2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

POLITE: Preferences Combined with Highlights in Reinforcement Learning

Simon Holk, Daniel Marta and Iolanda Leite

Abstract— **Many solutions to address the challenge of robot learning have been devised, namely through exploring novel ways for humans to communicate complex goals and tasks in reinforcement learning (RL) setups. One way that experienced recent research interest directly addresses the problem by considering human feedback as preferences between pairs of trajectories (sequences of state-action pairs). However, when simply attributing a single preference to a pair of trajectories that contain many agglomerated steps, key pieces of information are lost in the process. We amplify the initial definition of preferences to account for highlights: state-action pairs of relatively high information (high/low reward) within a preferred trajectory. To include the additional information, we design novel regularization methods within a preference learning framework. To this extent, we present our method which is able to greatly reduce the necessary amount of preferences, by permitting the highlighting of favoured trajectories, in order to reduce the entropy of the credit assignment. We show the effectiveness of our work in both simulation and a user study, which analyzes the feedback given and its implications. We also use the total collected feedback to train a robot policy for socially compliant trajectories in a simulated social navigation environment. We release code and video examples at https://sites.google.com/view/rl-polite**

I. Introduction

Teaching with preferences [1] offers an alternative approach in the field of inverse reinforcement learning (IRL) [2], [3], enabling humans to instruct robots by ranking different trajectories generated by a robot. This learning approach effectively relieves humans from the direct task of providing demonstrations. Teaching with preferences incorporates the essential elements of structural alignment [4], [5] and trajectory variation, which is necessary for effective learning. One major drawback of preference learning approaches stem from the assumption that if a trajectory is preferred, the entire trajectory is uniformly positively favored. However, valuable information might be collapsed in this assumption. Preferences are intertwined with their causality [6], [7], i.e., if one behavior is preferred over another, there may be a reason for it. Neglecting this may generate a distributional shift leading to *causal confusion* [8], [9], where additional interactions with an environment can yield worse performance. Not taking causality into account can lead to situations where one needs to over-query humans to pinpoint sparse goals.

To clarify our motivation, in Fig.1 we draw an example on how to clarify causal confution by introducing additional trajectory (sequences of state-actions pairs) information within preferences. Consider trajectories X and Y (see Fig.1.A). While trajectory X may be preferred over Y, the underlying reasoning is valuable and could be used to avoid learning by over-querying users. Consider additional trajectory highlights (a) and (b) (see Fig.1.B) which could be given as an explanation by a user, where highlight (a) is a high reward sequence, e.g, reaching a pick-up point or a coffee machine; and (b) is a low reward navigation sequence, passing inconveniently between a group of people. In theory, it should be possible for a human to provide a positive signal for (a) and a negative signal for (b) for a more informative reward profile. However, in the usual preference learning paradigm, there is no way to inform the learning agent about which part of a favored trajectory was responsible for their preference over another. This problem becomes amplified the sparser the rewards get. When a behavior is simply preferred, the entropy of the sequence of state-action pairs with respect to their reward value is maximum and represented by a uniform distribution. Consequently, we obtain a reward profile identical to trajectory X (see Fig.1.C); in this case, both (a) and (b) have the same relative value if the highlight information is not considered. Our method addresses this challenge by offering humans the option of choosing positive and negative parts of a trajectory segment, reducing the entropy of their initial preference; or maintaining said entropy, when the human is unable to recognize a part of the trajectory as more important.

In this work, we explore the possibility of achieving higher query efficiency within preference learning by introducing trajectory highlights. Our contributions are as follows:

1) **POLITE**: **P**references C**O**mbined with High**LI**ghts in Reinforcemen**T** L**E**arning (see Sec. IV), a novel preference learning algorithm which introduces novel regularization methods, by allowing humans to provide highlights in search of a causal relation.
2) We provide experimental results where we compare our work to the current state-of-the-art and baseline preference learning approaches.
3) We conduct a user study in a social navigation scenario to collect both human preferences and highlights, to test the robustness of our approach. We also analyze the highlights given and evaluate their effect on the estimated reward function and final policy when compared to just using preferences.

This research has been carried out as part of the Vinnova Competence Center for Trustworthy Edge Computing Systems and Applications at KTH, and partially supported by the Swedish Foundation for Strategic Research (SSF FFL18-0199) and the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

All of the authors are with the Division of Robotics, Perception and Learning, School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, 114 28 Stockholm, Sweden. The authors are also affiliated with Digital Futures. Mail addresses: {sholk, dlmarta, iolanda}@kth.se

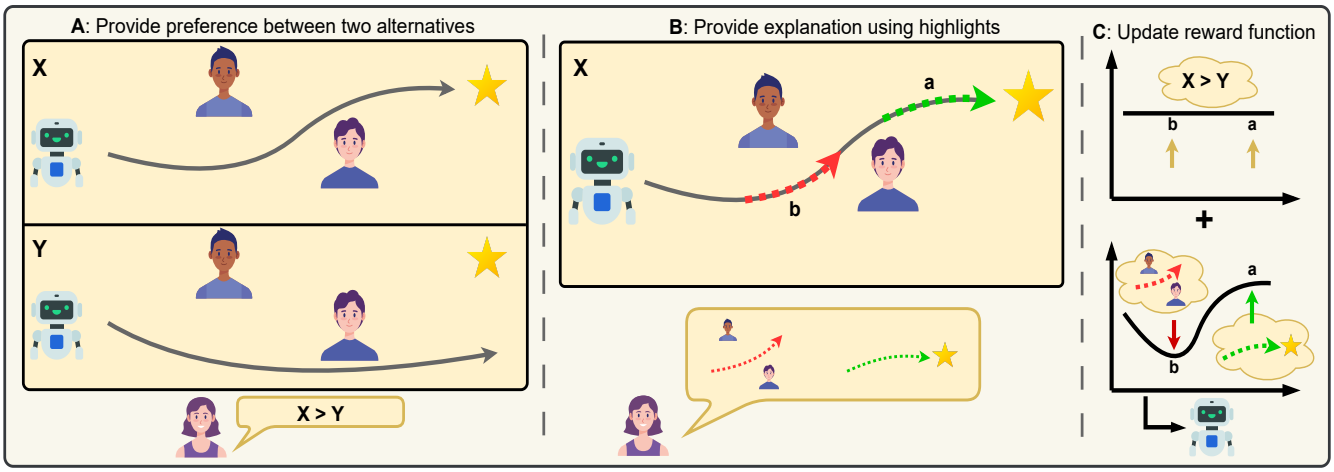

Fig. 1. Jointly integrating preferences and trajectory highlights. (A): A user provides a preference over two trajectories X and Y. (B): A user may choose to highlight part of the preferred trajectory either as good/bad or provide both highlights. (C): A reward function can be jointly updated both with preferences and highlights improving the credit assignment of the favored trajectory.

II. Related Work

Human-in-the-loop learning. Recent research in Human–robot interaction points to the direction that leveraging human knowledge when learning is not only desirable but extremely effective [10], [11], [12]. An interesting human-centric form of learning is through human advice, and more concretely, evaluative feedback [13]. Evaluative feedback can be provided by humans in scalar form [14], natural language [15], trajectory segmentation [16], or through commands and/or buttons signaling preferred behaviors [14], [17], [18]. For example, robots should benefit from having an explicit world model and acquire abstract reasoning [19], instead of building a policy directly which fails to generalize [20]. Not only should robots learn implicitly from features [21], but should be able to explicitly prompt humans for missing information [22]. One solution could be to guide humans through the decision-making process of the agent [23], or to allow humans to correct the robot's behavior during learning [24]. Otherwise, by relying on simple *mimicry*, relevant information related to causality may be collapsed and lead to a distribution shift during learning [8]. Consequently, learning internal and implicit representations by brute-force when over-querying humans, will lead to failure scenarios, lowering humans' trust [25].

We obtained inspiration from the above ideas and set out to design an algorithm that would take advantage of not only the implicit knowledge of preferences but also allow humans to explain their choice in search of a causal relation. Several other approaches are in line with this hypothesis, e.g, adopting divide-and-conquer strategies to simplify reward design [26], and that most tasks can be often represented by a simple reward function and additional constraints [27]. This decomposition property may enable robots to learn tasks of arbitrary complexity through human feedback [28]. There are many works that aim at generalization, by doing distillation of reward networks through demonstrations [29] and extrapolating information beyond sub-optimal demonstrations by learning a reward function from a ranked set [30], [31]. Another interesting work considers trajectories to be segmented by humans into contiguous parts in a Bayesian IRL framework [16]. Most similar approaches in robot learning are also human-centric by design, e.g, including humans in the loop when performing inference [32], or searching for more optimal ways from learning with human teachers in LfD [33], [34], [35], [36], [37]. While efficient, these approaches can be challenging to design for non-experts [38]. Our work presents an alternative approach to include human feedback through IRL with preferences and aims at reducing the number of total queries needed when prompting non-expert humans, leveraging a cause-and-effect relationship.

Learning from preferences. Considering Boltzmann rationals to account for large portions of trajectories, in order to use human feedback in a more sample-efficient manner, poses a promising research path [39], [40]. Consequently, frameworks were proposed to generalize different types of human feedback into Boltzmann rational feedback [41], [42]. Boltzmann noisily-rational decision models [43] have been considered to infer human behavior from goals and preferences, using trajectory distances instead of solely relying on rewards, for additional information gain. Utilizing preferences for learning enjoys an ever-growing body of literature [1]. However, there are many problems left to be solved, such as query selection and employing information strategies on the preferences selected. To address query selection, active learning approaches selecting trajectory pairs in order to remove volume from the distribution of potential rewards have been considered [44], [45]; when volume reduction would lead the robot to ask preferences on similar trajectories, information gain strategies were explored as alternatives [46], [47]. Current preference learning approaches can also be seen as a form of repeated inverse reinforcement learning [48]. The idea is to continuously learn a reward function by asking for preferences on novel iterations of a policy. Approaches that sequentially include humans in the learning loop when inferring reward functions, exhibit improved robustness and mitigate the possibility of humans providing heterogeneous

feedback [36], [49], [50]. In this manner, a reward function may be inferred from *tabula rasa* [49], or bootstrapped through imitation learning [51]. An alternative idea is to train a reward function from an initial set of queries gathered from a pre-trained policy [52], [53]. While our approach can be seen as a natural extension of the above works, our contributions are substantial, and open a path of leveraging the simplicity of using preferences; while exploring ways to reduce the inherent entropy generated by grouping large segments of state-action pairs.

III. Background

A. Preference Learning

In this work, we consider RL environments with continuous state spaces $\mathcal{S} \subseteq \mathbb{R}^n, n \in \mathbb{N}_+$ and action spaces $\mathcal{A} \subseteq \mathbb{R}^v, v \in \mathbb{N}_+$. The main purpose of preference learning is to infer a reward function $\hat{r}_\psi$ from human feedback, which models the true unknown reward function of the environment $r : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$. To learn a preference-based reward function, we build upon the work of Christiano et al. [49]. Preference learning is an interactive learning paradigm that enables humans to give structurally aligned [4] feedback by providing their preference amongst pairs of trajectory segments. Firstly, a trajectory τ is defined as sequences of states and actions, such that $\tau = ((s_0, a_0), (s_1, a_1), \ldots, (s_{k-1}, a_{k-1})) \in (\mathcal{S}, \mathcal{A})^k$, where $k \in \mathbb{N}_+$ represents the length of the trajectory from initial state s_0 to end state, obtained by following policy $\pi_\omega : S \rightarrow A$. Trajectories are collected into a trajectory dataset $\mathcal{D}_\tau = \{\tau_1, ..., \tau_l\}$, where l denotes the number of trajectories collected. From $\mathcal{D}_\tau$ we sample trajectory segments (or just segments for simplicity) σ which are composed of partial trajectories, i.e., $\sigma = ((s_0, a_0), \ldots, (s_m, a_m)) \in (\mathcal{S}, \mathcal{A})^m, m \in \mathbb{N}_+$, where we define the length of a segment by $m, 0 \leq m \leq k$. Subsequently, the collected segments form a segment dataset $\mathcal{D}_\sigma$. Moreover, pairs of segments $(\sigma_1, \sigma_2) \in \mathcal{D}_\sigma$ are randomly sampled, to form queries to be filled by humans for preferences. We denote $\succ$ as a preference operator as in [49]. We assume if a segment σ_1 is preferred over σ_2, i.e., $\sigma_1 \succ \sigma_2$, then the sum of rewards across segment σ_1 is higher than segment σ_2, i.e., $\sum_{(s,a)\in\sigma_1} r(s, a) > \sum_{(s,a)\in\sigma_2} r(s, a)$. A preference μ is a 2-D tuple of the form $\mu \in \{(1, 0), (0, 1), (0.5, 0.5), (0, 0)\}$ specifying which segment was preferred. One segment is preferred over the other in the case $\mu = (1, 0)$ or $\mu = (0, 1)$; if both are of equal value then $\mu = (0.5, 0.5)$; or $\mu = (0, 0)$ if unrelated. Thus, a query q is defined as a triple $q = (\sigma_1, \sigma_2, \mu)$. Queries are stored in a query dataset $\mathcal{D}_q$ which is used to sample and train the reward model $\hat{r}_\psi$. We model the reward function as a preference predictor between segment pairs. The reward model $\hat{r}_\psi$ is represented by a neural network with parameters ψ. The probability of a user preferring a trajectory over another can be seen as the exponential sum of the rewards of the preferred trajectory divided by the total amount of rewards of both trajectories:

$$\hat{P}[\sigma_1 \succ \sigma_2] = \frac{\exp\left(\sum_{\sigma_1} \hat{r}_\psi(s, a)\right)}{\exp\left(\sum_{\sigma_1} \hat{r}_\psi(s, a)\right) + \exp\left(\sum_{\sigma_2} \hat{r}_\psi(s, a)\right)} \quad (1)$$

The reward model $\hat{r}_\psi$ is optimized as a binary classifier, where the loss function is based on the user feedback, measuring the error between the predicted and the actual judgment.

$$\mathcal{L}_{\textbf{pref}}(\hat{r}_\psi) = -\sum_{(\sigma_1,\sigma_2,\mu)\in\mathcal{D}_q} \mu(1)\log\hat{P}[\sigma_1 \succ \sigma_2] + \mu(2)\log\hat{P}[\sigma_2 \succ \sigma_1] \quad (2)$$

IV. POLITE

In this section, we formally present **POLITE**: **P**references **CO**mbined with High**LI**ghts in Reinforcemen**T** L**E**arning, a preference learning algorithm which aims at improving query sample efficiency by asking humans for the causal reason of their preferred choice. While preference learning has plenty of potential, one of many drawbacks is related to the indetermination of which state-action pairs were categorically responsible for a human preference. The indetermination is more generally known in RL as the temporal credit assignment problem [1]. The temporal credit assignment determines which state-action pairs maximize the sum of expected returns. In a more typical RL setting, an agent explores the environment in a sequence of state-action pairs following policy π. Through discounted rewards, credit is trickled down the sequence leading up to the reward. However, in preference learning, once a preference has been attributed, the credit is uniformly associated with the entire trajectory segment. We developed POLITE to address this issue and propose an alternative formulation for preferences, to include relevant additional information. We present our approach threefold from the foundation to the final algorithm:

1) In Sec. IV-A we define highlights, an auxiliary and optional feedback signal in the context of preferences, and introduce highlighted queries.
2) In Sec. IV-B we propose how to use highlights as regularization terms in the learning objective of a reward model.
3) Finally, in Sec. IV-C we outline in great detail our contribution in the context of a preference learning framework.

A. Extending preferences with highlights

A highlight represents a rolling attention window that constitutes part of the preferred segment (see Fig.2). Consider the initial definition of a segment as a sequence of length m in Sec. III-A. Mathematically, we define a highlight h as a subsequence of a preferred segment. More concretely, highlights are subsequences of segments such that, $h_{i,j} = \sigma_{i,j} = ((s_i, a_i), \ldots, (s_j, a_j)) \in (\mathcal{S}, \mathcal{A})^{j-i}, i, j \in \mathbb{N}_+$ with $0 \leq i \leq j \leq m$. The length of a highlight $j - i = L$, where L represents the maximum considered length for a highlight. Additionally, we include a third option in case a human does not wish to highlight any part of the segment. Thus a highlight can be represented by:

$$h = \begin{cases} (s_0, a_0), \ldots, (s_j, a_j), & \text{if } j \leq L, \\ (s_{j-L}, a_{j-L}), \ldots, (s_j, a_j), & \text{if } j > L \\ \emptyset, \text{ if no highlight} \end{cases} \quad (3)$$

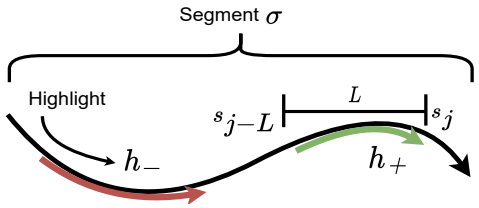

Fig. 2. Representation of highlights within a segment. The segment σ outlined by a curve contains two highlights, a negative in red h^- and a positive in green h^+. Both highlights have length L.

Furthermore, we define h^+ as a positive highlight yielding a positive sum of rewards, i.e., $\sum_{(s,a)\in h^+} r(s,a) > 0$. Likewise, we define h^- as a negative highlight with $\sum_{(s,a)\in h^-} r(s,a) < 0$. When humans are queried for preferences, they have an additional option to signal either a positive highlight h^+, a negative h^-, both, or none. A query as defined in Sec. III-A, constitutes a human preference μ alongside a pair of trajectory segments and is represented by $q = (\sigma_1, \sigma_2, \mu)$. We extend this definition with positive and negative highlights h^+, h^-, and refer to the new quintuple as highlighted queries $hq = (\sigma_1, \sigma_2, \mu, h^+, h^-)$. In POLITE highlighted queries are collected in a dataset denoted by $\mathcal{D}_{hq}$.

B. Regularizing the reward model with POLITE

The strategic application of regularization has been empirically demonstrated as an important factor in shaping state representations through the augmentation of the initial learning objective [54]. Auxiliary tasks refer to secondary tasks that are semi-related to the primary task but provide valuable training signals that can guide the learning of shared representations. These tasks are incorporated into the learning process to leverage the inherent structure and regularities present in the data. Integrating auxiliary tasks alongside the primary task has demonstrated an enhancement in learning and data efficiency [55], [56], [57], [58], [59]. Enhancing the loss function through regularization can be viewed as introducing an inductive bias, thereby constraining the search space over potential solutions. Recent work has revealed that utilizing human natural language to shape representations can lead to the development of human-like inductive biases and behaviors [5], [60]. In POLITE, we create a novel state representation task to reduce the entropy of preferred segments by including causal reasoning from a human teacher. We thereby learn a more informative representation that better distinguishes between high and low-value state-action sequences. By extracting more information from each preference we can effectively reduce the total amount of preferences μ. We specify the new task as additional regularization terms added to Eq.2. The purpose of the regularization is to better shape our reward function in order to detect valuable states and reward them accordingly. To put it simply, the objective is to maximize the reward of positive highlights h^+, and minimize the reward of negative highlights h^-. As in similar works which predict future rewards, we also apply a discount to states leading up to the last state s_j, to underline the importance of the highlight given by humans:

$$\mathcal{L}_+ = -\mathbb{E}_{h^+\sim\mathcal{D}_{hq}}\left[\sum_{l=0}^{L}\lambda^l \hat{r}_\psi(s_{j-l}, a_{j-l})\right] \tag{4}$$

$$\mathcal{L}_- = \mathbb{E}_{h^-\sim\mathcal{D}_{hq}}\left[\sum_{l=0}^{L}\lambda^l \hat{r}_\psi(s_{j-l}, a_{j-l})\right] \tag{5}$$

The final goal is to minimize $\mathcal{L}_+$ (see Eq.4) and $\mathcal{L}_-$ (see Eq.5) while maintaining the baseline preference learning loss $\mathcal{L}_{\textbf{pref}}$ (see Eq.2). We use highlighted query samples hq sampled from $\mathcal{D}_{hq}$ to optimize $\hat{r}_\psi$. Hyperparameters α_+ and α_- serve to weight both regularization terms, and thus the resulting learning objective is of the form:

$$\mathcal{L}_{\textbf{POLITE}} = \mathcal{L}_{\textbf{pref}} + \alpha_+\mathcal{L}_+ + \alpha_-\mathcal{L}_- \tag{6}$$

C. Preference learning with POLITE

Similarly to other preference-based RL methods [49], [52], POLITE (refer to Alg. 1) interleaves policy learning with reward learning. In step A (see Fig.3), policy π_ω interacts with an environment, generating (s_t, a_t, s_{t+1}) alongside estimations of $\hat{r}_\psi(s_t, a_t)$. We collect the resulting transitions $(s_t, a_t, s_{t+1}, \hat{r}_\psi(s_t, a_t))$ (arranged in trajectories) in a provisional buffer which is used to perform gradient descent on π_ω with respect to ω following PPO [61]. After training π_ω, we collect a large number of trajectory segments and store them in $\mathcal{D}_\sigma$. Next, in step B (see Fig.3) we perform a feedback session. We sample $N \in \mathbb{N}_+$ trajectory segments σ to acquire query preferences from humans. Afterward, we prompt humans for their desire of providing either positive h^+ or negative h^- highlights and store them in the highlighted queries dataset $\mathcal{D}_{hq}$. In step C, we iterate our reward model $\hat{r}_\psi$ by performing gradient descent on parameters ψ with $\mathcal{L}_{\textbf{POLITE}}$ as the learning objective. Finally, after obtaining a new iteration of $\hat{r}_\psi$ we resume to step A and repeat the algorithm to convergence.

Algorithm 1: POLITE

$\mathcal{D}_\sigma \leftarrow \emptyset$ $\mathcal{D}_{hq} \leftarrow \emptyset$
$\pi_\omega \leftarrow \text{train}(\pi_\omega, \hat{r}_\psi, env)$ ▷ // Step A
$\mathcal{D}_\sigma \leftarrow \text{sampleSegments}(\pi_\omega)$
while $|\mathcal{D}_{hq}| \leq N$ **do**
  $(\sigma_1, \sigma_2) \leftarrow \text{samplePairs}(\mathcal{D}_\sigma)$ ▷ // Step B
  $\mu \leftarrow \text{collectPreference}(\sigma_1, \sigma_2)$
  $(h_+, h_-) \leftarrow \text{collectHighlight}(\sigma_1, \sigma_2, \mu)$
  $\mathcal{D}_{hq} \leftarrow \mathcal{D}_{hq} \cup (\sigma_1, \sigma_2, \mu, h_+, h_-)$
for *each gradient step* **do**
  Sample minibatch from $\mathcal{D}_{hq}$ ▷ // Step C
  Optimize $\hat{r}_\psi$ $\mathcal{L}_{\textbf{POLITE}}$ with respect to ψ in Eq.(6)
return $\hat{r}_\psi$ ▷ // return $\hat{r}_\psi$

V. Experimental Evaluations

To determine the efficiency of POLITE we run simulated experiments to validate the hypotheses:

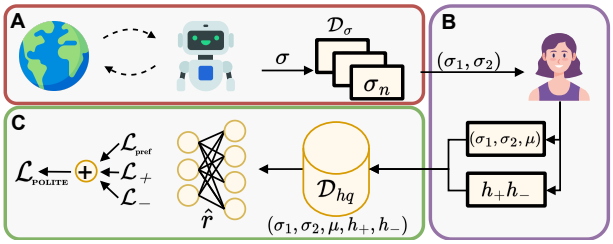

Fig. 3. Macroscopic representation of POLITE. Step A: We train policy π_ω and sample rollouts which are stored in $\mathcal{D}_\sigma$. Step B: We sample trajectory segments σ to query humans and collect both preferences and highlights. Step C: The highlighted queries are collected to form dataset $\mathcal{D}_{hq}$ and update the current reward model $\hat{r}_\psi$.

- **H1**: POLITE has better sample efficiency than the current state-of-the-art, by prompting humans fewer times while gathering additional information per preference given.
- **H2**: POLITE significantly outperforms other approaches by converging faster under the same settings.

We chose to run simulated experiments on two standard MuJoCo environments using OpenAI Gym [62]. We simulate human feedback using a synthetic oracle with access to the true reward function. Preferences of the synthetic oracle work as explained in Sec. III-A, where a segment is preferred over another if the cumulative reward over that segment is higher. We emulate a noisy human teacher by mislabeling 10% of the preferences received.

For POLITE, we extend our oracle to provide additional highlight feedback by keeping track of sequences of state-action pairs with the highest and lowest cumulative rewards for each preferred segment. If the rewards exceed a specific threshold, the oracle will highlight that part of the trajectory.

We evaluate the performance of POLITE against the current state-of-the-art. In addition, we vary the number of queries used to evaluate the sample efficiency. We compare three algorithms:

- POLITE (Our method): using highlighted queries as described in section IV.
- PEBBLE [52]: a state-of-the-art preference learning approach, results are obtained using the author's code.
- Preference Learning (Baseline): Preference learning based on the work of Christiano et al. [49], which can be considered an ablation of POLITE without highlights and exclusively using Eq.2 for the learning objective.

A. Synthetic Benchmark Results

In Walker2d, POLITE using 200 highlighted queries, outperforms baseline preference learning with 800 queries, effectively reducing the query count by 75% (Fig. 4a). Doubling the queries to 400 results in nearly double the performance compared to 800 queries from the baseline. Despite PEBBLE's improvements over the baseline, POLITE outperforms it with half the queries. These results show support for both H1 and H2 as we reach higher convergence with a greater query sample efficiency. Similarly, POLITE outperforms PEBBLE and baseline preference learning using equal queries in Cheetah (Fig. 4b). POLITE achieves near-baseline convergence with 75% fewer preferences, supporting H1 further. With 400 queries POLITE achieves higher convergence than baseline and PEBBLE, corroborating H2.

To understand the effect of combining highlights and preferences we performed an ablation on the Cheetah environment where the agent was given either preferences, highlights, or both. We observe a non-linear synergistic effect when combining preferences and highlights which allow the agent to learn more efficiently (see Fig. 4.c).

B. Evaluation with Real Human Feedback

In this section, we use a simulated robot performing social navigation to analyze how highlights provided from real human feedback effectively build a more informative reward function. To this extent, we validate the following hypotheses:

- H3: POLITE derives a substantially different reward function and policy compared to baseline preference learning.
- H4: POLITE is more effective at identifying and rewarding sparse goal states in the environment through the additional shaping of its reward function.

We run two conditions between-subject, POLITE and baseline preference learning, collecting 400 preferences each. Both conditions start from the same pool of segments and preference pairs generated from the same policy π_ω. The differences stem from how differently humans rate preference pairs and the addition of highlights in POLITE. In total, we recruited 40 (20 for POLITE, 20 for baseline) participants from AMT. All participants were from the United States and were paid a rate equivalent to $10 per hour.

To collect data, we use a social navigation environment introduced in [63]. The environment simulates a social scenario where a Pepper robot navigates through a corridor crowded with moving humans to reach a target destination. In the corridor, one intermittent goal and one end goal are placed, and three humans move around. The task of a participant is to observe pairs of videos with different robot behaviors and pick the preferred one. In the POLITE condition, they are also able to provide highlights.

C. High-level analysis of the resulting reward models

POLITE's curvature shows a more complexly shaped reward function that succeeds at distinguishing the positions of goals and humans (Fig 5a). We can explain the curvature by pointing out the location of the objects in the simulated environment: The first human is located around $z = -10$; the intermittent goal is around $z = -2.5$; two humans are around $z = [4, 15]$; the end goal is around $z = [10, 14]$; a final wall is located at $z = 15$. In contrast, the baseline has a relatively flat distribution throughout the corridor.

The KL-divergence increases (see Fig. 5c) the longer we train both reward functions. An increasing KL-divergence signifies an increase in discrepancy in the reward distribution between POLITE and baseline. The reward spread over the corridor alongside the increasing KL-divergence shows apparent differences in the reward functions supporting H3. POLITE shows its capability to identify and appropriately

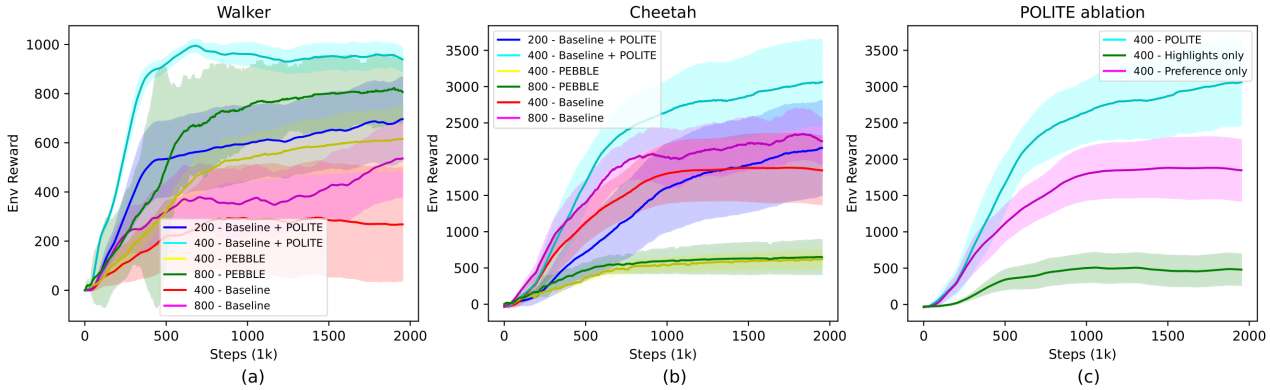

Fig. 4. Learning curves while varying the number of queries used for POLITE, PEBBLE, and baseline. Results are based on the (a) Walker2D and (b) Cheetah environments. The solid lines represent the mean and shaded areas of the standard error. (c) POLITE ablations in Cheetah: reward model trained on preferences, highlights and both.

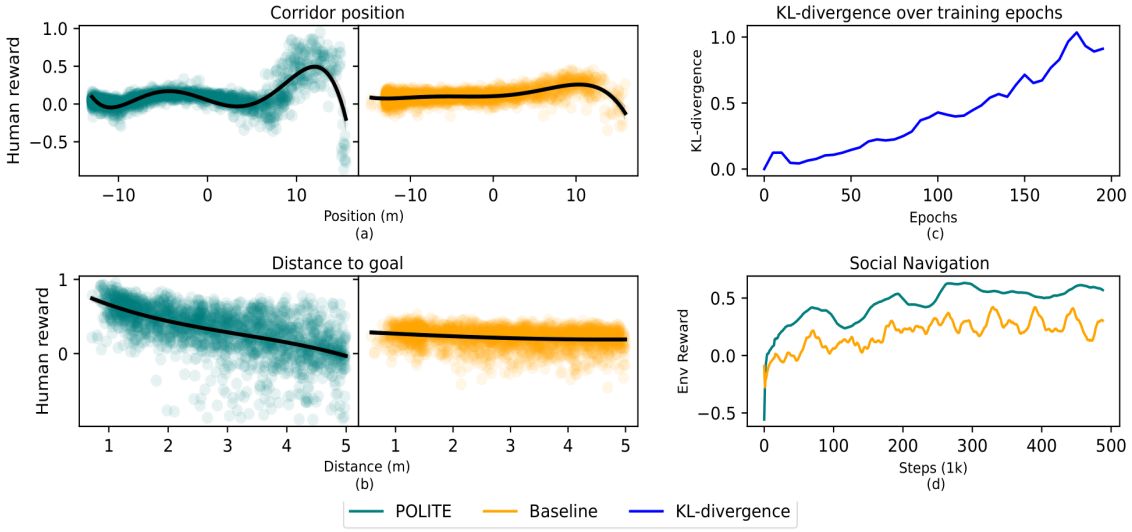

Fig. 5. (a-b) The reward distribution for POLITE and baseline based on location inside the corridor and distance to goal shows clear distinctions. (c) The KL-divergence validates that both reward models diverge. (d) POLITE thereby achieves higher convergence.

reward sparse goal states (Fig 5b). POLITE shows a non-linear relationship between distance to the goal and the corresponding reward received: the reward increases more rapidly the closer to the goal. The additional variance helps better distinguish the real value of highlighted states. POLITE's reward function displays a broader range of values, nearing 1 close to the goal and approaching 0 further away (Fig.5b). This spread is not observed in the baseline, which can hinder precise goal identification. The ability to better detect goals can be verified in Fig. 5d where POLITE outperforms the baseline, validating H4. In the baseline condition, participants didn't prefer any segment 16.5% of the time, while this decreased to 7% in the POLITE condition. Pearson's χ^2-test affirmed a significant relationship between the number of non-preferences and condition ($\chi^2 = 16.5, p < 0.001$). Participants were also willing to provide highlights in the POLITE condition, with positive and negative highlights given 63.44% and 47.31% of the time, respectively.

Finally, to obtain a sense of human effort, we inspect the time participants spent rating the queries. On average, it took people in the highlight condition 24.83 ± 6.48 minutes to rate the 20 queries and read through the tutorial. The baseline condition took an average of 18.93 ± 9.4 minutes. This results in a 31.17% increase in time for the highlight condition.

VI. Conclusions

We show, in simulation and by analyzing real human feedback, that highlights can help shape a more complex reward function, capturing implicit information that would otherwise have to be inferred by over-querying humans, leading to a drastic reduction of queries. Moreover, even with a significant reduction of queries, we still maintain or improve convergence when compared to regular preference learning and PEBBLE. The most interesting results come from Walker2d where POLITE can converge towards a higher reward compared to baseline preference learning with 75% fewer queries. We observe more expressiveness in POLITE's reward function reflected in the reward distributions. The reward function is more in line with the environment's structure by identifying goals and human-occupied areas. In addition, it identifies sparse goals and successfully attributes them to a higher reward. Finally, we observe that humans are more willing to express a preference for the POLITE approach over baseline preference learning.

REFERENCES

[1] C. Wirth, R. Akrour, G. Neumann, J. Fürnkranz *et al.*, "A survey of preference-based reinforcement learning methods," *Journal of Machine Learning Research*, vol. 18, no. 136, pp. 1–46, 2017.

[2] A. Y. Ng and S. Russell, "Algorithms for inverse reinforcement learning," in *Int. Conf. on Machine Learning*, 2000, pp. 663–670.

[3] S. Arora and P. Doshi, "A survey of inverse reinforcement learning: Challenges, methods and progress," *Artificial Intelligence*, vol. 297, p. 103500, 2021.

[4] S. Booth, S. Sharma, S. Chung, J. Shah, and E. L. Glassman, "Revisiting human-robot teaching and learning through the lens of human concept learning," in *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction*, 2022, pp. 147–156.

[5] K. Bhatia, A. Pananjady, P. Bartlett, A. Dragan, and M. J. Wainwright, "Preference learning along multiple criteria: A game-theoretic perspective," *Advances in neural information processing systems*, vol. 33, pp. 7413–7424, 2020.

[6] P. Spirtes, C. N. Glymour, R. Scheines, and D. Heckerman, *Causation, prediction, and search*. MIT press, 2000.

[7] F. Eberhardt, "Introduction to the foundations of causal discovery," *International Journal of Data Science and Analytics*, vol. 3, no. 2, pp. 81–91, 2017.

[8] P. De Haan, D. Jayaraman, and S. Levine, "Causal confusion in imitation learning," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[9] J. Tien, J. Z.-Y. He, Z. Erickson, A. D. Dragan, and D. Brown, "A study of causal confusion in preference-based reward learning," *arXiv preprint arXiv:2204.06601*, 2022.

[10] Q. Li, Z. Peng, H. Wu, L. Feng, and B. Zhou, "Human-ai shared control via policy dissection," *arXiv preprint arXiv:2206.00152*, 2022.

[11] R. Liu, C. Jia, G. Zhang, Z. Zhuang, T. Liu, and S. Vosoughi, "Second thoughts are best: Learning to re-align with human values from text edits," *Advances in Neural Information Processing Systems*, vol. 35, pp. 181–196, 2022.

[12] T. Xie, A. Saran, D. J. Foster, L. Molu, I. Momennejad, N. Jiang, P. Mineiro, and J. Langford, "Interaction-grounded learning with action-inclusive feedback," *arXiv preprint arXiv:2206.08364*, 2022.

[13] A. Najar and M. Chetouani, "Reinforcement learning with human advice: a survey," *Frontiers in Robotics and AI*, vol. 8, p. 584075, 2021.

[14] W. B. Knox and P. Stone, "Interactively shaping agents via human reinforcement: The tamer framework," in *Proceedings of the fifth international conference on Knowledge capture*, 2009, pp. 9–16.

[15] P. Sharma, B. Sundaralingam, V. Blukis, C. Paxton, T. Hermans, A. Torralba, J. Andreas, and D. Fox, "Correcting robot plans with natural language feedback," in *Robotics: Science and Systems (RSS)*, 2022.

[16] Y. Cui and S. Niekum, "Active reward learning from critiques," in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 6907–6914.

[17] W. B. Knox, P. Stone, and C. Breazeal, "Training a robot via human feedback: A case study," in *International Conference on Social Robotics*. Springer, 2013, pp. 460–470.

[18] E. Senft, P. Baxter, J. Kennedy, S. Lemaignan, and T. Belpaeme, "Supervised autonomy for online learning in human-robot interaction," *Pattern Recognition Letters*, vol. 99, pp. 77–86, 2017.

[19] M. Clark-Turner and M. Begum, "Deep reinforcement learning of abstract reasoning from demonstrations," in *2018 13th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2018, pp. 160–168.

[20] R. Choudhury, G. Swamy, D. Hadfield-Menell, and A. D. Dragan, "On the utility of model learning in hri," in *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2019, pp. 317–325.

[21] Y. Cui, Q. Zhang, A. Allievi, P. Stone, S. Niekum, and W. B. Knox, "The empathic framework for task learning from implicit human feedback," *arXiv preprint arXiv:2009.13649*, 2020.

[22] A. Bobu, M. Wiggert, C. Tomlin, and A. D. Dragan, "Feature expansive reward learning: Rethinking human input," in *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, 2021, pp. 216–224.

[23] D. Das, B. Kim, and S. Chernova, "Subgoal-based explanations for unreliable intelligent decision support systems," *arXiv preprint arXiv:2201.04204*, 2022.

[24] M. Hagenow, E. Senft, R. Radwin, M. Gleicher, B. Mutlu, and M. Zinn, "Corrective shared autonomy for addressing task variability," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 3720–3727, 2021.

[25] E. Hedlund, M. Johnson, and M. Gombolay, "The effects of a robot's performance on human teachers for learning from demonstration tasks," in *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, 2021, pp. 207–215.

[26] E. Ratner, D. Hadfield-Menell, and A. Dragan, "Simplifying reward design through divide-and-conquer," in *Proceedings of Robotics: Science and Systems*, Pittsburgh, Pennsylvania, June 2018.

[27] D. R. Scobee and S. S. Sastry, "Maximum likelihood constraint inference for inverse reinforcement learning," *arXiv preprint arXiv:1909.05477*, 2019.

[28] G. Wang, C. Trimbach, J. K. Lee, M. K. Ho, and M. L. Littman, "Teaching a robot tasks of arbitrary complexity via human feedback," in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 2020, pp. 649–657.

[29] L. Chen, R. Paleja, M. Ghuy, and M. Gombolay, "Joint goal and strategy inference across heterogeneous demonstrators via reward network distillation," in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 2020, pp. 659–668.

[30] D. Brown, W. Goo, P. Nagarajan, and S. Niekum, "Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations," in *International conference on machine learning*. PMLR, 2019, pp. 783–792.

[31] D. S. Brown, W. Goo, and S. Niekum, "Better-than-demonstrator imitation learning via automatically-ranked demonstrations," in *Conference on robot learning*. PMLR, 2020, pp. 330–359.

[32] M. Racca, V. Kyrki, and M. Cakmak, "Interactive tuning of robot program parameters via expected divergence maximization," in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 2020, pp. 629–638.

[33] B. D. Argall, S. Chernova, M. Veloso, and B. Browning, "A survey of robot learning from demonstration," *Robotics and autonomous systems*, vol. 57, no. 5, pp. 469–483, 2009.

[34] H. Ravichandar, A. S. Polydoros, S. Chernova, and A. Billard, "Recent advances in robot learning from demonstration," *Annual review of control, robotics, and autonomous systems*, vol. 3, pp. 297–330, 2020.

[35] 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," in *16th Robotics: Science and Systems, RSS 2020*. MIT Press Journals, 2020.

[36] M. L. Schrum, E. Hedlund-Botti, N. Moorman, and M. C. Gombolay, "Mind meld: Personalized meta-learning for robot-centric imitation learning." in *HRI*, 2022, pp. 157–165.

[37] M. Du, O. Y. Lee, S. Nair, and C. Finn, "Play it by ear: Learning skills amidst occlusion through audio-visual imitation learning," in *Robotics: Science and Systems (RSS)*, 2022.

[38] N. Gopalan, N. Moorman, M. Natarajan, and M. Gombolay, "Negative result for learning from demonstration: Challenges for end-users teaching robots with task and motion planning abstractions," in *Robotics: Science and Systems (RSS)*, 2022.

[39] W. B. Knox, S. Hatgis-Kessell, S. Booth, S. Niekum, P. Stone, and A. Allievi, "Models of human preference for learning reward functions," *arXiv preprint arXiv:2206.02231*, 2022.

[40] D. Marta, S. Holk, C. Pek, J. Tumova, and I. Leite, "Aligning human preferences with baseline objectives in reinforcement learning," in *International Conference on Robotics and Automation*, 2023.

[41] H. J. Jeon, S. Milli, and A. Dragan, "Reward-rational (implicit) choice: A unifying formalism for reward learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 4415–4426, 2020.

[42] S. Holk, D. Marta, and I. Leite, "Predilect: Preferences delineated with zero-shot language-based reasoning in reinforcement learning," *arXiv preprint arXiv:2402.15420*, 2024.

[43] A. Bobu, D. R. Scobee, J. F. Fisac, S. S. Sastry, and A. D. Dragan, "Less is more: Rethinking probabilistic models of human behavior," in *Proceedings of the 2020 acm/ieee international conference on human-robot interaction*, 2020, pp. 429–437.

[44] D. Sadigh, A. D. Dragan, S. Sastry, and S. A. Seshia, *Active preference-based learning of reward functions*, 2017.

[45] D. Marta, S. Holk, C. Pek, J. Tumova, and I. Leite, "Variquery: Vae segment-based active learning for query selection in preference-based reinforcement learning," in *2023 IEEE/RSJ International Conference*

on Intelligent Robots and Systems (IROS). IEEE, 2023, pp. 7878–7885.

[46] E. Bıyık, N. Huynh, M. J. Kochenderfer, and D. Sadigh, "Active preference-based gaussian process regression for reward learning," in *Robotics: Science and Systems (RSS)*, 2020.

[47] E. Bıyık, D. P. Losey, M. Palan, N. C. Landolfi, G. Shevchuk, and D. Sadigh, "Learning reward functions from diverse sources of human feedback: Optimally integrating demonstrations and preferences," *The International Journal of Robotics Research*, vol. 41, no. 1, pp. 45–67, 2022.

[48] K. Amin, N. Jiang, and S. Singh, "Repeated inverse reinforcement learning," *Advances in neural information processing systems*, vol. 30, 2017.

[49] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei, "Deep reinforcement learning from human preferences," *Advances in neural information processing systems*, vol. 30, 2017.

[50] E. Båvenstrand and J. Berggren, "Performance evaluation of imitation learning algorithms with human experts," 2019.

[51] B. Ibarz, J. Leike, T. Pohlen, G. Irving, S. Legg, and D. Amodei, "Reward learning from human preferences and demonstrations in atari," *Advances in neural information processing systems*, vol. 31, 2018.

[52] K. Lee, L. Smith, and P. Abbeel, "Pebble: Feedback-efficient interactive reinforcement learning via relabeling experience and unsupervised pre-training," *arXiv preprint arXiv:2106.05091*, 2021.

[53] J. Hejna and D. Sadigh, "Few-shot preference learning for human-in-the-loop rl," *arXiv preprint arXiv:2212.03363*, 2022.

[54] T. De Bruin, J. Kober, K. Tuyls, and R. Babuška, "Integrating state representation learning into deep reinforcement learning," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 1394–1401, 2018.

[55] M. Jaderberg, V. Mnih, W. M. Czarnecki, T. Schaul, J. Z. Leibo, D. Silver, and K. Kavukcuoglu, "Reinforcement learning with unsupervised auxiliary tasks," *arXiv preprint arXiv:1611.05397*, 2016.

[56] P. Mirowski, R. Pascanu, F. Viola, H. Soyer, A. J. Ballard, A. Banino, M. Denil, R. Goroshin, L. Sifre, K. Kavukcuoglu *et al.*, "Learning to navigate in complex environments," *arXiv preprint arXiv:1611.03673*, 2016.

[57] J. Matas, S. James, and A. J. Davison, "Sim-to-real reinforcement learning for deformable object manipulation," in *Conference on Robot Learning*. PMLR, 2018, pp. 734–743.

[58] E. Shelhamer, P. Mahmoudieh, M. Argus, and T. Darrell, "Loss is its own reward: Self-supervision for reinforcement learning," *arXiv preprint arXiv:1612.07307*, 2016.

[59] L. Pinto and A. Gupta, "Learning to push by grasping: Using multiple tasks for effective learning," in *2017 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2017, pp. 2161–2168.

[60] S. Kumar, C. G. Correa, I. Dasgupta, R. Marjieh, M. Y. Hu, R. Hawkins, J. D. Cohen, K. Narasimhan, T. Griffiths *et al.*, "Using natural language and program abstractions to instill human inductive biases in machines," *Advances in Neural Information Processing Systems*, vol. 35, pp. 167–180, 2022.

[61] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[62] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, "Openai gym," *arXiv preprint arXiv:1606.01540*, 2016.

[63] D. Marta, C. Pek, G. I. Melsión, J. Tumova, and I. Leite, "Human-feedback shield synthesis for perceived safety in deep reinforcement learning," *IEEE Robotics and Automation Letters*, vol. 7, no. 1, pp. 406–413, 2021.