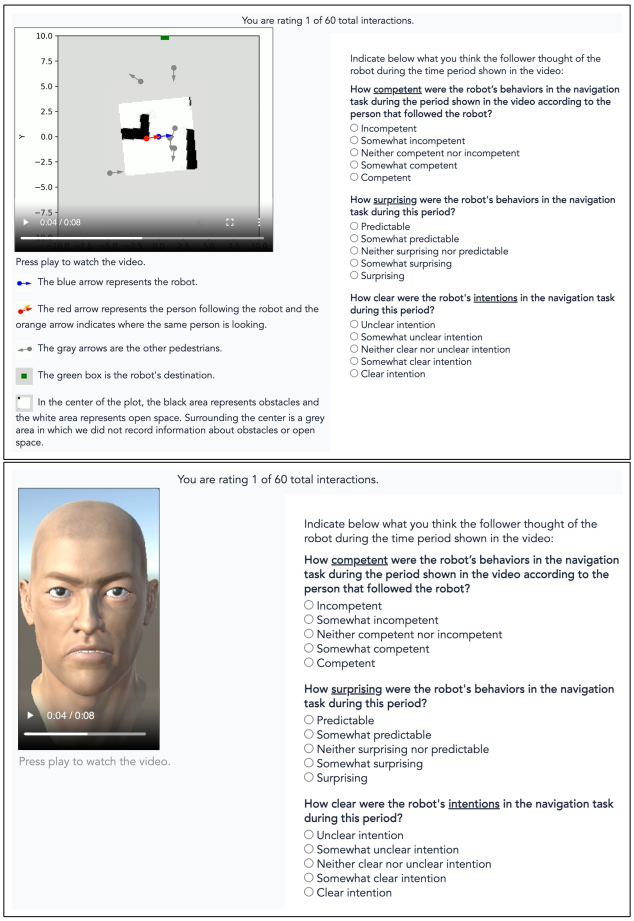

Fig. 4. Layout of the interfaces used for video annotation for the human baseline. *Left:* Layout used for the *Nav.-Only* annotation condition, showing the navigation rendering on the left, and questions on the right. *Right:* Layout for the *Facial.-Only* condition.

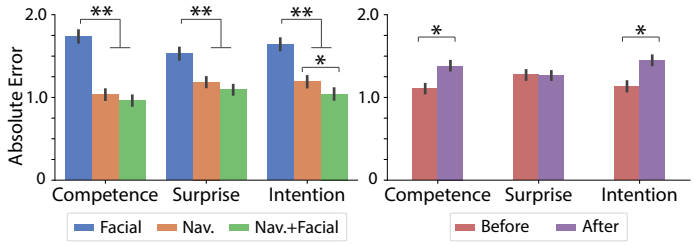

Fig. 5. Errors for annotators' predictions by Annotation Conditions (*left*) and Before/After Robot Behavior Change (*right*). (**) and (*) denote $p < 0.0001$ and $p < 0.05$, respectively.

Table 1. Machine learning methods and human annotation (HA) performance on 120 examples. Methods: Random (R) sampling from the distribution of labels in the training set, Random Forest (RF), Multi-Layer Perceptron (MLP), Graph Neural Network (GNN), and Transformer (T). Arrows indicate that higher (↑) and lower (↓) results are better. Cells with (-) do not have results because a GNN trained on facial features only was effectively an MLP. The Best and Second results are highlighted.

		F_1-Score ($\mu \pm \sigma$) ↑			Accuracy ($\mu \pm \sigma$) ↑			Mean Absolute Error ($\mu \pm \sigma$) ↓		
		Facial	Nav.	Nav.+Facial	Facial	Nav.	Nav.+Facial	Facial	Nav.	Nav.+Facial
Competence	HA	0.16 ± 0.0	0.28 ± 0.1	0.29 ± 0.2	0.19 ± 0.1	0.40 ± 0.1	0.42 ± 0.1	1.74 ± 0.2	1.03 ± 0.3	0.99 ± 0.4
	R	0.18 ± 0.0	0.19 ± 0.0	0.17 ± 0.0	0.21 ± 0.0	0.21 ± 0.0	0.20 ± 0.0	1.73 ± 0.1	1.75 ± 0.1	1.81 ± 0.1
	RF	0.19 ± 0.0	0.37 ± 0.0	0.38 ± 0.0	0.33 ± 0.0	0.52 ± 0.0	0.52 ± 0.0	1.43 ± 0.0	0.88 ± 0.0	0.82 ± 0.0
	MLP	0.23 ± 0.0	0.29 ± 0.1	0.25 ± 0.1	0.28 ± 0.0	0.48 ± 0.0	0.44 ± 0.1	1.66 ± 0.1	1.07 ± 0.3	1.19 ± 0.4
	GNN	-	0.31 ± 0.1	0.33 ± 0.0	-	0.43 ± 0.1	0.39 ± 0.1	-	1.22 ± 0.3	1.04 ± 0.0
	T	0.21 ± 0.0	0.33 ± 0.0	0.33 ± 0.0	0.30 ± 0.0	0.43 ± 0.0	0.41 ± 0.1	1.58 ± 0.1	0.97 ± 0.0	0.95 ± 0.0
Surprise	HA	0.18 ± 0.0	0.24 ± 0.1	0.25 ± 0.1	0.20 ± 0.1	0.30 ± 0.1	0.32 ± 0.1	1.53 ± 0.3	1.19 ± 0.2	1.12 ± 0.2
	R	0.19 ± 0.0	0.21 ± 0.0	0.17 ± 0.0	0.20 ± 0.0	0.21 ± 0.0	0.18 ± 0.0	1.64 ± 0.1	1.60 ± 0.1	1.68 ± 0.1
	RF	0.29 ± 0.0	0.38 ± 0.0	0.34 ± 0.0	0.30 ± 0.0	0.40 ± 0.0	0.34 ± 0.0	1.30 ± 0.0	0.93 ± 0.0	0.98 ± 0.0
	MLP	0.24 ± 0.0	0.26 ± 0.1	0.24 ± 0.1	0.25 ± 0.0	0.30 ± 0.0	0.29 ± 0.1	1.23 ± 0.1	1.12 ± 0.2	1.08 ± 0.1
	GNN	-	0.29 ± 0.0	0.27 ± 0.0	-	0.30 ± 0.0	0.28 ± 0.0	-	1.13 ± 0.1	1.07 ± 0.1
	T	0.27 ± 0.0	0.29 ± 0.0	0.32 ± 0.1	0.28 ± 0.0	0.31 ± 0.0	0.33 ± 0.1	1.37 ± 0.1	1.07 ± 0.1	1.04 ± 0.1
Intention	HA	0.18 ± 0.0	0.25 ± 0.1	0.28 ± 0.1	0.21 ± 0.1	0.37 ± 0.2	0.41 ± 0.1	1.64 ± 0.2	1.19 ± 0.4	1.07 ± 0.2
	R	0.21 ± 0.1	0.19 ± 0.0	0.17 ± 0.0	0.23 ± 0.1	0.22 ± 0.0	0.19 ± 0.0	1.70 ± 0.1	1.73 ± 0.1	1.80 ± 0.1
	RF	0.28 ± 0.0	0.28 ± 0.0	0.24 ± 0.0	0.37 ± 0.0	0.43 ± 0.0	0.41 ± 0.0	1.45 ± 0.0	1.13 ± 0.0	1.14 ± 0.0
	MLP	0.27 ± 0.0	0.26 ± 0.1	0.22 ± 0.0	0.31 ± 0.0	0.41 ± 0.1	0.39 ± 0.1	1.86 ± 0.1	1.31 ± 0.3	1.51 ± 0.5
	GNN	-	0.28 ± 0.0	0.29 ± 0.0	-	0.37 ± 0.0	0.35 ± 0.0	-	1.32 ± 0.1	1.25 ± 0.1
	T	0.24 ± 0.0	0.29 ± 0.1	0.32 ± 0.0	0.33 ± 0.0	0.41 ± 0.0	0.40 ± 0.0	1.63 ± 0.1	1.21 ± 0.1	1.20 ± 0.1

human annotators were better at telling the directionality of robot performance ratings than at predicting their exact magnitude.

Finally, we investigated the performance of human annotations over the span of data collection because prior work suggests that the expressiveness of people engaged in human-robot interactions can change over time [15], e.g., potentially due to changes in their expectations about the robot or due to fatigue. Figures 6(a)–(c) show the evolution of mean absolute errors for the human annotators' predictions over 10-minute intervals of interaction, considering each performance dimension. In general, human performance was very stable, suggesting no major bias over time in participant's spatial behavior or facial expressions. Interestingly, the results also suggested that improvements in performance with an individual feature did not necessarily translate in improvements on the *Nav.+Facial* condition. Humans may have combined the information from the different implicit feedback modalities in subtle ways when making their predictions about how participants in VR perceived the robot.

5.2 Can Machine Learning Methods Predict Impressions of Robot Performance as Well as Humans?

We compared human prediction performance with a variety of classifiers, including a random forest and neural networks.

Method: Machine learning (ML) models were evaluated on the same samples shown to the human annotators ($n = 120$). The rest of the data was used for training ($n = 2280$) and validation ($n = 569$). We trained one model for each combination of feature sets shown to the human annotators (*Facial-Only*, *Nav.-Only*, and *Nav.+Facial*). The *Nav.* feature set included occupied space near the robot, which we encoded using a ResNet-18 representation

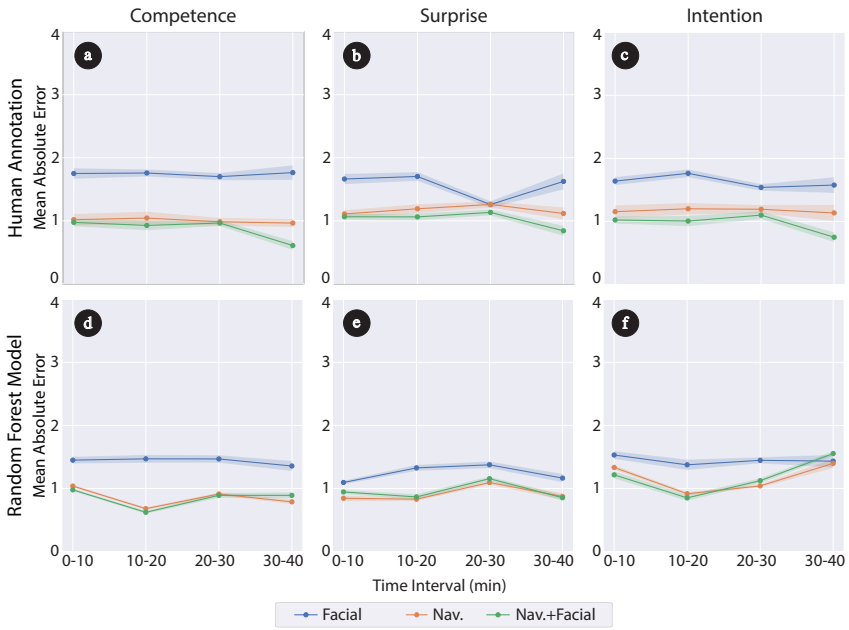

Fig. 6. Mean Absolute Errors (MAE) of human annotation and Random Forest (RF) results over 10-minute intervals of the data collection sessions. MAE was computed for all data samples in each interval, and then the average and standard errors of MAE were calculated considering the performance of the 10 unique annotators (for human annotation results in (a)–(c)) or the 10 Random Forest models trained with different seeds in Table 1 (RF results in (d)–(f)).

[33]. We repeated training for each model 10 times with varying random seeds. The Random Forest (RF) used 100 trees and the depth was grown until leaves had less than 2 samples. The neural networks had a number of parameters on the same order of magnitude: 5.4×10^6 for a Multi-Layer Perceptron (MLP), 2.1×10^6 for a message-passing Graph Neural Network (GNN) [7], and 6.5×10^6 for a Transformer (T) [75]. Networks were trained using minibatch gradient descent with the Adam optimizer and cross-entropy loss. Learning rate, batch size, and dropout were chosen using grid search with validation-based early stopping [57]. We also compared all these models with a random sampling baseline.

Results: As is shown in Table 1, ML models outperformed both human-level performance and random baseline in all cases when measured via F_1-Score. When measured using Accuracy and Mean Absolute Error, ML models performed the best, except for Intention when using *Nav.+Facial* features. These outcomes indicate that our implicit feedback data contained useful information that can be leveraged by ML models to predict users' impressions of robot performance. Further, ML models trained with *Nav.-Only* and *Nav.+Facial* features outperformed those trained with *Facial-Only* features. This finding aligns with our observation in Sec. 5.1 on the criticality of the *Nav.* features in comparison to the *Facial* features on performance prediction.

Figures 6(d)–(f) show the evolution of mean absolute errors for the Random Forest model, which generally performed the best, over 10-minute intervals of interaction during the data collection. Similar to the results from human annotators (Figures 6(a)–(c), Sec. 5.1), the error for the RF model did not fluctuate drastically, although the performance for Intention prediction with *Nav.* and *Nav.+Facial* features decreased in the last two time intervals of data collection (having higher mean absolute error). The decrease in performance could be the result of a distribution shift, especially in the last interval which had the fewest number of samples because not all

interactions took the full 40 minutes. Also, a good proportion of the samples in the last time interval showed the end of navigation tasks, at which point the participants could have been more sensitive to robot navigation in the wrong direction. Indeed, there was a higher proportion of lower ratings for Intention in the last interval than in the other intervals, as shown in the Appendix.

To better understand differences in the prediction performance between ML and human annotators, we first identified the examples annotated by humans for which there was a difference greater than 1 in Mean Absolute Error between human annotators and the RF model that tended to perform best. Then, we inspected the 8-second navigation renderings of these data examples, as in Fig. 4 (left). Among examples where the RF model performed better than humans, 64% exhibited a major behavior pattern for the robot that persisted despite minor deviations. For example, the robot navigated effectively to the goal most of the time, but was occasionally blocked and had to move around the obstacles. We hypothesize that ML did better in these cases because machine learning can leverage regularities in the data when making predictions without potentially getting distracted with the minor deviations. Among the examples where human annotators performed better, 68% showed the robot exhibiting more than one behavior (*Nav-Stack*, *Spinning*, or *Wrong-Way*) or the interaction involved unconventional reactions from humans, such as people interfering in the navigation task. We suspect that humans were better in these cases because they can leverage their prior knowledge about the world to better reason about uncommon variations in the data. For the RF, uncommon observations can be out-of-distribution samples that result in more prediction errors, especially considering the limited size of our dataset.

Taken together, these results motivated us to focus the analysis in the next section on the aggregate, overall results rather than the interval-based results.

5.3 Can Machine Learning Generalize to Unseen Users?

We investigated how well learning models could predict performance by a user whose data was held out from training.

Method: We used the models and training scheme from Sec. 5.2 with all features (*Nav.+Facial*), but split the data using leave-one-out cross-validation. For each fold, the data for one participant was used as the test set and the remaining examples were split between training (80%) and validation (20%). We searched for new hyperparameters and computed results both on 5-classes and on binary classification. Binary targets and prediction labels were computed as in Sec. 5.1.

Results: Fig. 7 reports F_1-Scores over all folds. The models generalized to unseen people with only a slight reduction in performance in comparison to Table 1. Also, the average F_1-Score across all performance dimensions improves from 0.25 in the multiclass case to 0.62 in the binary case. This makes the ML predictions more usable in practice. For example, in the future, we envision deploying the trained ML on new users (as in Fig. 2b) in order to detect low robot performance. This could be an indication that the robot made a mistake, triggering interaction recovery behaviors like apologies or explanations [70], which could increase trust on the system [17].

6 REAL-WORLD DEMONSTRATION

To investigate whether we could predict human impressions of robot performance in other, more realistic scenarios than those observed in our VR data collection, we conducted a real-world demonstration with a modified Pioneer 3-DX mobile base. More specifically, we conducted a data collection with the mobile robot in two semi-public indoor environments of Yale's University, and analyzed how well a random forest model could predict human impressions of robot performance in the real-world setup. This real-world data collection, as further described below, was approved by our local Institutional Review Board.

The system that we built for real-world data collection was designed in consideration of: 1) we wanted to induce naturalistic interactions between the robot and pedestrians; and 2) we wanted to support the same data

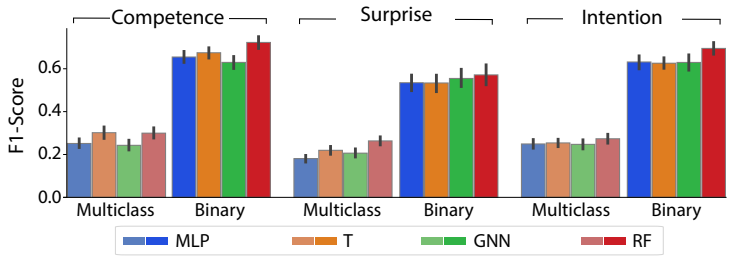

Fig. 7. ML models trained on *Nav.+Facial* features using leave-one-out cross-validation and evaluated on the held-out participant's data. F_1-Scores are computed over 5 classes (Multiclass) and 2 classes (Binary). Error bars represent the standard errors calculated from the F_1-Scores per leave-one-out fold. See the text for details.

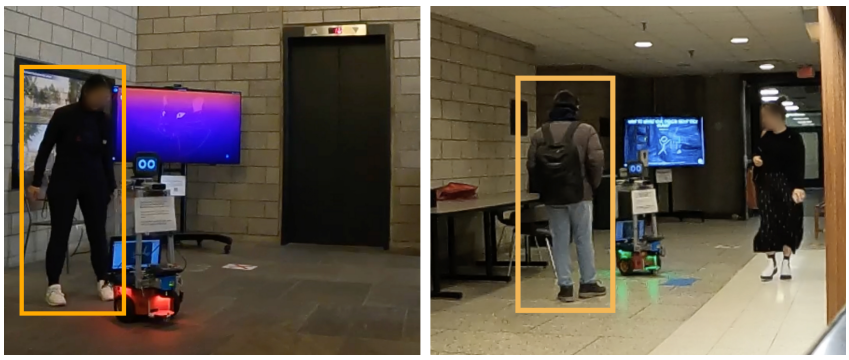

Fig. 8. Real-world data collection in two indoor spaces of Yale University. The orange box highlights the follower, i.e., the person that followed the robot during navigation tasks. Other people could pass by the follower and the robot as in the *right* image during data collection. The robot had lights to indicate when it was navigating (green, *right* image) or had paused navigation (red, *left* image).

collection protocol used with SEAN, as in Section 4.2. Therefore, we did not recruit participants prior to the data collection. Instead, we operated the robot and, as pedestrians walked nearby, we asked them if they would be willing to follow the robot for a short period and answer brief surveys. In total, 45 pedestrians agreed to follow the robot for this demonstration.

Mobile Robot: The Pioneer 3-DX robot is a differential-drive mobile base and, thus, it moves in a similar way to the Fetch robot used in our VR data collection. We added to the Pioneer robot lights that illuminated green to indicate that it was navigating towards a location, and red to indicate that it had paused navigation. Over the Pioneer base, we built a frame that held a robotic screen face (similar to [43, 76]) on the very top of the robot, which allowed to easily distinguish the front of the platform. The frame also held two Kinect Azure RGB-D cameras right below the robot head. Each camera had a 120-degree field of view. One was pointed forward and the other was pointed backwards, which allowed the robot to track people in front and behind it using the Kinect SDK. Additionally, the bottom section of the frame held a 2D LMS-100 Sick LiDAR and a gaming laptop with an Intel Core i7-8750H CPU, 32 GiB of RAM, and an Nvidia GeForce GTX 1070 GPU. The laptop ran the Robot Operating System to control the robot using the ROS navigation stack [58] with social cost layers [47], which enabled the robot to avoid collisions with nearby people. Fig. 8 and our supplementary video show the robot in this demonstration effort.

Table 2. F_1-Score ($\mu \pm \sigma$) for Random Forest models trained using *Nav.-Only* features from either the *Real*-world data, or *VR* data considering the nearest 5 people to the robot (as explained in Sec. 6) Results include multi-class classification based on the 5-point Likert responses (*Multi-cls*) and binary classification (*Binary*). Column 1 corresponds to training on VR data and evaluating on VR data (VR→VR), Column 2 corresponds to training on VR data and evaluating on real data (VR→Real), and Column 3 is training and evaluating on real data (Real→Real).

		(1) VR→VR	(2) VR→Real	(3) Real→Real
Multi-cls	Competence	0.30 ± 0.09	0.21 ± 0.18	0.27 ± 0.35
	Surprise	0.27 ± 0.08	0.26 ± 0.21	0.26 ± 0.27
	Intention	0.26 ± 0.08	0.20 ± 0.28	0.24 ± 0.34
Binary	Competence	0.69 ± 0.10	0.56 ± 0.41	0.61 ± 0.34
	Surprise	0.59 ± 0.18	0.58 ± 0.36	0.58 ± 0.33
	Intention	0.65 ± 0.08	0.55 ± 0.40	0.60 ± 0.40

Demonstration Protocol: We waited for pedestrians to walk by the robot in two locations on a university campus. One location was a subterranean pedestrian tunnel or concourse; the other one was an L-shaped entrance corridor to a building. When pedestrians passed by, we asked them if they would be interested in following the robot as it navigated to a nearby goal marked by a red cross on the ground. For those that agreed, we instructed them that the robot would navigate when it showed a green light. After short intervals of time, it would pause navigation, showing a red light, and they would be asked a few quick questions about their impressions of the action that the robot just performed using a mobile device. The device showed the same questions about robot competence, surprising behavior and clear intent (on a 5-point Likert responding format) as in our VR data collection. Also, the robot navigation behaviors and the timing of questions about robot performance matched those in Sec. 4.2.

Data: We focused on capturing *Nav.-Only* features (that described the navigation behavior of the robot and humans, as in Sec. 5.1) for two reasons. First, our prior results with VR data suggested facial expression features were not as critical to make predictions over human impressions of robot performance than the other features. Second, facial expressions were often occluded, providing no information to the robot. In total, we collected 235 examples from this real world demonstration, each consisting of *Nav.-Only* features and associated survey responses.[2]

ML Models: Our primary aim was to understand the applicability of our approach to infer impressions of robot performance in the real world. However, there were important differences in our VR and real-world data collection setups as a result of real-world constraints. For example, the real robot had a more limited field of view compared to the simulation where the ground truth motion for all people in the environment was available. Moreover, the real-world environments were less densely populated than simulation.

Therefore, to fairly compare our results across simulation and the real world, we trained two types of Random Forest classifiers, given that the RF model generally performed best in Table 1. One type of model was trained using VR data but we limited the field of view of the robot to 120-degrees forward and backward as well as the maximum number of nearby people input to the model to five individuals. The other type of Random Forest model (with the same parameters) was trained using real-world data. Both types of models were trained considering 5-classes, with binary targets and prediction tables being computed as in Sec. 5.1.

[2]The real-world data that we collected from this demonstration is available at: https://sean-together.interactive-machines.com/.

Results: Table 2 shows the F_1-Score of models evaluated on the same type of data they were trained on (Sim or Real). For these results, we used leave-one-person-out cross-validation to train and evaluate generalization to new robot followers. That is, data from one person was held out for each fold. Also, Table 2 shows the performance of the model trained in simulation on real-world data. In this case, a RF model was trained using all the VR data from the VR→VR case, and then evaluated on the test set for the leave-one-person-out folds for the real-world data. As one would naturally expect based on our prior results with VR data, binary classification resulted in higher performance than multi-class classification in all these cases.

In general, performance was higher for models trained and evaluated in simulation (Column 1), which could be the result of having more VR data than real-world data. The results for models trained and evaluated on real data (Column 3) were close to those that considered simulation data only (Column 1). This suggested that our methodology to collect real-world data and the RF model are promising for inferring impressions of robot performance in the real world. Finally, reasonable performance was obtained for the model that was trained with VR data and tested on real-world data (Column 2). This highlights the potential of sim-to-real transfer of machine learning models trained on spatial features as well as the potential of using our VR data to build computational models that predict human perceptions of robot performance in real-world interactions.

7 IMPLICATIONS FOR REAL-WORLD APPLICATIONS

We hope that future work leverages our findings to build effective models for mapping implicit human feedback to users' impressions of robot performance in real-world social navigation tasks. To this end, we first recommend prioritizing robust people tracking and pose estimation over computing fine-grained facial expressions, especially when computational resources may be limited. Reasoning about spatial behavior features in the context of the task can facilitate achieving reasonable prediction performance with lower sensor requirements. Also, occlusions are likely more common for facial expressions than body tracking, as we observed in our real-world demonstration.

Second, it is important to consider the granularity of the predictions over impressions of robot performance. We began our work gathering impressions of robot performance on a 5-point Likert responding format, which we believed could reveal subtle aspects of human perceptions during navigation. However, we found that predicting impressions of robot performance over 5 classes was challenging for both humans and ML models. While human prediction performance could have been affected by specific details of the visualizations that we used to gather our human baseline results, it is worth considering less granular feedback to favor prediction performance during robot deployments. In particular, for more practical usage of human feedback, we recommend building models that start by identifying poor robot performance (performing binary classification) and then, on top of that, try to predict more granular impressions of robot performance.

Finally, if a robot is executing multiple behaviors, we recommend considering whether the robot switched behaviors recently when reasoning about performance predictions. As in our results, predicting performance recently after a behavior change can be more difficult than before, when the behavior was more consistent.

8 LIMITATIONS AND FUTURE WORK

Our work has several limitations that point to interesting future directions. In particular, we obtained human baselines for prediction performance, but used only a limited set of feature combinations that described interactions in a single VR environment and two real-world environments. In the future, it would be interesting to consider a broader set of feature categories in a more diverse range of environments. For instance, future work could investigate the value of more detailed human pose features (e.g., [82]) across a wider range of scenarios (public plazas or hospitals) where humans may behave differently due to their activity, stress or other factors.

Facial expressions and the nuance of human motion are challenging to capture. In our data collection with virtual reality, we were limited by the features captured by the Vive Pro Eye VR headset, which describe the geometry of the face through blend shapes. We visualized this data by rendering the features on a virtual avatar head, and this could have affected the perception of subtle human facial expressions. In the future, it would be interesting to utilize more advanced devices such as the recently released Apple Vision Pro to create other datasets of implicit human feedback. The new Apple device can sense faces in a way that allows rendering higher quality avatars for users, and the data it captures could potentially improve the accuracy and robustness of ML models that predict robot performance.

In the future, inferred performance predictions could be used to adapt robot behavior. For example, a robot could use binary robot performance predictions as instantaneous rewards that guide changes in robot behavior to better align what the robot does with human preferences [22, 42, 48]. When the predictions indicate low robot performance or suggest drastic changes in impressions of the robot's behavior, the robot could also opt for querying users explicitly about its performance to verify the predictions. Perhaps the responses can also be used to improve the prediction model.

9 CONCLUSION

This work contributes the SEAN TOGETHER Dataset, consisting of observations of human-robot interactions in VR, including implicit human feedback, and corresponding performance ratings in guided robot navigation tasks. Our analyses with VR data revealed that facial expressions can help predict impressions of the robot, but spatial behavior features in the context of the navigation task were more critical for these inferences. Our experiments also demonstrated the ability of humans and ML models to infer perceived robot performance from interaction observations. A general trend that we observed throughout this work was that predicting the directionality of impressions of robot performance (as a binary classification task) was easier and, thus, seemed more practical than predicting exact performance ratings (on a 5-point scale).

As part of this work, we also conducted a real-world demonstration that showed the applicability of machine learning in predicting human perceptions of a mobile robot in indoor environments. We did not capture facial expression features for this demonstration, but rather focused on capturing features that described the navigation behavior of the robot and humans based on our prior findings. Both the models trained with VR data and real-world data showed promising generalization capabilities when evaluated on real-world data, confirming the potential of machine learning for predicting impressions of robot performance from implicit feedback signals in social robot navigation. Our datasets, accompanying analyses, and demonstration facilitate future research on more scalable supervision of robot navigation behavior. Potentially, robots could use implicit human feedback as supervision to interactively improve their behavior in the future.

ACKNOWLEDGMENTS

This work was partially supported by the National Science Foundation (Grant No. IIS-1924802, IIS-2143109, and IIS-2106690) and Google. We are grateful to Carolina Parada and Leila Takayama for their valuable feedback.

APPENDIX

This appendix first shows the distributions of ground truth labels provided by participants over 10-minute intervals of the data collection sessions. Second, we provide additional details about the annotation tool that we used to collect human annotators' predictions of users' impressions of robot performance, analyze inter-rater reliability for the human annotations, and discuss further findings from the human annotation samples. Then, we provide the full list of features used for predicting human impression of robot performance, with a brief description of each feature. Lastly, we describe the specific model architectures that we used and our training

procedure. These details are included in this document to facilitate better understanding of our methodology and reproducibility of our work.

A DISTRIBUTION OF GROUND-TRUTH LABELS GIVEN BY PARTICIPANTS OVER 10-MINUTE INTERVALS

Fig. A1(a) shows the distributions of ground-truth labels provided by the participants of our VR data collection (Sec. 4 in the paper) over 10-minute intervals of the data collection sessions. Fig. A1(b) shows the distribution of labels from the 120 samples that we randomly drew for human annotation (Sec. 5.1 and 5.2 in the paper). Overall, the distributions of labels over different intervals are similar, except for 30-40 min, which is close to the end of navigation tasks.

B HUMAN ANNOTATION

B.1 Annotation Interface

To reduce misalignment between human annotators, we conducted a couple of pilots for our data collection with the team and lab members, through which we improved our annotation interface and instructions. Fig. A2 and A3 show the instruction pages in our annotation tool.

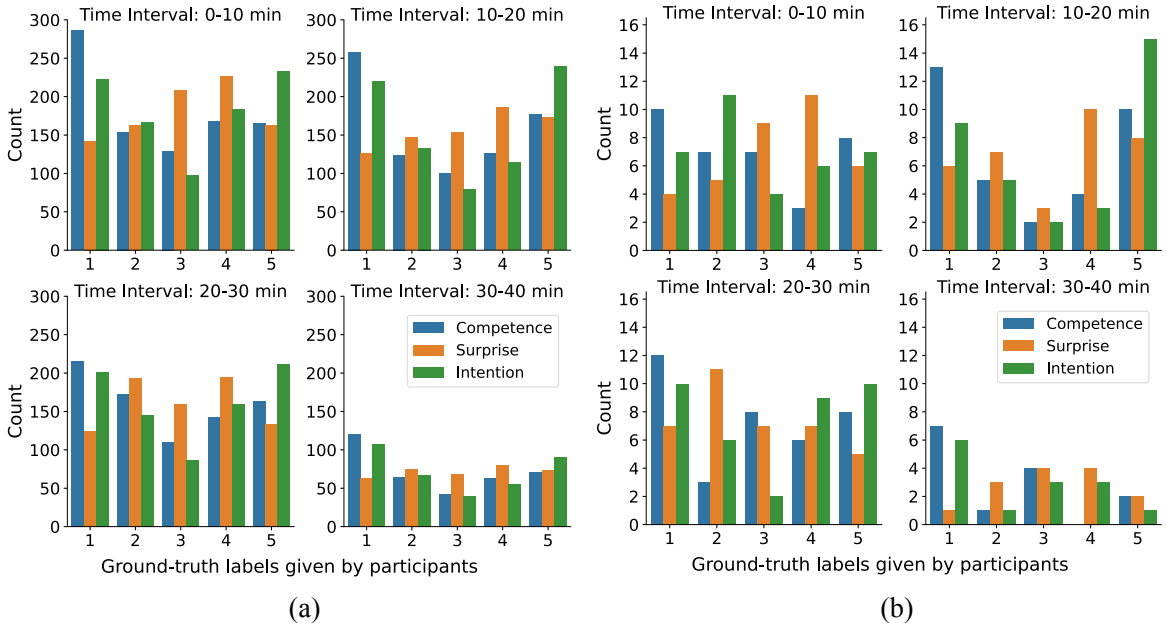

Fig. A1. (a) Distributions of ground truth labels provided by the participants that experienced the human-robot interactions in VR over 10-minute intervals of the data collection sessions; (b) Distributions of ground truth labels used for the human annotation. They are a subset of those in (a). Each plot shows data over 10-minute intervals of the data collection.

Fig. A2. Instruction page 1 that provides the background of social navigation data collected in VR.

B.2 Annotation Reliability

For each visualization of a data sample, we asked 10 different human annotators to provide their predictions on it, which allowed us to compare their prediction performance statistically as reported in the paper. In addition to those results, we also evaluated the reliability of the human annotations. More specifically, we used Krippendorff's alpha to measure the inter-rater reliability for our ordinal labels, which led to an α of 0.67 for competence, 0.54 for surprise, and 0.68 for intention, respectively. These values indicate a moderate to substantial level of agreement among the annotators.

Fig. A4 shows the distribution of labels given by human annotators on the 120 data samples considered for our human baseline.

C FURTHER FINDINGS FROM HUMAN ANNOTATION SAMPLES

Upon reviewing the data that we had collected, we realized that the renderings of participants' faces were shown to annotators with the face mirrored. We were concerned this could have led to confusion among the annotators when they evaluated the gaze direction of the face rendering in comparison to the navigation rendering (e.g., as in Fig. 3 of the paper). Therefore, we repeated the data collection for the *Nav.+Facial* condition, but with the face image not mirrored.

Results for the mirror and not-mirrored data are shown in Table A1. The results in the not-mirrored case only had a subtle difference in comparison to the mirrored case. This suggest that the gaze direction was not an issue and validate the reproducibility of our human annotation experiments.

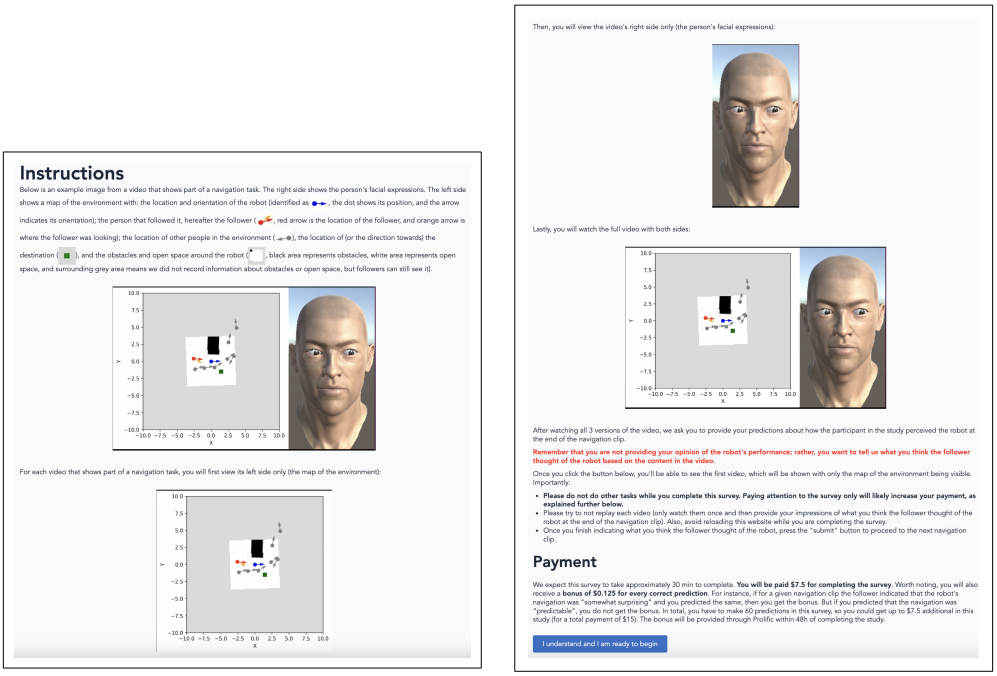

Fig. A3. Instruction page 2 that details the annotation procedures and participant's compensation. The left image is the top of the page while the right image is the continuation of the left image.

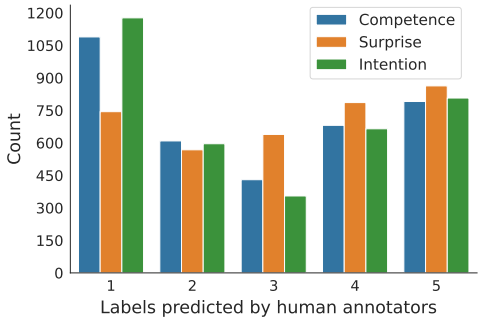

Fig. A4. Distribution of labels predicted by human annotators.

As discussed in Sec. 5.2 of the paper, for the 120 data samples shown to the human annotators, the performance of machine learning models for predicting Intention with *Nav.* and *Nav.+Facial* features decreased in the last two time intervals of data collection. As suggested by Fig. A1(b), a change in the distribution of ground truth labels can be observed in the interval of 30-40 min. This interval also had the fewest number of samples due to how the data was collected, because not all interactions took the full 40 minutes. Also, a good proportion of the samples in the last time interval showed the end of navigation tasks, at which point the participants could have been more sensitive to robot navigation in the wrong direction. Indeed, there was a higher proportion of lower ratings

Table A1. The F_1 Score, the Accuracy, and the Mean Absolute Error of using mirrored and not-mirrored facial rendering when annotating with *Nav.+Facial* features.

	F_1-Score ($\mu \pm \sigma$) ↑		Accuracy ($\mu \pm \sigma$) ↑		Mean Absolute Error ($\mu \pm \sigma$) ↓	
	Mirrored	Not-Mirrored	Mirrored	Not-Mirrored	Mirrored	Not-Mirrored
Competence	0.30 ± 0.1	0.29 ± 0.2	0.43 ± 0.1	0.42 ± 0.1	0.96 ± 0.3	0.99 ± 0.4
Surprise	0.25 ± 0.1	0.25 ± 0.1	0.33 ± 0.1	0.32 ± 0.1	1.09 ± 0.2	1.12 ± 0.2
Intention	0.30 ± 0.1	0.28 ± 0.1	0.42 ± 0.1	0.41 ± 0.1	1.04 ± 0.3	1.07 ± 0.2

for Intention in the last interval than in the other intervals. Such a distribution shift can make the prediction task harder for both the annotators and the machine learning models.

D FEATURE EXTRACTION

The following sections describe in detail the features used for predicting human impressions of robot performance:

Participants' Facial Expression Features:

- *gaze_origin_mm_[left, right]_[x, y, z]*: The gaze origins of left and right eyes, measured in millimeters.
- *gaze_direction_normalized_[left, right]_[x, y, z]*: The normalized gaze directions of left and right eyes.
- *pupil_diameter_mm_[left, right]*: The pupil diameters of left and right eyes, measured in millimeters.
- *eye_openness_[left, right]*: The openness of left and right eyes.
- *pupil_position_in_sensor_area_[left, right]_[x, y]*: The pupil positions of left and right eyes in the sensor area.
- *gaze_origin_mm_combined_[x, y, z]*: The combined gaze origin of left and right eyes, measured in millimeters.
- *gaze_direction_normalized_combined_[x, y, z]*: The normalized combined gaze direction of left and right eyes.
- *pupil_diameter_mm_combined*: The combined pupil diameter of the left and right eyes, measured in millimeters.
- *eye_openness_combined*: The combined eye openness of the left and right eyes.
- *pupil_position_in_sensor_area_combined_[x, y]*: The combined pupil position of left and right eyes in the sensor area.
- *eye_[wide, squeeze, frown]_[left, right]*: The extent of eye wide, squeeze, and frown of left and right eyes.
- *jaw_[right, left, forward, open]*: The extent of jaw being right, left, forward, and open.
- *mouth_ape_shape*: The extent of mouth ape shape.
- *mouth_[upper, lower]_[right, left]*: The extent of upper and lower part of mouth moving to the right and left.
- *mouth_[upper, lower]_overturn*: The extent of upper and lower part of mouth overturning.
- *mouth_pout* The extent of mouth pouting.
- *mouth_smile_[right, left]*: The extent of mouth smiling on the right and left side.
- *mouth_sad_[right, left]*: The extent of mouth being sad on the right and left side.
- *cheek_puff_[right, left]*: The extent of cheek puffing on the right and left side.
- *cheek_suck*: The extent of cheek sucking on the right and left side.
- *mouth_upper_[upright, upleft]*: The extent of the upper part of mouth moving upright and upleft.
- *mouth_lower_[downright, downleft]*: The extent of the lower part of mouth moving downright and downleft.

- *mouth_[upper, lower]_inside*: The extent of the upper and lower part of mouth moving inside.
- *mouth_lower_overlay*: The extent of the lower part of mouth overlaying.
- *tongue_[longstep1, longstep2]*: The extent of the person's tongue stretching long.
- *tongue_[down, up, right, left, roll]*: The extent of the person's tongue moving down, up, right, left, and rolling.
- *tongue_[upleft, upright, downleft, downright]_morph*: The extent of the person's tongue morphing upleft, upright, downleft, and downright.

Spatial Behavior Features:

- *participant_pose_[x, y, cos(θ), sin(θ)]*: The 2D position and orientation of the participant, computed relative to the robot.
- *nearby_agents_pose_[x, y, cos(θ), sin(θ)]*: The 2D positions and orientations of the other automatically-controlled avatars within a 7.2m radius, computed relative to the robot.

Goal Features:

- *goal_[x, y]*: The 2D position of the navigation destination in a coordinate frame attached to the robot.

Occupancy Features:

- *map_resnet18*: The cropped section of the 2D map around the robot (of 7.2m × 7.2m) to describe the occupancy of nearby space by static objects, encoded by ResNet-18 [33].

E MODEL ARCHITECTURE AND HYPERPARAMETERS

For training our machine learning models, the input for each example corresponded to an 8-second window of features (navigation, facial, or both types of features). The synchronized multi-modal data was re-sampled at 5Hz, resulting in an input sequence of 40 timesteps. This helped reduce the length of the series of features input to the model, facilitating learning in practice. The targets for the examples corresponded to the ground truth labels for robot performance at the end of the window. These ground truth labels were provided by the participants in our VR data collection.

For part of our evaluation, we converted the 5-point ratings in the ground truth labels to binary ratings (e.g., as reported in Sec. 5.1 of the paper) as well as converted the output of machine learning models and human ratings from a 5-point scale to a binary output. This conversion was used to evaluate how well human annotators and machine learning models could predict the directionality of robot performance (rather than focusing on the exact performance level indicated in the ground truth labels). The binary classification task could be useful in the future for identifying situations were a robot makes mistakes during navigation and potentially engaging in recovery behaviors.

To facilitate future reproducibility, the next paragraphs provide more details about the specific architectures implemented for the deep learning models considered in our work:

MLP architecture. Our MLP model first encoded the input features at each timestep with a dense linear layer with 256 hidden units. The encoded sequence was then concatenated across timesteps and fed into three nonlinear dense layers with 512, 256, and 64 hidden units, respectively, and with Leaky ReLU activation. Finally, it was passed into a linear layer that output the logits corresponding to the 5 categories of labels to be classified.

Transformer architecture. Our Transformer model first passed the input sequence through a BatchNorm layer, and then encoded at each timestep with a dense layer with 256 hidden units. Positional encoding was applied to the encoded sequence, which was then fed into 2 transformer encoder layers with 4 heads, a feed-forward dimension of 512, and ReLU activation. The output was then concatenated across timesteps and fed into three nonlinear dense layers with 512, 256, and 64 hidden units, respectively, and with Leaky ReLU activation. Finally,

Table A2. The best learning rate, batch size, and dropout of Multi-Layer Perceptron (MLP), Graph Neural Network (GNN), and Transformer (T), chosen using grid search with validation-based early stopping. "Annotation Samples" values correspond to the hyper-parameters for the deep learning models reported in Table I of the paper (Sec. 5.2). "Leave-One-Out" values correspond to the hyper-parameters for the results in Fig. 6 of the paper (Sec. 5.3).

	Learning Rate		Batch Size		Dropout	
	Annotation Samples	Leave-One-Out	Annotation Samples	Leave-One-Out	Annotation Samples	Leave-One-Out
MLP	0.003	0.001	256	512	0.1	0.1
GNN	0.003	0.001	512	512	0.0	0.5
T	0.003	0.0003	512	512	0.0	0.0

the result was passed into a linear layer that output the logits corresponding to the 5 categories of labels to be classified.

Graph Neural Network (GNN) architecture. We constructed a bidirectional, fully-connected graph in order to utilize the relational inductive bias present in the data and process this data using a GNN. Input sequences of temporal data were first divided into three groups, corresponding to the node features, the edge features, and the global features. Node features consisted of the positions and orientations of the participant and nearby agents relative to the robot. Edge features between every pair of nodes in the graph consisted of the Euclidean distance between the two connected nodes. Global features consisted of all other features for a given experiment. A feedforward network with 64 hidden units was created for each of the three groups of temporal data. Then, each time step of each type of temporal data was encoded using the corresponding feedforward network.

The architecture of our model consisted of two message-passing layers [7]. Each edge update function and each node update function was composed of a feedforward network with a ReLU activated, single hidden layer of 64 units. All of the node representations for a graph that resulted from the final message-passing layer were concatenated with the global feature representations that resulted from the temporal encoding of the input global features. Finally, a classification head of three, Leakly ReLU activated, nonlinear dense layers with 512, 256, and 64 hidden units, respectively, was used to output the logits corresponding to the 5 categories of labels to be classified.

The deep learning models were trained using the Cross-Entropy (CE) loss against the ground truth labels. To update model parameters, we used the AdamW optimizer with a weight decay coefficient of 0.01. The best learning rates, batch sizes, and dropout rates found by hyperparameter search are shown in Table A2. All the results described in the paper were obtained with an Intel Core i7 10700K 8-Core 3.6GHz desktop computer that had an NVIDIA 24GB GeForce RTX 3090 GPU.

REFERENCES

[1] Peter Anderson, Ayush Shrivastava, Joanne Truong, Arjun Majumdar, Devi Parikh, Dhruv Batra, and Stefan Lee. 2021. Sim-to-real transfer for vision-and-language navigation. In *Conference on Robot Learning*. PMLR, 671–681.

[2] Georgios Angelopoulos, Alessandra Rossi, Claudia Di Napoli, and Silvia Rossi. 2022. You Are In My Way: Non-verbal Social Cues for Legible Robot Navigation Behaviors. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 657–662.

[3] Reuben M Aronson and Henny Admoni. 2018. Gaze for error detection during human-robot shared manipulation. In *Fundamentals of Joint Action workshop, Robotics: Science and Systems*. 5.

[4] Chatchalita Asavanant and Hiroyuki Umemuro. 2021. Personal space violation by a robot: An application of expectation violation theory in human-robot interaction. In *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE, 1181–1188.

[5] Eleanor Avrunin and Reid Simmons. 2014. Socially-appropriate approach paths using human data. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 1037–1042.

[6] Christoph Bartneck, Tony Belpaeme, Friederike Eyssel, Takayuki Kanda, Merel Keijsers, and Selma Šabanović. 2020. *Human-robot interaction: An introduction.* Cambridge University Press.

[7] Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. 2018. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261* (2018).

[8] Aniket Bera, Tanmay Randhavane, and Dinesh Manocha. 2019. Improving Socially-aware Multi-channel Human Emotion Prediction for Robot Navigation.. In *CVPR Workshops*. 21–27.

[9] Homanga Bharadhwaj, Zihan Wang, Yoshua Bengio, and Liam Paull. 2019. A data-efficient framework for training and sim-to-real transfer of navigation policies. In *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 782–788.

[10] Erdem Bıyık, Aditi Talati, and Dorsa Sadigh. 2022. Aprel: A library for active preference-based reward learning algorithms. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 613–617.

[11] Rita Borgo, Bongshin Lee, Benjamin Bach, Sara Fabrikant, Radu Jianu, Andreas Kerren, Stephen Kobourov, Fintan McGee, Luana Micallef, Tatiana von Landesberger, et al. 2017. Crowdsourcing for information visualization: Promises and pitfalls. In *Evaluation in the Crowd. Crowdsourcing and Human-Centered Experiments: Dagstuhl Seminar 15481, Dagstuhl Castle, Germany, November 22–27, 2015, Revised Contributions*. Springer, 96–138.

[12] Martim Brandao, Gerard Canal, Senka Krivić, Paul Luff, and Amanda Coles. 2021. How experts explain motion planner output: a preliminary user-study to inform the design of explainable planners. In *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE, 299–306.

[13] Cynthia Breazeal, Nick DePalma, Jeff Orkin, Sonia Chernova, and Malte Jung. 2013. Crowdsourcing human-robot interaction: New methods and system evaluation in a public environment. *Journal of Human-Robot Interaction* 2, 1 (2013), 82–111.

[14] Kate Candon, Jesse Chen, Yoony Kim, Zoe Hsu, Nathan Tsoi, , and Marynel Vázquez. 2023. Nonverbal Human Signals Can Help Autonomous Agents Infer Human Preferences for Their Behavior. In *Proceedings of the 22nd International Conference on Autonomous Agents and Multiagent Systems*.

[15] Kate Candon, Nicholas C. Georgiou, Helen Zhou, Sidney Richardson, Qiping Zhang, Brian Scassellati, and Marynel Vázquez. 2024. REACT: Two Datasets for Analyzing Both Human Reactions and Evaluative Feedback to Robots Over Time. arXiv:2402.00190 [cs.RO]

[16] Colleen M Carpinella, Alisa B Wyman, Michael A Perez, and Steven J Stroessner. 2017. The robotic social attributes scale (rosas) development and validation. In *Proceedings of the 2017 ACM/IEEE International Conference on human-robot interaction*. 254–262.

[17] Yuhang Che, Allison M Okamura, and Dorsa Sadigh. 2020. Efficient and trustworthy social navigation via explicit and implicit robot–human communication. *IEEE Transactions on Robotics* 36, 3 (2020), 692–707.

[18] Mohamed Chetouani. 2021. Interactive Robot Learning: An Overview. *ECCAI Advanced Course on Artificial Intelligence* (2021), 140–172.

[19] HeeSun Choi, Cindy Crump, Christian Duriez, Asher Elmquist, Gregory Hager, David Han, Frank Hearl, Jessica Hodgins, Abhinandan Jain, Frederick Leve, et al. 2021. On the use of simulation in robotics: Opportunities, challenges, and suggestions for moving forward. *Proceedings of the National Academy of Sciences* 118, 1 (2021), e1907856118.

[20] Jack Collins, Shelvin Chand, Anthony Vanderkop, and David Howard. 2021. A review of physics simulators for robotic applications. *IEEE Access* 9 (2021), 51416–51431.

[21] Yuchen Cui, Pallavi Koppol, Henny Admoni, Scott Niekum, Reid Simmons, Aaron Steinfeld, and Tesca Fitzgerald. 2021. Understanding the relationship between interactions and outcomes in human-in-the-loop machine learning. In *International Joint Conference on Artificial Intelligence*.

[22] Yuchen Cui, Qiping Zhang, Brad Knox, Alessandro Allievi, Peter Stone, and Scott Niekum. 2021. The empathic framework for task learning from implicit human feedback. In *Conference on Robot Learning*. PMLR, 604–626.

[23] Anca D Dragan, Shira Bauman, Jodi Forlizzi, and Siddhartha S Srinivasa. 2015. Effects of robot motion on human-robot collaboration. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. 51–58.

[24] Anca D Dragan, Kenton CT Lee, and Siddhartha S Srinivasa. 2013. Legibility and predictability of robot motion. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 301–308.

[25] Anthony Francis, Claudia Pérez-d'Arpino, Chengshu Li, Fei Xia, Alexandre Alahi, Rachid Alami, Aniket Bera, Abhijat Biswas, Joydeep Biswas, Rohan Chandra, et al. 2023. Principles and guidelines for evaluating social robot navigation algorithms. *arXiv preprint arXiv:2306.16740* (2023).

[26] Yuxiang Gao and Chien-Ming Huang. 2022. Evaluation of socially-aware robot navigation. *Frontiers in Robotics and AI* 8 (2022), 721317.

[27] Rachel Gockley, Jodi Forlizzi, and Reid Simmons. 2007. Natural person-following behavior for social robots. In *Proceedings of the ACM/IEEE international conference on Human-robot interaction*. 17–24.

[28] Giorgio Grisetti, Cyrill Stachniss, and Wolfram Burgard. 2007. Improved techniques for grid mapping with rao-blackwellized particle filters. *IEEE transactions on Robotics* 23, 1 (2007), 34–46.

[29] Balint Gucsi, Danesh S Tarapore, William Yeoh, Christopher Amato, and Long Tran-Thanh. 2020. To ask or not to ask: A user annoyance aware preference elicitation framework for social robots. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 7935–7940.

[30] Hatice Gunes, Massimo Piccardi, and Maja Pantic. 2008. From the Lab to the Real World: Affect Recognition Using Multiple Cues and Modalities. In *Affective Computing*. IntechOpen.
[31] Edmund T Hall and Edward T Hall. 1966. *The hidden dimension*. Vol. 609. Anchor.
[32] David A. Harville. 1977. Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. *J. Amer. Statist. Assoc.* 72, 358 (1977), 320–338. http://www.jstor.org/stable/2286796
[33] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
[34] Padraig Higgins, Ryan Barron, Stephanie Lukin, Don Engel, and Cynthia Matuszek. 2023. A Collaborative Building Task in VR vs. Reality. In *Proc. of the International Symposium on Experimental Robotics (ISER)* (Chiang Mai, Thailand).
[35] Tom P Huck, Christoph Ledermann, and Torsten Kröger. 2021. Testing robot system safety by creating hazardous human worker behavior in simulation. *IEEE Robotics and Automation Letters* 7, 2 (2021), 770–777.
[36] Angel Hsing-Chi Hwang and Andrea Stevenson Won. 2021. IdeaBot: investigating social facilitation in human-machine team creativity. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.
[37] Tetsunari Inamura, Yoshiaki Mizuchi, and Hiroki Yamada. 2021. VR platform enabling crowdsourcing of embodied HRI experiments–case study of online robot competition. *Advanced Robotics* 35, 11 (2021), 697–703.
[38] Walther Jensen, Simon Hansen, and Hendrik Knoche. 2018. Knowing you, seeing me: Investigating user preferences in drone-human acknowledgement. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–12.
[39] Haresh Karnan, Anirudh Nair, Xuesu Xiao, Garrett Warnell, Sören Pirk, Alexander Toshev, Justin Hart, Joydeep Biswas, and Peter Stone. 2022. Socially compliant navigation dataset (scand): A large-scale dataset of demonstrations for social navigation. *IEEE Robotics and Automation Letters* 7, 4 (2022), 11807–11814.
[40] Adam Kendon. 1988. Goffman's approach to face-to-face interaction. *Erving Goffman: Exploring the interaction order* (1988).
[41] Ross A Knepper, Christoforos I Mavrogiannis, Julia Proft, and Claire Liang. 2017. Implicit communication in a joint action. In *Proceedings of the 2017 acm/ieee international conference on human-robot interaction*. 283–292.
[42] W Bradley Knox and Peter Stone. 2009. Interactively shaping agents via human reinforcement: The TAMER framework. In *Proceedings of the fifth international conference on Knowledge capture*. 9–16.
[43] Alexander Lew, Sydney Thompson, Nathan Tsoi, and Marynel Vázquez. 2023. Shutter, the Robot Photographer: Leveraging Behavior Trees for Public, In-the-Wild Human-Robot Interactions. *arXiv preprint arXiv:2302.00191* (2023).
[44] Rui Li, Marc van Almkerk, Sanne van Waveren, Elizabeth Carter, and Iolanda Leite. 2019. Comparing human-robot proxemics between virtual reality and the real world. In *2019 14th ACM/IEEE international conference on human-robot interaction (HRI)*. IEEE, 431–439.
[45] Jinying Lin, Zhen Ma, Randy Gomez, Keisuke Nakamura, Bo He, and Guangliang Li. 2020. A review on interactive reinforcement learning from human social feedback. *IEEE Access* 8 (2020), 120757–120765.
[46] Shih-Yun Lo, Katsu Yamane, and Ken-ichiro Sugiyama. 2019. Perception of pedestrian avoidance strategies of a self-balancing mobile robot. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 1243–1250.
[47] David V Lu, Dave Hershberger, and William D Smart. 2014. Layered costmaps for context-sensitive navigation. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 709–715.
[48] James MacGlashan, Mark K Ho, Robert Loftin, Bei Peng, Guan Wang, David L Roberts, Matthew E Taylor, and Michael L Littman. 2017. Interactive learning from policy-dependent human feedback. In *International conference on machine learning*. PMLR, 2285–2294.
[49] Roberto Martín-Martín, Hamid Rezatofighi, Abhijeet Shenoi, Mihir Patel, J Gwak, Nathan Dass, Alan Federman, Patrick Goebel, and Silvio Savarese. 2019. Jrdb: A dataset and benchmark for visual perception for navigation in human environments. *arXiv preprint arXiv:1910.11792* (2019).
[50] Christoforos Mavrogiannis, Patrícia Alves-Oliveira, Wil Thomason, and Ross A Knepper. 2022. Social momentum: Design and evaluation of a framework for socially competent robot navigation. *ACM Transactions on Human-Robot Interaction (THRI)* 11, 2 (2022), 1–37.
[51] Christoforos Mavrogiannis, Francesca Baldini, Allan Wang, Dapeng Zhao, Pete Trautman, Aaron Steinfeld, and Jean Oh. 2023. Core challenges of social robot navigation: A survey. *ACM Transactions on Human-Robot Interaction* 12, 3 (2023), 1–39.
[52] Emily McQuillin, Nikhil Churamani, and Hatice Gunes. 2022. Learning socially appropriate robo-waiter behaviours through real-time user feedback. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 541–550.
[53] Daxton Mitchell, HeeSun Choi, and Justin M Haney. 2020. Safety Perception and Behaviors during Human-Robot Interaction in Virtual Environments. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 64. SAGE Publications Sage CA: Los Angeles, CA, 2087–2091.
[54] Noriaki Mitsunaga, Christian Smith, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. 2008. Adapting robot behavior for human–robot interaction. *IEEE Transactions on Robotics* 24, 4 (2008), 911–916.
[55] Oskar Palinko, Francesco Rea, Giulio Sandini, and Alessandra Sciutti. 2016. Robot reading human gaze: Why eye tracking is better than head tracking for human-robot collaboration. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 5048–5054.

[56] Sören Pirk, Edward Lee, Xuesu Xiao, Leila Takayama, Anthony Francis, and Alexander Toshev. 2022. A protocol for validating social navigation policies. *arXiv preprint arXiv:2204.05443* (2022).
[57] Lutz Prechelt. 2002. Early stopping-but when? In *Neural Networks: Tricks of the trade*. Springer, 55–69.
[58] Morgan Quigley, Ken Conley, Brian Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler, and Andrew Ng. 2009. ROS: an open-source Robot Operating System. *ICRA Workshop on Open Source Software* 3.
[59] Claire Rivoire and Angelica Lim. 2016. The delicate balance of boring and annoying: Learning proactive timing in long-term human robot interaction. (2016).
[60] Dorsa Sadigh, Shankar Sastry, Sanjit A Seshia, and Anca D Dragan. 2016. Planning for autonomous cars that leverage effects on human actions.. In *Robotics: Science and systems*, Vol. 2. Ann Arbor, MI, USA, 1–9.
[61] Alessandra Sciutti, Martina Mara, Vincenzo Tagliasco, and Giulio Sandini. 2018. Humanizing human-robot interaction: On the importance of mutual understanding. *IEEE Technology and Society Magazine* 37, 1 (2018), 22–29.
[62] Masahiro Shiomi, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. 2010. A larger audience, please!—Encouraging people to listen to a guide robot. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 31–38.
[63] Maia Stiber. 2022. Effective Human-Robot Collaboration via Generalized Robot Error Management Using Natural Human Responses. In *Proceedings of the 2022 International Conference on Multimodal Interaction*. 673–678.
[64] Maia Stiber, Russell Taylor, and Chien-Ming Huang. 2022. Modeling Human Response to Robot Errors for Timely Error Detection. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 676–683.
[65] Maia Stiber, Russell H. Taylor, and Chien-Ming Huang. 2023. On Using Social Signals to Enable Flexible Error-Aware HRI. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction* (Stockholm, Sweden) *(HRI '23)*. Association for Computing Machinery, New York, NY, USA, 222–230. https://doi.org/10.1145/3568162.3576990
[66] Walter W Stroup. 2012. *Generalized linear mixed models: modern concepts, methods and applications*. CRC press.
[67] Aamodh Suresh, Angelique Taylor, Laurel D Riek, and Sonia Martinez. 2023. Robot Navigation in Risky, Crowded Environments: Understanding Human Preferences. *arXiv preprint arXiv:2303.08284* (2023).
[68] Xiang Zhi Tan, Samantha Reig, Elizabeth J Carter, and Aaron Steinfeld. 2019. From one to another: how robot-robot interaction affects users' perceptions following a transition between robots. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 114–122.
[69] Andrea L Thomaz and Cynthia Breazeal. 2008. Teachable robots: Understanding human teaching behavior to build more effective robot learners. *Artificial Intelligence* 172, 6-7 (2008), 716–737.
[70] Leimin Tian and Sharon Oviatt. 2021. A taxonomy of social errors in human-robot interaction. *ACM Transactions on Human-Robot Interaction (THRI)* 10, 2 (2021), 1–32.
[71] Russell Toris, David Kent, and Sonia Chernova. 2014. The robot management system: A framework for conducting human-robot interaction studies through crowdsourcing. *Journal of Human-Robot Interaction* 3, 2 (2014), 25–49.
[72] Pete Trautman, Jeremy Ma, Richard M Murray, and Andreas Krause. 2015. Robot navigation in dense human crowds: Statistical models and experimental studies of human–robot cooperation. *The International Journal of Robotics Research* 34, 3 (2015), 335–356.
[73] Nathan Tsoi, Mohamed Hussein, Olivia Fugikawa, JD Zhao, and Marynel Vázquez. 2021. An approach to deploy interactive robotic simulators on the web for hri experiments: Results in social robot navigation. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 7528–7535.
[74] Nathan Tsoi, Alec Xiang, Peter Yu, Samuel S Sohn, Greg Schwartz, Subashri Ramesh, Mohamed Hussein, Anjali W Gupta, Mubbasir Kapadia, and Marynel Vázquez. 2022. Sean 2.0: Formalizing and generating social situations for robot navigation. *IEEE Robotics and Automation Letters* 7, 4 (2022), 11047–11054.
[75] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
[76] Marynel Vázquez, Yofti Milkessa, Michelle M Li, and Neha Govil. 2020. Gaze by Semi-Virtual Robotic Heads: Effects of Eye and Head Motion. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 11065–11071.
[77] Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. 2009. Social signal processing: Survey of an emerging domain. *Image and vision computing* 27, 12 (2009), 1743–1759.
[78] Lennart Wachowiak, Peter Tisnikar, Gerard Canal, Andrew Coles, Matteo Leonetti, and Oya Celiktutan. 2022. Analysing eye gaze patterns during confusion and errors in human–agent collaborations. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 224–229.
[79] Yukang Yan, Chun Yu, Wengrui Zheng, Ruining Tang, Xuhai Xu, and Yuanchun Shi. 2020. Frownonerror: Interrupting responses from smart speakers by facial expressions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.
[80] Qiping Zhang, Austin Narcomey, Kate Candon, and Marynel Vázquez. 2023. Self-Annotation Methods for Aligning Implicit and Explicit Human Feedback in Human-Robot Interaction. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. 398–407.

[81] Qiping Zhang, Nathan Tsoi, and Marynel Vázquez. 2023. SEAN-VR: An Immersive Virtual Reality Experience for Evaluating Social Robot Navigation. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. 902–904.

[82] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. 2019. Deephuman: 3d human reconstruction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7739–7749.