Towards Realistic Oracles: Modeling Teacher Behavior in Human-Interactive Robot Learning

Martijn R. van der Wijck[1] and Kim Baraka[1,*] and Daniel F. Preciado Vanegas[2,*]

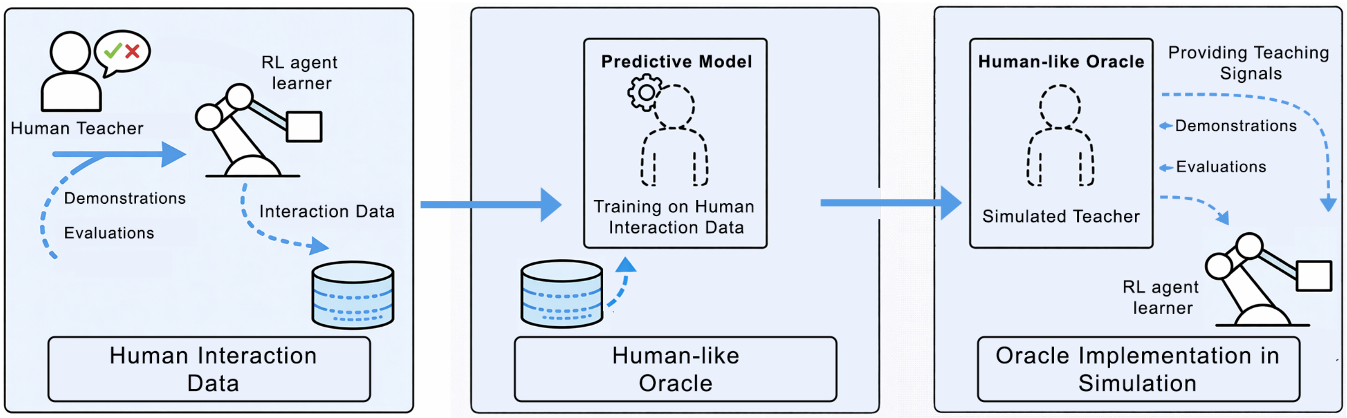


Fig. 1. Overview of the proposed approach. Human interaction data is used to train a predictive model of teaching behavior, which is deployed as a simulated teacher. This simulated teacher provides teaching signals (demonstrations and evaluative feedback) to guide the learning agent.

Abstract— **Human-Interactive Robot Learning (HIRL) often relies on idealized oracle teachers, despite evidence that real human teaching behavior is variable and imperfect. This paper investigates whether human teaching behavior can be modeled from interaction data and reused as a realistic simulated teacher for robot learning. Using data from a prior human-robot interaction study, we train classifiers to predict when a human teacher intervenes and whether they provide evaluative feedback or demonstrations. We compare a flat multi-class classifier with a layered modeling approach that separates intervention prediction from intervention type prediction. The layered model achieves the best predictive performance under severe class imbalance and offers improved interpretability. When integrated into a Panda-gym environment, the simulated teacher successfully guides learning across three tasks. Although it provides fewer interventions and less variability, learning efficiency in successful sessions remains comparable to human teachers. These results show that human teaching behavior can be modeled to create realistic simulated teachers that enable scalable and repeatable evaluation of interactive learning systems.**

I. INTRODUCTION

Human-Interactive Robot Learning (HIRL) [1] has become an important approach for training robotic agents through real-time interaction with human users [1], [2]. By allowing humans to provide guidance in the form of feedback or demonstrations, learning agents can adapt more efficiently and align their behavior with human expectations. This paradigm is particularly relevant in Human-Robot Interaction (HRI) settings, where robots must learn not only task success but also behaviors that are intuitive and compatible with how people naturally teach.

Despite this, many HIRL studies rely on simulated oracle teachers instead of human participants. Such oracles are a practical and widely used tool, as they enable fast, scalable, and repeatable experimentation without the time, cost, and logistical challenges of recruiting human users, particularly during early-stage development and benchmarking. However, these oracles are typically derived from optimal or idealized policies and do not reflect how humans actually teach. Recent work on human feedback modeling highlights that commonly used assumptions about human behavior in reinforcement learning are overly simplistic and fail to capture the variability and context-dependence of real human interaction [3]. As a result, simulated teachers based on such assumptions risk misrepresenting how humans provide guidance in practice.

This issue is closely related to the sim-to-real gap in robotics, where systems developed in simulation often fail to generalize to real-world conditions due to mismatches in variability and system dynamics [4]. In HIRL, a similar gap arises when simulated teachers fail to capture the temporal and behavioral characteristics of human teaching. As a result, conclusions drawn from such simulations may not generalize well to studies involving real human participants.

These limitations highlight the need for simulated teachers that prioritize realism over optimality, capturing how humans actually behave rather than how they should behave. Such models can serve as more faithful substitutes for human

[1]M.R.W. and K.B. are with the Department of Computer Science, Faculty of Science, Vrije Universiteit Amsterdam, The Netherlands k.baraka@vu.nl
[2]D.F.P.V. is with the Department of Communication Science, Faculty of Social Sciences and Humanities, Vrije Universiteit Amsterdam, The Netherlands
*These authors contributed equally as supervising authors.

teachers, enabling more reliable and human-aligned evaluation in interactive robot learning research.

In this work, we investigate whether human teaching behavior can be modeled from interaction data and reused as a realistic simulated teacher for robot learning. We build on a dataset collected by Christofi et al. [5], in which human participants taught a virtual robotic arm using evaluative feedback and demonstrations across three goal-reaching tasks. Rather than modeling an optimal policy, our goal is to capture how humans decide when to intervene and which type of intervention to provide. Fig. 1 provides an overview of the proposed system, illustrating how human feedback data is used to train a predictive teaching model that is deployed as a simulated teacher during reinforcement learning. Specifically, this work addresses the following research question:

- **RQ** To what extent can human teaching behavior be modeled from interaction data to predict whether and how teachers intervene during robot learning?

To address this question, we model human teaching behavior at the level of individual decision steps and compare three approaches: a logistic regression baseline, a flat multi-class classifier, and a layered modeling approach that first predicts whether a human teacher intervenes and subsequently predicts the type of intervention. The layered approach is designed to reflect the structure of human teaching decisions while improving interpretability under strong class imbalance. We evaluate the trained models by integrating the best-performing teacher model into the Panda-gym robotic arm environment, replacing the human teacher during learning and comparing its behavior and effectiveness to that of the original human participants.

The contributions of this paper are threefold. First, we demonstrate that human teaching behavior can be captured using a data-driven model that reproduces key intervention patterns observed in human teachers. Second, we show that a layered modeling approach improves both predictive performance and interpretability compared to standard flat classification approaches. Third, we show that the resulting simulated teacher can be used to evaluate reinforcement learning agents across multiple tasks while preserving key characteristics of human teaching behavior.

By modeling and simulating human teaching behavior, this work contributes to the development of more realistic and scalable evaluation frameworks for interactive robot learning. Rather than relying on idealized oracle teachers, our approach enables repeatable experiments that better reflect the variability and imperfections of human teaching, supporting more reliable evaluation of interactive learning agents.

II. Related Work

A. Human Teaching Behavior in Interactive Robot Learning

Human teaching behavior in interactive robot learning has been shown to be structured, adaptive, and influenced by the interaction context, rather than resembling static oracle-like behavior. Early work by Thomaz and Breazeal [6] demonstrated that people tend to provide more positive than negative feedback when teaching robots, reflecting natural human teaching preferences. Knox et al. [7] further showed that human teachers adapt their feedback strategies over time, providing more frequent evaluative input during early learning phases and reducing interventions as the agent's performance improves. These findings highlight that human teaching behavior is reactive and context-dependent, shaped by the learner's progress rather than fixed instructional rules.

More recent work has focused on how humans structure their teaching behavior across different modalities and interaction contexts. Christofi et al. [8] show that human teachers make strategic decisions when balancing demonstrations and evaluative feedback, with clear patterns in both the timing and frequency of interventions. Their findings indicate that feedback is often provided earlier in the interaction, while demonstrations are used more selectively, suggesting that human teaching behavior is temporally structured and influenced by task progression.

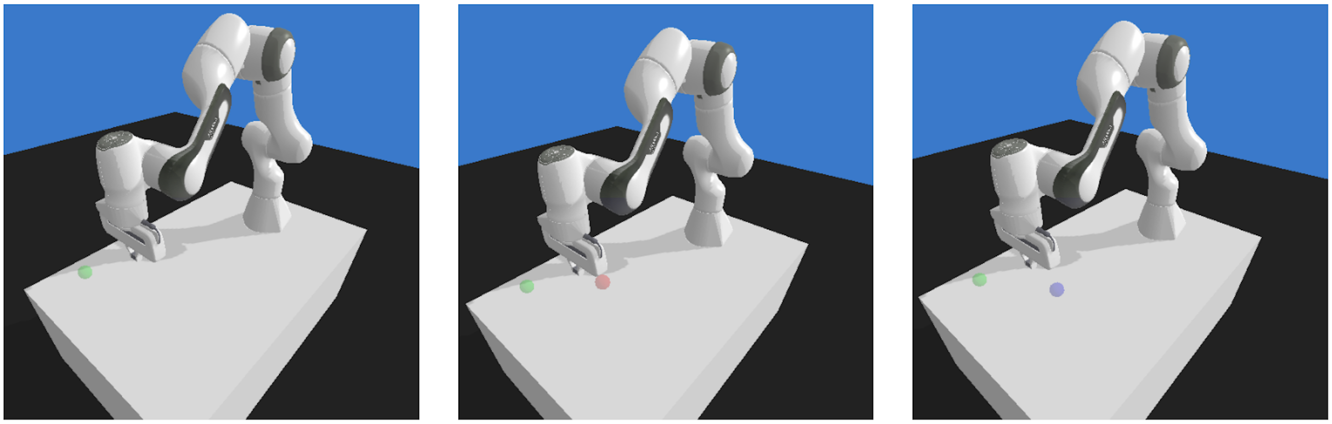
Subsequent studies have also emphasized the substantial variability that exists between individual teachers. Ayub et al. [9] reported the presence of diverse teaching styles in a large-scale study on continual learning, identifying differences ranging from proactive, feedback-heavy teaching to more hands-off approaches. While these differences did not strongly affect final task performance, they suggest that human teaching behavior is far from uniform and cannot be adequately represented by a single idealized model.

However, while prior work has characterized the structure, timing, and variability of human teaching behavior, and proposed evaluation methodologies using simulated users [10], relatively little work has focused on learning predictive models directly from human interaction data that can reproduce these behaviors as realistic simulated teachers.

B. Human-Like Oracles and Simulated Teachers

To reduce reliance on continuous human participation, many reinforcement learning studies employ simulated teachers or oracles that provide feedback or demonstrations. Early approaches often assumed perfect oracles that deliver immediate and accurate guidance, simplifying experimentation and accelerating training [11]. Techniques such as the "Oz of Wizard" paradigm have similarly been used to approximate human input during system development without requiring real users [12]. While effective for early-stage evaluation, such approaches oversimplify real human teaching behavior.

Simulated teachers are commonly used as a practical alternative to human participants, particularly during early-stage development and evaluation of interactive reinforcement learning systems. Human-in-the-loop experiments are often expensive, time-consuming, and difficult to repeat under controlled conditions, motivating the use of simulated users that can approximate human interaction [10], [2]. Such simulated users enable rapid experimentation, reproducibility, and systematic variation of interaction characteristics such as feedback frequency and accuracy.



(a) Task 1 - Reach green sphere

(b) Task 2 - Avoid red (right) sphere and reach green (left) sphere

(c) Task - 3 Reach blue (right) sphere first and then green (left) sphere

Fig. 2. Visual representation of the robotic arm in the Panda-gym environment executing the three tasks (adapted from Christofi et al. [5]).

While simulated users are not a replacement for real human teachers, they provide an efficient and controlled framework for evaluating learning algorithms under predefined human-like constraints [10]. This makes them a valuable tool for benchmarking and analysis, especially when large-scale or repeated experiments are required.

Recent work has further explored the use of imperfect simulated teachers for scalable evaluation. Lee et al. [13] introduced B-Pref, a benchmark for preference-based reinforcement learning that models different forms of teacher irrationality rather than assuming perfect preferences. However, these simulated teachers are based on predefined behavioral assumptions, whereas the present work aims to learn teaching behavior directly from observed human interaction data.

Human feedback in practice is often delayed, sparse, and inconsistent, and this mismatch limits the extent to which simulation results transfer to studies involving real human teachers [10], [11], [14]. Huang et al. [11] proposed a parametric model that captures variability in human feedback along dimensions such as frequency, delay, strictness, bias and accuracy. Their findings revealed discrepancies between self-reported teaching preferences and observed behavior, illustrating the complexity of accurately modeling human instruction. These efforts demonstrate the value of simulated teachers that reflect human imperfections, enabling scalable evaluation while remaining grounded in realistic interaction patterns.

C. Teaching as a Decision-Making Process

Another line of research has examined teaching itself as a structured decision-making process. Torrey and Taylor [15] introduced a teacher-student reinforcement learning framework in which agents provide advice under a limited budget, showing that the timing and strategy of interventions play a critical role in learning efficiency. Related work on policy shaping interprets feedback as probabilistic guidance rather than direct reward, allowing agents to learn effectively even when feedback is noisy or sparse [16].

Layered and hierarchical learning approaches further support the idea that learning and by extension teaching, can be decomposed into structured stages [17]. Together, these studies suggest that teaching decisions can be modeled in terms of *when* to intervene and *how* to intervene. Building on these insights the present work develops a predictive model of human teaching behavior based on observed interaction data. Rather than optimizing teaching strategies, we aim to capture and simulate how humans naturally teach, bridging descriptive analyses of human instruction with practical implementations of realistic simulated teachers for interactive robot learning.

III. Methodology

Our approach builds on prior work in interactive robot learning by Christofi et al. [5] and aims to model and simulate human teaching behavior in a virtual robotic arm environment. The proposed system consists of three components: (1) extracting behavioral features from human-robot interaction data, (2) training predictive models of human teaching decisions, and (3) integrating the best performing model into a reinforcement learning environment as a simulated human teacher to a reinforcement learning agent. To support reproducibility, the implementation used in this work is publicly available.[1]

A. Learning Environment and Tasks

All experiments are conducted in a Panda-gym simulation environment using a virtual Franka Emika-Panda robotic arm [18]. The agent is trained on three goal-reaching tasks of increasing complexity (see Fig. 2): a simple reaching task, a reach task with an obstacle, and a sequential task requiring the agent to reach two targets in order. The learning agent uses Q-learning combined with policy shaping, allowing it to integrate evaluative feedback and demonstrations into its action selection process [16]. Environmental rewards are sparse, with a penalty of -1 per action until task completion, reflecting the original human teaching setup.

[1] https://github.com/mwk207/modeling-teacher-behavior-hirl

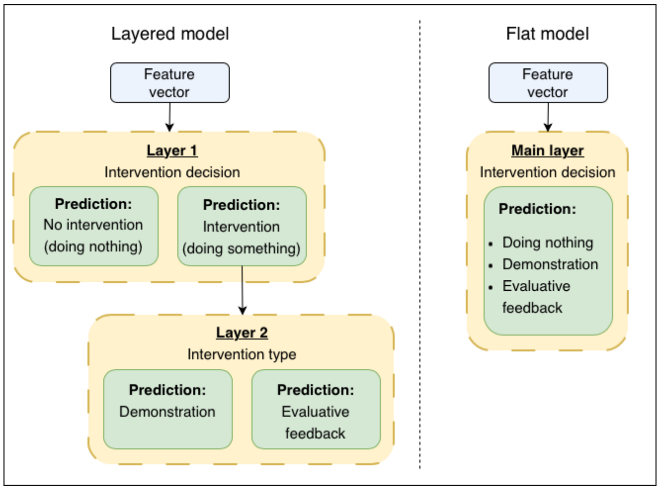


Fig. 3. Comparison between the flat and layered modeling approaches. The flat model directly predicts one of three teaching labels from the feature vector. The layered model decomposes this decision into two stages: first predicting whether an intervention occurs, and subsequently predicting the type of intervention, resulting in the same three possible outputs.

1) Human Teaching Dataset: We use a subset of interaction data from Christofi et al. [5], consisting of 30 participants (21 male-identifying, 9 female-identifying; age range 22–59 years) who taught the robotic arm using evaluative feedback and demonstrations. Participants were allowed to intervene using two teaching modalities: binary evaluative feedback "good" / "bad") and demonstrations consisting of short action sequences. Teaching budgets were limited to 150 demonstration steps and 75 feedback signals per session.

Each participant interacted with the robotic arm in a single session consisting of multiple episodes. An episode is defined as a single attempt by the agent to complete the task, ending either when the goal is reached or when the time limit is exceeded. A session therefore represents a sequence of episodes under the guidance of a single human teacher.

Each timestep in the dataset is labeled as one of three teaching behaviors: no intervention, demonstration, or feedback. The dataset is highly imbalanced, with approximately 85% of timesteps corresponding to no intervention, 9% to demonstrations, and 6% to feedback. This imbalance poses a challenge for predictive modeling and motivates the use of specialized classification strategies.

2) Feature Extraction: To model human teaching decisions, we extract a set of behavior-based features at each timestep. These features are designed to reflect both task progression and interaction history that may influence human teaching decisions, including normalized cumulative demonstrations followed, normalized cumulative feedback followed, elapsed time within the episode, and the Euclidean distance to the goal. All features are computed consistently across tasks and episodes to support generalization.

3) Teaching Behavior Models: We frame the prediction of human teaching behavior as a three-class classification problem, where each timestep is assigned one of the following labels: *no intervention*, *demonstration*, or *evaluative feedback*. These labels represent the possible teaching decisions available to the teacher at each decision point.

For all models, the dataset was split into a training and test set, with 80% of the data used for training and 20% reserved for evaluation. This split was applied consistently across all modeling approaches to ensure fair comparison of predictive performance. The input to the model consists of a feature vector, i.e., a set of behavior-based features describing the current state of the interaction.

We compare three modeling approaches: a logistic regression baseline, a flat Random Forest classifier predicting all three labels directly, and a layered Random Forest model. The flat model directly maps the feature vector to one of the three teaching labels in a single prediction step, without explicitly modeling the intermediate decision of whether to intervene. The layered model decomposes the prediction into two stages and can be interpreted as a hierarchical decision process.

Fig. 3 illustrates the difference between the flat and layered modeling approaches. While the flat model directly predicts one of the three teaching labels from the feature vector, the layered model decomposes this decision into two sequential stages of intervention detection followed by intervention type prediction. The first stage predicts whether an intervention occurs at a given timestep, corresponding to a binary decision between *doing nothing* and *doing something*. If an intervention is predicted, a second stage model is applied to determine the intervention type: *demonstration* or *evaluative feedback*. The final prediction of the layered model is therefore one of the same three teaching labels: doing nothing, demonstration, or evaluative feedback.

This hierarchical structure addresses class imbalance by allowing each stage to focus on a simpler decision and improves interpretability by separating intervention detection from modality selection.

4) Model Integration as Simulated Teacher: To evaluate whether the modeled teacher can function as a realistic substitute for human input, we integrate the trained model into the Panda-gym environment, replacing the human teaching during learning. At each timestep, the agent queries the model for a teaching decision based on the current feature vector. When feedback is predicted, an evaluative signal ("good" / "bad") is injected into the policy shaping module. If there exists historical feedback for the current discretized state, the model draws on the majority label from the recorded human teacher dataset by Christofi et al. [5]. If no historical feedback is available, a fallback feedback signal is generated. The agent's step() function then incorporates this feedback by adjusting its action-selection probabilities via a softmax blend of Q-values and feedback strength.

When a demonstration is predicted, the agent's policy is temporarily suspended and a short action sequence is executed in place of the agent's policy. The system first checks whether a matching demonstration for the current state exists in the same recorded human teacher dataset [5]; if found, it replays it. If no matching demonstration is available, it samples a sequence length from the Gaussian distribution fitted to human demonstration lengths and generates actions (dataset-derived or random) that reduce the discretized distance to the goal. During playback, the agent's policy is still paused, and each demonstrated action is logged. Once the demonstration sequence completes, control returns to the agent's Q-learning policy.

This setup results in a fully automated human-in-the-loop simulation, allowing direct comparison between learning under human teachers and learning under the modeled teacher in a controlled and repeatable manner.

IV. Results

A. Prediction of Human Teaching Behavior

We first evaluated the performance of three models in predicting human teaching behavior from interaction data: a logistic regression baseline, a flat Random Forest classifier, and a layered Random Forest classifier. Performance was assessed using precision, recall and F1 score for each intervention class (no intervention, demonstration, feedback). Table 1 reports the precision, recall and F1 scores for all models across the intervention classes.

TABLE I

Classification performance (precision, recall, and F1 score) of the three models across all intervention classes.

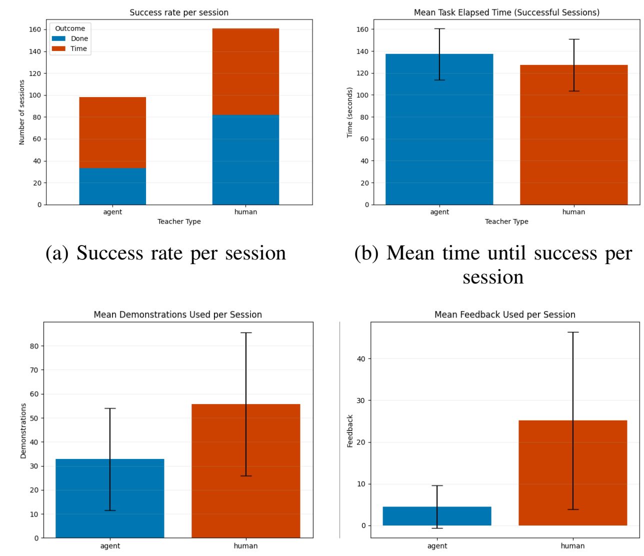
Model	Class	Precision	Recall	F1 Score
Logistic Regression	Nothing	0.87	0.37	0.52
	Demonstration	0.09	0.62	0.16
	Feedback	0.13	0.33	0.19
Flat Random Forest	Nothing	0.99	0.96	0.97
	Demonstration	0.75	0.81	0.78
	Feedback	0.97	0.96	0.97
Layered Random Forest	Nothing	0.97	0.97	0.97
	Demonstration	0.98	0.97	0.98
	Feedback	0.77	0.80	0.78

The logistic regression baseline performed substantially worse than both Random Forest approaches, particularly for the demonstration and feedback classes, with most misclassifications occurring between feedback and demonstration. The no-intervention class was predicted with high accuracy in all models, reflecting the class imbalance in the dataset.

Although the flat Random Forest achieved comparable, and in some cases slightly higher classification scores than the layered model, the layered approach was selected for further use due to its improved interpretability. Feature importance analysis revealed that cumulative reward and elapsed game time were the strongest predictors of whether an intervention occurred, while cumulative feedback followed and elapsed time were most influential in distinguishing between feedback and demonstration interventions. This separation allowed the layered model to provide insight into different aspects of human teaching decisions.

B. Simulated Teacher versus Human Teacher Performance

Next we evaluated the layered model as a simulated human-like teacher by integrating it into the Panda-gym environment and comparing learning outcomes to those achieved under human teachers. Performance was compared across three tasks using success rate, task completion time, total reward, number of interventions, and number of attempts until success. Fig. 4 summarizes these performance measures for both teacher conditions.



(a) Success rate per session

(b) Mean time until success per session

(c) Mean demonstrations per session

(d) Mean feedback per session

Fig. 4. Performance comparison between simulated teacher agents and human teachers. (a) Success rate per session, divided into completed sessions ("Done") and those exceeding the time limit ("Time"); (b) mean elapsed time until goal completion in successful sessions; (c) mean demonstrations per session; (d) mean feedback interventions per session.

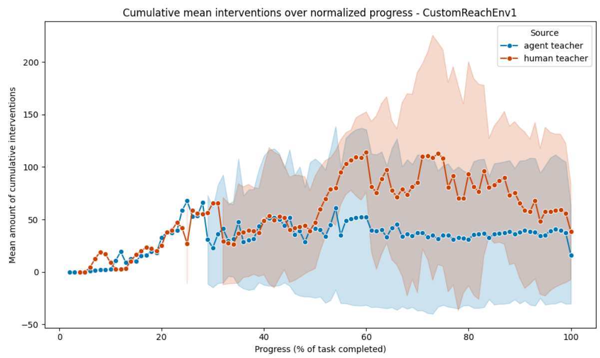
Human teachers achieved a significantly higher overall success rate than the simulated teacher. However, among successful sessions, no significant difference was observed in task completion time or number of attempts until success. This suggests that when the simulated teacher successfully guided the agent, learning efficiency was comparable to that observed under human teaching.

Human teachers provided significantly more teaching interventions than the simulated teacher did, particularly in the form of evaluative feedback. Despite this, the simulated teacher achieved comparable learning performance with fewer interventions. While the simulated teacher yielded higher overall total rewards across all sessions, this difference was not statistically significant when considering only successfully completed sessions.

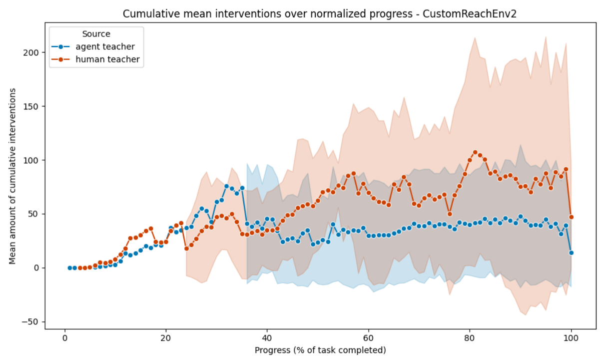
C. Intervention Timing and Teaching Patterns

To further examine differences in teaching behavior over time, Fig. 5 shows the cumulative mean number of teaching interventions plotted against normalized task progress for each task. Across all tasks, the simulated teacher exhibited a more conservative and stable intervention pattern, with lower cumulative counts and reduced variability compared to human teachers.

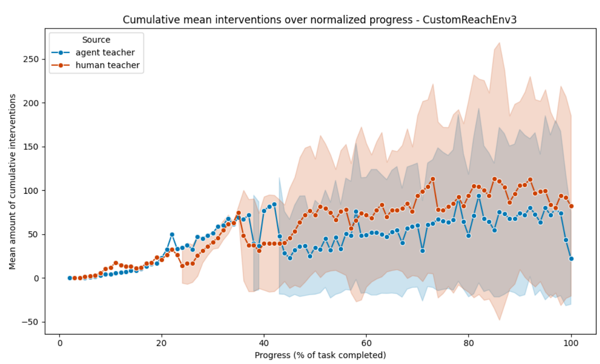
In the first task, both teacher types showed low intervention levels early in the task, after which human teachers intervened more frequently as task progress increased.



(a) Interventions progress task 1



(b) Interventions progress task 2



(c) Interventions progress task 3

Fig. 5. Cumulative mean interventions over normalized task progress for each task. Shaded regions indicate standard deviation, showing that the simulated teacher exhibits more stable and consistent behavior, while human teachers display greater variability.

In the second and third tasks, which involved obstacles or sub-goals, the timing of interventions produced by the simulated teacher closely aligned with those of human teachers, despite consistently lower intervention rates. Human teachers displayed greater variability, particularly in later stages of the tasks.

Overall, these results indicate that while the simulated teacher intervenes less frequently than human teachers, it reproduces key temporal patterns of human teaching behavior, especially in tasks with increased structural complexity.

V. Discussion

A. *Modeling Human Teaching Behavior*

The results demonstrate that human teaching behavior can be captured using a data-driven predictive model trained on interaction data. The layered Random Forest classifier achieved strong performance despite severe class imbalance and provided improved interpretability by separating the decision of *whether* to intervene from the decision of *how* to intervene. Feature importance analysis revealed that cumulative reward and elapsed time were most influential in predicting intervention occurrence, while cumulative feedback followed and elapsed time were key in distinguishing between feedback and demonstrations. These findings support the view of teaching as a structured decision-making process rather than a static policy.

B. *Simulated Teachers versus Human Teachers*

When deployed as a simulated teacher, the layered model successfully guided the reinforcement learning agent to task completion across all three goal-reaching tasks. Compared to human teachers, the simulated teacher intervened less frequently, particularly through evaluative feedback, and exhibited lower variability. While this resulted in a lower overall success rate, learning efficiency in successful sessions, measured by completion time and number of attempts, was comparable between teacher types. This suggests that the modeled teacher captures essential aspects of human teaching behavior without behaving as a perfect oracle. The remaining gap between the simulated and human teachers may stem from only modeling the decision of whether and how to intervene, but not the quality of the intervention itself. While the classifier predicts when demonstrations or feedback should be provided, the content and duration of these teaching signals are only approximated. This may introduce additional suboptimality beyond the prediction task itself. Interestingly, comparable learning efficiency was achieved with fewer interventions, suggesting that human teachers may provide redundant or overly frequent guidance, consistent with prior findings that human teaching behavior is not necessarily optimal, but shaped by heuristics and interaction preferences [7], [6].

C. *Temporal Patterns and Human-Likeness*

Analysis of cumulative interventions over normalized progress revealed that, although the simulated teacher was more conservative, its intervention timing closely aligned with human teaching patterns, particularly in tasks involving obstacles or sub-goals. Human teachers showed greater variability, especially in later task stages, whereas the simulated teacher followed more stable intervention trajectories. These results suggest that the model captures not only aggregate teaching frequency, but also key temporal dynamics of human instruction.

D. *Implications for Interactive Robot Learning*

The ability to simulate human-like teaching behavior has important implications for interactive robot learning research. Behavior-based teacher models enable scalable and repeatable experimentation without continuous involvement of human participants, reducing experimental cost while preserving key aspects of human teaching behavior. By moving beyond idealized oracle teachers, such models provide a more realistic setting for evaluating interactive learning algorithms and help narrow the gap between simulation-based evaluation and real human interaction.

E. Limitations and Future Work

Several limitations should be considered. Differences in temporal granularity between the human-teacher dataset and simulated environment complicate direct comparisons. Furthermore, the small, imbalanced dataset may bias predictions toward non-intervention. The evaluation was also limited to a single robotic platform and a small set of tasks. Finally, the simulated teacher does not explicitly model individual differences between human teachers, which may limit its ability to capture the full diversity of teaching strategies observed in real-world settings.

Future work could address these limitations by collecting larger and more diverse datasets and evaluating the approach across additional robotic platforms, task domains, and user populations. Prior research has shown that human teachers exhibit diverse teaching styles when using demonstrations and feedback [8]. Future studies could therefore investigate personalized, clustered, or parameterized teacher models that explicitly represent different teaching behaviors. Future work could also model the characteristics of the teaching signals themselves and explore inverse RL to infer underlying teaching objectives rather than solely reproducing observed teaching behavior. Finally, incorporating cross-teacher validation and additional baseline comparisons could provide a more comprehensive assessment of robustness and generalization.

VI. CONCLUSION

This paper investigated whether human teaching behavior can be modeled and reused as a simulated teacher for interactive robot learning. Using human-robot interaction data, we trained predictive models to estimate when teachers intervene and which teaching modality they use. A layered Random Forest model proved strong performance and improved interpretability. When integrated into a Panda-gym reinforcement learning environment, the simulated teacher achieved performance comparable to human teachers, despite providing fewer interventions and exhibiting less variability. Although it does not fully capture the adaptability of human instruction, the close alignment with human intervention timing demonstrates that key aspects of human teaching behavior can be modeled successfully.

By focusing on realism rather than optimality, this work provides a foundation for developing simulated teachers that can serve as effective evaluation tools for interactive robot learning research. This enables scalable and repeatable experimentation under human-like teaching conditions, supporting more reliable assessment of learning algorithms, while reducing the need for costly human-subject experiments.

ACKNOWLEDGMENT

The authors would like to thank Konstantinos Christofi for collecting and making available the human-robot interaction dataset that formed the basis of this work.

REFERENCES

[1] K. Baraka et al., “Human-Interactive Robot Learning: Definition, Challenges, and Recommendations,” ACM Trans. Hum.-Robot Interact., vol. 15, no. 2, pp. 1–31, 2026.

[2] L. Zhang, Z. Ji, N. Waytowich, and B. Chen, “GUIDE: Real-Time Human-Shaped Agents,” in Advances in Neural Information Processing Systems 37, Vancouver, BC, Canada: Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024, pp. 138959–138980.

[3] D. Lindner and M. El-Assady, “Humans are not Boltzmann Distributions: Challenges and Opportunities for Modelling Human Feedback and Interaction in Reinforcement Learning,” 2022, arXiv.

[4] W. Zhao, J. P. Queralta, and T. Westerlund, “Sim-to-Real Transfer in Deep Reinforcement Learning for Robotics: a Survey,” in 2020 IEEE Symposium Series on Computational Intelligence (SSCI), Canberra, ACT, Australia: IEEE, Dec. 2020, pp. 737–744.

[5] K. Christofi and K. Baraka, “Uncovering Patterns in Humans that Teach Robots through Demonstrations and Feedback,” in Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction, Boulder CO USA: ACM, 2024, pp. 332–336.

[6] A. L. Thomaz and C. Breazeal, “Teachable robots: Understanding human teaching behavior to build more effective robot learners,” Artificial Intelligence, vol. 172, no. 6–7, pp. 716–737, 2008.

[7] W. B. Knox, B. D. Glass, B. C. Love, W. T. Maddox, and P. Stone, “How Humans Teach Agents: A New Experimental Perspective,” Int J of Soc Robotics, vol. 4, no. 4, pp. 409–421, 2012.

[8] K. Christofi, C. Tichelaar, D. F. Preciado, and K. Baraka, “Human Teaching Patterns in Interactive Robot Learning from Multiple Teaching Modalities,” in 2025 34th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), Eindhoven, Netherlands: IEEE, 2025, pp. 1181–1187.

[9] A. Ayub, J. Mehta, Z. D. Francesco, P. Holthaus, K. Dautenhahn, and C. L. Nehaniv, “How Do Human Users Teach a Continual Learning Robot in Repeated Interactions?,” in 2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), Busan, Republic of Korea: IEEE, 2023, pp. 1975–1982.

[10] A. Bignold, F. Cruz, R. Dazeley, P. Vamplew, and C. Foale, “An Evaluation Methodology for Interactive Reinforcement Learning with Simulated Users,” Biomimetics, vol. 6, no. 1, Art. no. 13, 2021.

[11] J. Huang, R. M. Aronson, and E. S. Short, “Modeling Variation in Human Feedback with User Inputs: An Exploratory Methodology,” in Proc. of the 2024 ACM/IEEE International Conference on Human-Robot Interaction, Boulder CO USA: ACM, 2024, pp. 303–312.

[12] A. Steinfeld, O. C. Jenkins, and B. Scassellati, “The oz of wizard: simulating the human for interaction research,” in Proc. of the 4th ACM/IEEE international conference on Human robot interaction, La Jolla California USA: ACM, 2009, pp. 101–108.

[13] K. Lee, L. Smith, A. Dragan, and P. Abbeel, “B-PREF: Benchmarking Preference-Based Reinforcement Learning,” in Proc. of the 35th Conference on Neural Information Processing Systems (NeurIPS), Track on Datasets and Benchmarks, 2021.

[14] S. Krening and K. M. Feigh, “Interaction Algorithm Effect on Human Experience with Reinforcement Learning,” ACM Trans. Hum.-Robot Interact., vol. 7, no. 2, pp. 1–22, 2018.

[15] L. Torrey and M. Taylor, “Teaching on a budget: agents advising agents in reinforcement learning,” in AAMAS ’13: Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems - Volume 1, New York, New York, 2013, pp. 1053–1060.

[16] S. Griffith, K. Subramanian, J. Scholz, C. L. Isbell, and A. Thomaz, “Policy Shaping: Integrating Human Feedback with Reinforcement Learning,” in Proc. of the 27th International Conference on Neural Information Processing Systems (NIPS), USA, 2013, pp. 2625–2633.

[17] C. N. Silla and A. A. Freitas, “A survey of hierarchical classification across different application domains,” Data Min Knowl Disc, vol. 22, no. 1–2, pp. 31–72, 2011.

[18] Q. Gallouédec, N. Cazin, E. Dellandréa, and L. Chen, “panda-gym: Open-source goal-conditioned environments for robotic learning,” 4th Robot Learning Workshop: Self-Supervised and Lifelong Learning at NeurIPS, 2021.