Reciprocal MIND MELD: Improving Learning From Demonstration via Personalized, Reciprocal Teaching

Anonymous Author(s)
Affiliation
Address
email

Abstract:
Endowing robots with the ability to learn novel tasks via demonstrations will increase the accessibility of robots for non-expert, non-roboticists. However, research has shown that humans can be poor teachers [1], making it difficult for robots to effectively learn from humans. If the robot could instruct humans how to provide better demonstrations, then humans might be able to effectively teach a broader range of novel, out-of-distribution tasks. In this work, we introduce Reciprocal MIND MELD, a framework in which the robot learns the way in which a demonstrator is suboptimal and utilizes this information to provide feedback to the demonstrator to improve upon their demonstrations. We additionally develop an Embedding Predictor Network which learns to predict the demonstrator's personalized embedding online without the need for optimal labels. In a series of human-subject experiments, we demonstrate that robotic feedback can effectively improve human demonstrations ($p < .001$) and that robotic feedback translates into better learning outcomes for a robotic agent on novel tasks ($p = .045$).

Keywords: meta-learning, personalization, imitation learning

1 Introduction

When an individual purchases an in-home cleaning robot, the robot will have to be taught many novel tasks over an extended period of time. The user may have to teach the robot how to move dishes from the dishwasher to the proper location in the cabinets or how to wash the windows and take out the trash. Simply pre-programming these tasks may not be an adequate solution as different users may have differing preferences for how their robot should operate. Therefore, to effectively meet the needs of the end-user, the robot must be capable of successfully learning new tasks quickly via demonstration. Prior work has shown that robots learn well via robot-centric learning from demonstration (LfD) [2] because the robot can effectively learn how to recover from mistakes. While robot-centric LfD works well when the demonstrations are high quality, prior work has also shown that humans tend to provide low-quality corrective demonstrations to robots [3, 4, 5, 6, 7].

Researchers are increasingly focused on designing new algorithms to learn from suboptimal demonstrations [4, 8, 9, 10, 11, 12, 13, 14, 15]. In Chen et al., the authors introduce Self-Supervised Reward Regression (SSRR) which improves upon an agent's ability to learn from suboptimal demonstrations by characterizing the relationship between noise and performance [16]. Brown et al. [17] and Myers et al. [18] improve upon the ability to learn from suboptimal demonstrations by learning a reward function from a ranked set of demonstrations. Schrum et al. introduced MIND MELD [7] which meta-learns an individual-specific embedding describing a teacher's suboptimal tendencies and was shown to outperform prior work in LfD. (See Appendix for more discussion on related works.)

While several approaches have attempted to improve upon a teacher's ability to provide high quality demonstrations via tutorials and video-based instruction [19, 20], prior work has primarily focused

Submitted to the 6th Conference on Robot Learning (CoRL 2022). Do not distribute.

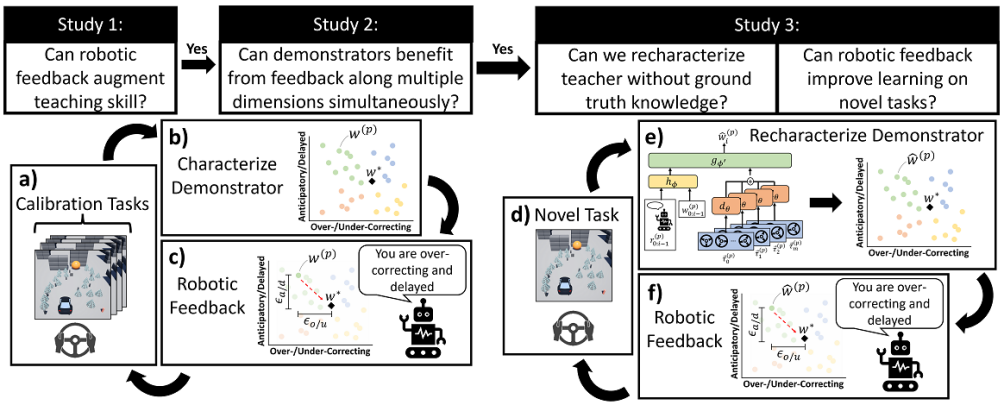


Figure 1: This figure illustrates an overview of our methodology and study designs.

on correcting for suboptimality after-the-fact rather than directly improving the teaching abilities of end-users. We hypothesize that correcting for suboptimal demonstrations under-the-hood (e.g., via a machine-learning algorithm) may not be the best long-term strategy because doing so may 1) contribute to end-users' lack of functional understanding, 2) reinforce suboptimal tendencies, and 3) result in poor performance on out-of-distribution tasks and novel robotic platforms [3, 21, 22].

Consequently, there is a need for a framework that can coach demonstrators to become better teachers. By doing so, we aim to increase functional understanding, increase quality of teachers' demonstrations, and improve end-users' ability to teach novel robotic platforms. To solve this problem, we employ the idea of *reciprocal teaching* from the teaching and education literature [23]. Reciprocal teaching describes a learning strategy in which the student takes on the role of the teacher. We develop a reciprocal teaching framework, Reciprocal Mutual Information Driven Meta-Learning from Demonstration (Reciprocal MIND MELD). Reciprocal MIND MELD learns a *semantically meaningful* personalized embedding that describes the way in which a demonstrator is suboptimal (Fig. 1b and e). Based upon this personalized embedding, Reciprocal MIND MELD provides feedback to the demonstrator to improve their teaching abilities (Fig. 1c and f) and is capable of dynamically updating the demonstrator's personalized embedding. In our work, we contribute the following:

1. We propose Reciprocal MIND MELD, a novel method for providing feedback to demonstrators to improve their teaching abilities via a personalized embedding.
2. We develop an Embedding Predictor Network (EPN) to dynamically update the demonstrator's personalized embedding after receiving feedback (Fig. 1e).
3. We demonstrate that Reciprocal MIND MELD can improve an individual's demonstrations across multiple dimensions of suboptimality ($p < .001$), accurately estimate a demonstrator's new embedding ($p = .002$), and improve learning outcomes of the robot ($p = .045$).

2 Preliminaries

Reciprocal MIND MELD is inspired by the MIND MELD architecture demonstrated in previous work [7]. The objective of MIND MELD is to learn a personalized embedding to describe the way in which a demonstrator is suboptimal in a robot-centric learning from demonstration paradigm in which the demonstrator provides corrective feedback to the robot. MIND MELD then utilizes this embedding to map a demonstrator's suboptimal demonstrations to demonstrations closer to the optimal. The MIND MELD architecture is trained via *calibration tasks*, which are used to learn the mapping from suboptimal labels to better labels and learn the personalized embedding representing an individual demonstrator. The calibration tasks consist of a set of pre-recorded policy rollouts with known optimal labels. Participants provide corrective demonstrations to the robot during these rollouts to direct the robot to a goal. MIND MELD learns to map the participant provided corrective labels to higher-quality labels while simultaneously inferring the personalized embedding, $w^{(p)}$, representing an individual, p's, suboptimal style. To ensure that $w^{(p)}$ can represent various and

distinct feedback styles, MIND MELD maximizes a lower bound on mutual information between the way in which a demonstrator is suboptimal and $w^{(p)}$ via variational inference [24]. Additional details on the MIND MELD architecture can be found in the Appendix.

While prior work demonstrated that MIND MELD is capable of improving upon suboptimal demonstrations, MIND MELD suffers from several key limitations: 1) MIND MELD corrects for suboptimality under-the-hood and does not convey to the demonstrator how best to improve upon their suboptimal tendencies, and 2) MIND MELD assumes that demonstrators are static (i.e., the way in which they are suboptimal does not change over time). Reciprocal MIND MELD overcomes these limitations by 1) providing actionable robotic feedback to the demonstrator to improve upon the quality of their demonstrations and 2) dynamically updating the estimate of their personalized embedding online in order to account for changes in suboptimal tendencies and teaching ability.

3 Methodology

Because humans have a greater ability to generalize to novel tasks and domains than a machine-learning algorithm [22], our objective is to provide demonstrators with knowledge about how to improve their demonstrations rather than correcting suboptimality under-the-hood. We propose an approach to reason about a demonstrator's embedding and provide robotic feedback derived from their embedding that is intended to improve upon their demonstration abilities. In keeping with prior work [25], we investigate the abilities of our approach in a driving simulator domain (see Appendix for additional domain details). We break the problem of improving upon a demonstrator's teaching abilities into three research questions.

RQ1: Can robotic feedback improve upon a demonstrator's teaching abilities?
RQ2: What is the best method to provide robotic feedback to improve teaching abilities?
RQ3: Does robotic feedback result in improved learning outcomes on novel tasks and over time?

3.1 Semantically Meaningful Embedding Space

Prior work [7, 25] has illustrated that MIND MELD learns embeddings that correlate with suboptimal tendencies. These suboptimal tendencies are dependent on the specific robotic domain but are agnostic to the task. For example, in a driving simulator domain, prior work [25] has shown that demonstrators tend to over- or under-correct and provide anticipatory or delayed feedback. Our goal is to learn a *semantically meaningful* embedding space (i.e., a space that can be translated into actionable feedback) and then utilize the location of the demonstrator's embedding within the embedding space to provide actionable robotic feedback.

To learn a semantically meaningful embedding space whose dimensions reflect suboptimal tendencies, we add an additional network head, $p_\psi(w^{(p)}) = \hat{\vec{m}}^{(p)}$, to the MIND MELD architecture to estimate the suboptimal tendency, $\vec{m}^{(p)}$, (i.e., the magnitude by which the demonstrator over-/under-corrects and is anticipatory/delayed). We utilize a mean squared error (MSE) loss, $L(\psi, w) = \frac{1}{N}\sum_i ||p_\psi(w^{(i)}) - \vec{m}^{(i)}||_2^2$, to train the network to predict the suboptimal tendency, $\vec{m}^{(p)}$, given the personalized embedding. This loss helps to ensure that the dimensions of the embedding space are semantically meaningful and can therefore be translated into actionable robotic feedback. Under IRB approval, we leverage the calibration dataset collected in Schrum et al. [25] to learn a semantically meaningful embedding space. This dataset consists of 76 participants who provided demonstrations on a set of calibration tasks. The suboptimal magnitude, $\vec{m}^{(p)}$, is determined via dynamic time warping (DTW) [26] between the participants' feedback

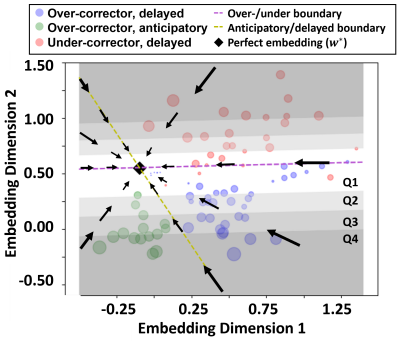


Figure 2: The learned embedding space and decision boundaries. Q1-Q4 indicate quartiles 1-4 for the over-/under-correcting dimension.

and optimal labels from the calibration tasks. Because MIND MELD outputs the difference between the participant's corrective label and the optimal label, the perfect demonstrator's embedding, w^*, is defined as the embedding which minimizes the output of the MIND MELD architecture, $\mathcal{M}$: $w^* = \text{argmin}_{w^{(p)}} \sum_{t,p} \mathcal{M}_{\theta,\phi'}(a^{(p)}_{t-\Delta t:t+\Delta t}, w^{(p)})$, where $a^{(p)}_{t-\Delta t:t+\Delta t}$ is a sequence of demonstrations.

Our next objective is to determine the semantically meaningful dimensions of the embedding space. We train a support vector machine (SVM) with a linear kernel to learn the decision boundaries which best separate the demonstrators into their respective suboptimal categories (over- vs. under-correctors and delayed vs. anticipatory). We utilize an SVM to learn the decision boundaries so that we can add the additional constraint that the classifier must pass through the point representing the perfect demonstrator, w^*. The distance between the embedding and the decision boundary along the suboptimal dimension determines the magnitude by which the demonstrator is suboptimal.

Fig. 2 depicts our embedding space with the linear classifiers separating over-correctors from under-correctors and delayed from anticipatory. The size of the point represents the magnitude by which the demonstrator is suboptimal in the over-/under-correcting dimension as determined by DTW. A similar plot showing the magnitude for the anticipatory/delayed dimension can be found in the Appendix. The plot illustrates that demonstrators who are more suboptimal in the over-/under-correcting dimension (as represented by larger points) are farther away from the over-/under-correcting decision boundary, supporting our hypothesis that distance from the decision boundary can be used to measure the degree of suboptimality. To further support our claim, we apply Spearman's correlation and find that distance from the decision boundary strongly correlates with the magnitude of suboptimality in both the over-/under-correcting dimension ($\rho = .84$, $p < .001$) and in the anticipatory/delayed dimension ($\rho = .93$, $p < .001$).

3.2 Robotic Feedback

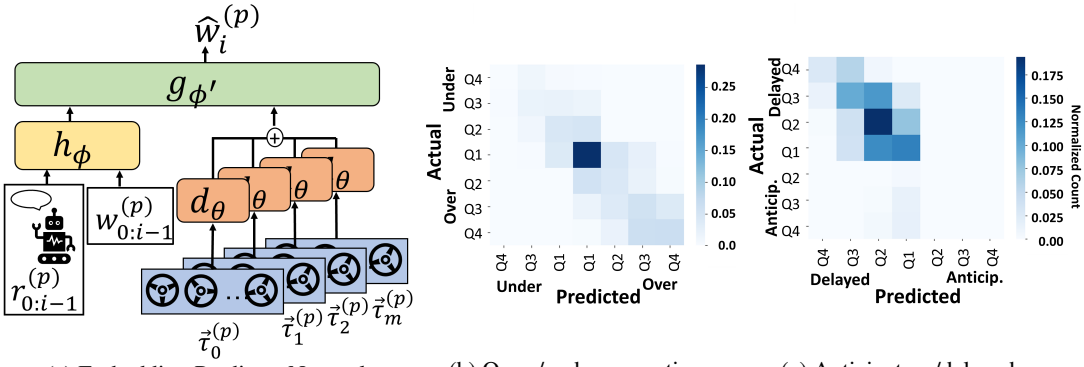
To determine the feedback the robot should provide, we calculate the distance, ϵ, along the semantically meaningful dimension between the personalized embedding, $w^{(p)}$, and perfect embedding, w^*, as shown in Fig. 1. In the driving domain, we are interested in $\epsilon^{(i)}_{o/u}$ and $\epsilon^{(i)}_{a/d}$, which define the distance between the demonstrator's embedding and the hypothetical perfect demonstrator's embedding in the over-/under-correcting dimension and the anticipatory/delayed dimension respectively after the ith round of feedback. In our framework, the feedback is proportional to the distance from w^*.

To convert $\epsilon^{(i)}_{o/u}$ into actionable and intelligible robotic feedback, we discretize the range of $\epsilon^{(i)}_{o/u}$ by splitting the embeddings from the previously collected calibration participants into quartiles as shown in Fig. 2. Our objective is to move a participant's embedding so that they are in the range denoting the 25% of calibration participants who are the least suboptimal (i.e., quartile one). Participants who fall in a quartile farther from the decision boundary receive feedback proportional to their quartile. For example, if a participant falls in the fourth quartile in the over-/under-correcting dimension, the robot will instruct the participant to turn the wheel a lot less whereas if they fall in the second quartile, they will be instructed to turn the wheel slightly less. A table showing the feedback for each quartile and dimension can be found in Appendix Table 1.

3.3 Online Embedding Estimate

To determine if additional feedback should be provided to the demonstrator and if so, the form of the feedback, we must update our estimate of $w^{(p)}$ after each iteration of robotic feedback. One option to update our estimate of the embedding is to have the demonstrator redo the calibration tasks. However, doing so is time consuming and increases the workload of the demonstrator.

Instead, we propose to dynamically update the embedding online using an LSTM-based architecture which extracts salient features from the demonstrations to estimate the personalized embedding rather than relying on calibration tasks which require known, optimal labels. For example, the velocity and magnitude with which the demonstrator turns the steering wheel are two salient features which can inform the estimate of the new embedding. We call this network the Embedding



(a) Embedding Predictor Network (b) Over-/under-correcting (c) Anticipatory/delayed

Figure 3: Fig. 3a illustrates our EPN architecture. Fig. 3b and 3c show our network's ability to predict the quartile that the embedding falls within on holdout test tasks for the over-/under-correcting dimension and anticipatory/delayed dimension respectively.

Predictor Network (EPN) (Fig. 3a). The input to the EPN is the set of new demonstrations, $\tau_{0:m}^{(p)}$, the demonstrator's previous embeddings, $w_{0:i-1}^{(p)}$, and the robotic feedback that was previously provided to the demonstrator, $r_{0:i-1}^{(p)}$. The output of the EPN is an estimate of the new personalized embedding, $\hat{w}_i^{(p)}$. This network utilizes two LSTM subnetworks, h_ϕ and d_θ, the output of which is then fed into subnetwork, $g_{\phi'}$, made up of linear layers with ReLU activations. The inputs to h_ϕ are $w_{0:i-1}^{(p)}$ and $r_{0:i-1}^{(p)}$. Each trajectory, $\tau_t^{(p)}$, is fed into an LSTM subnetwork, d_θ. We then average across the outputs of d_θ and feed the result into $g_{\phi'}$ which produces our embedding estimate, $w_i^{(p)}$. We choose to average across the outputs of d_θ so that our network is agnostic to the number of trajectory inputs.

We train our predictor network on the data collected in Studies 1 and 2 as described in Section 4. Fig. 3b and 3c show the ability of the network to accurately predict the quartile of suboptimality in the over-/under-correcting dimension and the anticipatory/delayed dimension respectively.

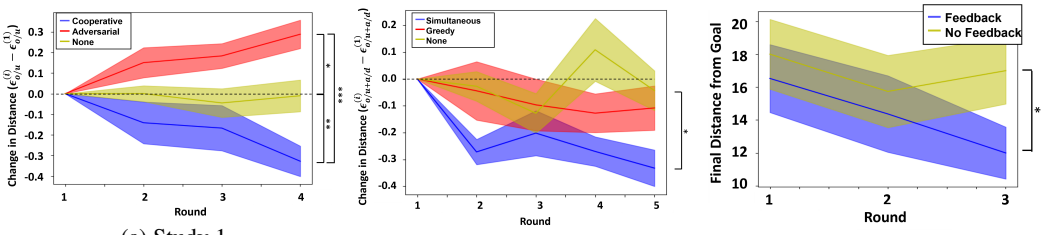
4 Human-Subjects Studies, Results, and Discussion

To determine if our Reciprocal MIND MELD framework is able to improve upon a demonstrator's ability to provide high-quality demonstrations, we conduct three human-subjects studies. The objective of Study 1 is to determine if we are able to shift a demonstrator's embedding via robotic feedback in one dimension (i.e., the over-/under-correcting dimension). In Study 2, we investigate if, and how best, we can shift a demonstrator's embedding in two dimensions. In Study 3, we determine if 1) robotic feedback derived from our EPN rather than the calibration tasks is a good metric of teacher suboptimality and 2) if robotic feedback improves teaching outcomes over time (RQ3). In our analysis, we check all parametric models for normality and homoscedasticity. Model details, tests for assumptions, and additional results are in the Appendix.

4.1 Study 1 (RQ1)

Our objective in Study 1 is to demonstrate that robotic feedback can effectively modulate a participant's teaching characteristics. In this study, we start by investigating feedback only in the over-/under-correcting dimension. After completing demographic information and pre-study surveys, participants complete four rounds of the calibration tasks to measure how their embedding is changing and receive robotic feedback between each round. Between each round, participants complete trust [27] and fluency [28] surveys to determine their subjective perceptions of the robot.

Conditions: In the *Cooperative* condition, the robot provides feedback to improve the demonstrator's teaching. In the *Adversarial* condition, the robot provides feedback to make the participant a worse demonstrator. In the *None* condition, the participant does not receive any feedback.



(a) Study 1 (b) Study 2 (c) Study 3

Figure 4: Fig. 4a and 4b show the difference between the embedding distance at round i, and the embedding distance at round one for Study 1 and Study 2 respectively. Fig. 4c shows the final distance from the goal for the robot after each round of Study 3.

Results: We recruited 27 participants (Mean age = 24.15, SD = 3.4; 37.0% Female) to participate in our study. Fig. 4a shows the change in the distance ($\epsilon^{(i)}_{o/u} - \epsilon^{(1)}_{o/u}$) in the over-/under-correcting dimension between round one and rounds one through four. We plot $\epsilon^{(i)}_{o/u} - \epsilon^{(1)}_{o/u}$ to show how participants change irrespective of their initial teaching skill. We find that the distance at round one, $\epsilon^{(1)}_{o/u}$, is significantly greater from the distance, $\epsilon^{(4)}_{o/u}$, at round four in Cooperative ($\chi^2(1) = 5.44$, $p = .020$) and significantly less in Adversarial ($F(1, 8) = 20.1$, $p = .002$). We find no significant difference in the None condition.

We next apply a one-way ANOVA with Tukey post-hoc to compare $\Delta\epsilon_{o/u}$ across conditions, where $\Delta\epsilon_{o/u} = \epsilon^{(4)}_{o/u} - \epsilon^{(1)}_{o/u}$. $\epsilon^{(1)}_{o/u}$ is the distance from the perfect embedding in round one and $\epsilon^{(4)}_{o/u}$ is the distance in round four. We find that the Adversarial condition results in the embedding shifting significantly farther from the perfect embedding between rounds one to four ($F(2, 24) = 20.2$, $p < .001$) compared to Cooperative ($p < .001$) and None ($p = .014$). We additionally find that Cooperative results in the embedding shifting significantly closer to the perfect embedding ($p = .009$) compared to None. Together, these findings indicate that our approach is capable of modulating a person's teaching style in either direction along the suboptimal dimension. Further, the results in the None condition shows that participants are not simply improving due to repeated interactions.

By applying a one-way ANOVA with Tukey post-hoc, we find that participants' trust increased significantly more ($F(2, 24) = 5.15$, $p = .014$) in Cooperative compared to Adversarial ($p = .020$) and None ($p = .038$). Additionally, we find that participants report a significantly greater positive change in fluency ($F(2, 24) = 5.10$, $p = .014$) in Cooperative compared to Adversarial ($p = .017$). **Takeaway: Robotic feedback can effectively improve a participant's teaching abilities.**

4.2 Study 2 (RQ2)

In Study 2, we next determine how best to provide robotic feedback to both prevent cognitive overload and efficiently improve upon a participant's teaching abilities. Our study design follows the same procedure as Study 1, in which participants complete five rounds of the calibration tasks and receive robotic feedback between each round.

Conditions: In the *Simultaneous* condition, the robot provides feedback related to both the over-/under-correcting and the anticipatory/delayed dimensions. In the *Greedy* condition, the robot only provides feedback related to the condition in which the participant is worst (i.e., furthest from w^*). In the *None* condition, the participant receives no feedback.

Results: We recruited 39 participants (Mean age = 22.46, SD = 3.3; 38.5% Female) to participate in our study. Fig. 4b shows the overall change in the distance ($\epsilon^{(i)}_{o/u+a/d} - \epsilon^{(1)}_{o/u+a/d}$) in the over-/under-correcting and anticipatory/delayed dimension between round one and rounds one through five. We conduct a repeated measures ANOVA and find that the distance at round one, $\epsilon^{(1)}_{o/u+a/d}$, is significantly greater from the distance, $\epsilon^{(5)}_{o/u+a/d}$, at round five in the Simultaneous condition ($F(1, 12) = 22.3$), $p < .001$). We do not find significance for Greedy or None. We next apply a

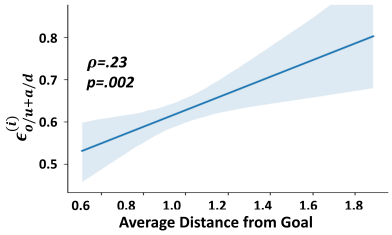


Figure 5: This figure plots the correlation between the embedding distance, $\epsilon^{(i)}_{o/u+a/d}$, and distance from the goal.

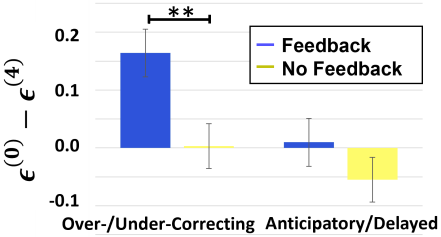


Figure 6: This figure shows the change in $\epsilon_{o/u}$ and $\epsilon_{a/d}$ between the first and last calibration tasks for Feedback and No Feedback.

one-way ANOVA with Tukey post-hoc to compare $\Delta\epsilon_{o/u+a/d}$ across conditions ($F(2, 36) = 3.77$, $p = .033$). We find that the Simultaneous condition results in the embedding shifting significantly closer to the perfect embedding between rounds one to five compared to None ($p = .034$). We do not find significance between None and Greedy or Simultaneous and Greedy.

By applying a one-way ANOVA with Tukey post-hoc, we find that participants' trust ($F(2, 36) = 3.81$, $p = .032$) and team fluency significantly increased ($F(2, 36) = 7.23$, $p = .002$) in Simultaneous compared to None ($p = .029$, $p = .002$ respectively). **Takeaway: Providing feedback in both dimensions simultaneously produces better results for both objective and subjective metrics.**

4.3 Study 3 (RQ3)

In Study 3, we aim to show that our approach and the results from Study 1 and 2 translate to improved learning outcomes for an LfD agent on novel tasks. Participants first complete the calibration tasks to obtain an initial estimate of their embedding, $w_0^{(p)}$, and determine $\epsilon^{(0)}_{o/u}$ and $\epsilon^{(0)}_{a/d}$. Next, the robot provides feedback to the participant intended to improve their demonstrations in both the over-/under-correcting and the anticipatory/delayed dimension given our positive findings for the *Simultaneous* condition in Study 2. Participants then train the robot for three rounds in three different novel environments (i.e., new start and goal locations) for six demonstrations each. Between each environment, we estimate the participant's new embedding, $w_i^{(p)}$, via the EPN, and calculate $\epsilon^{(i)}_{o/u}$ and $\epsilon^{(i)}_{a/d}$ after each round, $i \in \{1, 2, 3\}$. The robot provides robotic feedback based upon the new estimate of the participant's embedding derived from the EPN which is trained on the data collected in Studies 1 and 2. At the end of the study, the participants redo the calibration tasks to determine $\epsilon^{(4)}_{o/u}$ and $\epsilon^{(4)}_{a/d}$. By redoing the calibration tasks, we are able to obtain a ground truth estimate of how the quality of their demonstrations has changed over the course of the study.

Conditions: In the *Feedback* condition, the robot provides feedback to the participant about their demonstrations after each round of the study. In the *No Feedback* condition, the robot still interacts with the participant but does not provide demonstration feedback.

Results: We recruited 60 participants (Mean age = 21.9, SD = 2.89; 28.3% Female). Fig. 4c shows the robot's final distance from the goal for Feedback and No Feedback for rounds 1-3. Participants in the Feedback condition achieve a lower final distance to the goal in round one despite starting off as worse demonstrators on average as measured via the calibration tasks at the beginning of the study (Mean $\epsilon^{(0)}_{o/u+a/d}$ in Feedback: 0.93, Mean $\epsilon^{(0)}_{o/u+a/d}$ in No Feedback: 0.89). Additionally, the final distance of the robot improves over the rounds in Feedback whereas in No Feedback, the robot improves slightly then gets worse in the final round. In round three, the robot achieves a significantly lower final distance from the goal ($Z = -2.0$, $p = .045$) compared to No Feedback.

To determine if the personalized embeddings as estimated by the EPN are a good metric of performance, we compute the correlation between the distance, $\epsilon^{(i)}_{o/u+a/d}$, of the estimated embedding from the perfect embedding and performance as measure by the average distance from the goal for each round, i. Fig. 5 shows that there is a statistically significant correlation ($\rho = .23, p = .002$) be-

tween embedding distance and performance, suggesting that the embeddings estimated by the EPN are a good measure of demonstrator suboptimality.

Next, we investigate the overall change in the quality of the participants' demonstrations as measured via the first set of calibration tasks (conducted at the beginning of the study) and the last set (conducted at the end). Fig. 6 shows the change, $\epsilon^{(0)} - \epsilon^{(4)}$, for the over-/under-correcting dimension and the anticipatory/delayed dimension. We find that participants became statistically significantly better in the Feedback condition ($t(52) = 2.62, p = .006$) compared to the No Feedback condition in the over-/under-correcting dimension. While we do not find significance in the anticipatory/delayed dimension, we do find that participants improve in the Feedback condition whereas they become worse in the No Feedback condition in this dimension.

Lastly, we investigate Feedback versus No Feedback in terms of subjective metrics. We find that Feedback statistically significantly increases trust throughout the study compared to No Feedback ($Z = -2.33$, $p = .019$). We also find that participants rate Feedback as requiring significantly less workload (as measured via NASA TLX [29]) than No Feedback ($t(58.0) = -1.79$, $p = .039$). **Takeaway: Feedback derived from our Embedding Predictor Network improved participant teaching and resulted in better learning outcomes for the robot in novel tasks.**

4.4 Discussion and Limitations

In Study 1, we demonstrated we can shift a demonstrator's embedding both farther from and closer to the perfect embedding depending on whether the demonstrator received feedback from an Adversarial robot or a Cooperative robot respectively ($p < .001$) (RQ1). In Study 2, we found that providing feedback intended to improve upon both dimensions of suboptimality simultaneously is the best strategy and does not cause participants to suffer from an undue level of cognitive overload ($p < .001$) (RQ2). Studies 1 and 2 present strong evidence that robotic feedback is capable of improving upon the quality of a teacher's demonstrations, suggesting that a robot will learn better from a teacher who has received robotic feedback. In Study 3, we tested this hypothesis and investigated the abilities of our EPN to update our estimate of the participant's embedding during novel tasks. We found that final distance of the robot from the goal improves as the demonstrator receives more feedback about their demonstrations ($p = .045$). Overall, we have demonstrated that robotic feedback derived from our Reciprocal MIND MELD architecture results in better learning outcomes for a robot.

Limitations: A limitation of Reciprocal MIND MELD is that domain knowledge is required to determine the dimensions of suboptimality in a specific domain. However, robotic domains share many similarities in terms of the control interfaces and the potential for suboptimality, suggesting that the dimensions in one domain will likely be similar in others. In this work, we investigate verbal feedback to improve upon demonstration quality. However, prior work has suggested that alternative methods of providing feedback such as videos may be more effective at improving a teacher's ability to provide high-quality demonstrations [19]. We leave to future work an investigation of the best modality for providing feedback to a human demonstrator. Additionally, we only explored Reciprocal MIND MELD's abilities over short time frame with mostly college aged students. In future work, we propose to sample from a more diverse participant pool over a longer time horizon.

5 Conclusion

We introduce Reciprocal MIND MELD, a novel LfD framework for providing robotic feedback to a human demonstrator based upon a personalized embedding to improve suboptimal teaching tendencies. We demonstrate our approach in a series of three human-subject experiments in which we show that robotic feedback can improve upon the quality of a teacher's demonstrations, providing feedback in multiple dimensions simultaneously is the most effective method, and robotic feedback results in improved learning outcomes for a robot. Additionally, we show that our Embedding Predictor Network is capable of accurately estimating the updated personalized embedding online, thus enabling continuous feedback to be provided to the demonstrator.

References

[1] M. Laskey, S. Staszak, W. Y. S. Hsieh, J. Mahler, F. T. Pokorny, A. D. Dragan, and K. Goldberg. SHIV: Reducing supervisor burden in DAgger using support vectors for efficient learning from demonstrations in high dimensional state spaces. *Proceedings - IEEE International Conference on Robotics and Automation*, 2016-June:462–469, 2016. ISSN 10504729. doi:10.1109/ICRA.2016.7487167.

[2] S. Ross, G. J. Gordon, and J. A. Bagnell. No-regret reductions for imitation learning and structured prediction. *Aistats*, 15:627–635, 2011. URL http://proceedings.mlr.press/v15/ross11a/ross11a.pdf.

[3] S. Amershi, M. Cakmak, W. B. Knox, and T. Kulesza. Power to the people: The role of humans in interactive machine learning. *AI Magazine*, 35(4):105–120, 2014. ISSN 07384602. doi:10.1609/aimag.v35i4.2513.

[4] J. Spencer, S. Choudhury, M. Barnes, M. Schmittle, M. Chiang, P. Ramadge, and S. Srinivasa. Learning from Interventions: Human-robot interaction as both explicit and implicit feedback. 2020. doi:10.15607/rss.2020.xvi.055.

[5] M. Laskey, C. Chuck, J. Lee, J. Mahler, S. Krishnan, K. Jamieson, A. Dragan, and K. Goldberg. Comparing human-centric and robot-centric sampling for robot deep learning from demonstrations. *Proceedings - IEEE International Conference on Robotics and Automation*, pages 358–365, 2017. ISSN 10504729. doi:10.1109/ICRA.2017.7989046.

[6] J. Berggren. Performance Evaluation of Imitation Learning Algorithms with Human Experts. 2019.

[7] M. L. Schrum, E. Hedlund, and M. C. Gombolay. Improving Robot-Centric Learning from Demonstration via Personalized Embeddings. *arXiv*, 2021. 2110.03134.

[8] H. He, H. Daumé, and J. Eisner. Imitation Learning by Coaching. *Conference on Neural Information Processing Systems*, pages 1–9, 2012.

[9] M. Kelly, C. Sidrane, K. Driggs-Campbell, and M. J. Kochenderfer. HG-DAgger: Interactive imitation learning with human experts. *Proceedings - IEEE International Conference on Robotics and Automation*, 2019-May:8077–8083, 2019. ISSN 10504729. doi:10.1109/ICRA.2019.8793698.

[10] K. Menda, K. Driggs-Campbell, and M. J. Kochenderfer. EnsembleDAgger: A Bayesian Approach to Safe Imitation Learning. *IEEE International Conference on Intelligent Robots and Systems*, (2):5041–5048, 2019. ISSN 21530866. doi:10.1109/IROS40897.2019.8968287.

[11] M. Kaiser, B. B. Group, H. Friedrich, and U. Dillmann. Obtaining good performance from a bad teacher. 1996. URL https://www.researchgate.net/publication/2448780.

[12] B. Burchfiel, C. Tomasi, and R. Parr. Distance minimization for reward learning from scored trajectories. *AAAI Conference on Artificial Intelligence*. URL www.aaai.org.

[13] M. Valko, M. Ghavamzadeh, A. Lazaric, M. P. Deisenroth, C. Szepesvári, and J. Peters. Semi-supervised apprenticeship learning. *JMLR: Workshop and Conference Proceedings*, 24:131–141, 2012.

[14] D. S. Brown, W. Goo, P. Nagarajan, and S. Niekum. Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations. *International Conference on Machine Learning*, 2019. URL https://github.com/hiwonjoon/.

[15] D. H. Grollman, Aude, and G. Billard. Robot learning from failed demonstrations. *International Journal of Social Robotics*, 2012.

[16] L. Chen, R. R. Paleja, and M. C. Gombolay. Learning from suboptimal demonstration via self-supervised reward regression. In *CoRL*, 2020.

[17] D. S. Brown, W. Goo, P. Nagarajan, and S. Niekum. Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations. 4 2019. URL http://arxiv.org/abs/1904.06387.

[18] V. Myers, E. Bıyık, and N. Anari. Learning multimodal rewards from rankings.

[19] M. Cakmak and L. Takayama. Teaching people how to teach robots: The effect of instructional materials and dialog design. pages 431–438. IEEE Computer Society, 2014. ISBN 9781450326582. doi:10.1145/2559636.2559675.

[20] A. Sena and M. Howard. Quantifying teaching behavior in robot learning from demonstration. *International Journal of Robotics Research*, 39:54–72, 1 2020. ISSN 17413176. doi:10.1177/0278364919884623.

[21] M. D. A. P. S. Farace. Department of business and management chair of web analytics and marketing humanization builds trust: the effect of human-like chatbots on the willingness to disclose personal information online.

[22] R. Geirhos, C. R. M. Temme, J. Rauber, H. H. Schütt, M. Bethge, and F. A. Wichmann. Generalisation in humans and deep neural networks. *Conference on Neural Information Processing*. URL https://github.com/rgeirhos/generalisation-humans-DNNs.

[23] P. Annemarie Sullivan. Improving the reading comprehension of junior high students through the reciprocal teaching of comprehension-monitoring strategies. *University of Illinois at Urbana-Champaign. ProQuest Dissertations Publishing*, 1982.

[24] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in Neural Information Processing Systems*, pages 2180–2188, 2016. ISSN 10495258.

[25] M. L. Schrum, E. Hedlund-Botti, N. Moorman, and M. C. Gombolay. MIND MELD: Personalized Meta-Learning for Robot-Centric Imitation Learning. *ACM/IEEE International Conference on Human-Robot Interaction*, 2022.

[26] S. Salvador and P. Chan. FastDTW: Toward Accurate Dynamic Time Warping in Linear Time and Space. 2004.

[27] J.-Y. Jian, A. Bisantz, and C. Drury. Foundations for an Empirically Determined Scale of Trust in Automated Systems. *International Journal of Cognitive Ergonomics*, 4:53–71, 2000. doi:10.1207/S15327566IJCE0401_04.

[28] G. Hoffman. Evaluating Fluency in Human-Robot Collaboration. *IEEE Transactions on Human-Machine Systems*, 49(3):209–218, 2019. ISSN 21682291. doi:10.1109/THMS.2019.2904558.

[29] S. G. Hart and L. E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. *Human mental workload*, 1(3):139–183, 1988.