2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 23-27, 2022, Kyoto, Japan

2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) | 978-1-6654-7927-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/IROS47612.2022.9981076

Task Decoupling in Preference-based Reinforcement Learning for Personalized Human-Robot Interaction

Mingjiang Liu and Chunlin Chen

Abstract— **Intelligent robots designed to interact with humans in the real world need to adapt to the preferences of different individuals. Preference-based reinforcement learning (RL) has shown great potential for teaching robots to learn personalized behaviors from interacting with humans without a meticulous, hand-crafted reward function, replaced by learning reward based on a human's preferences between two robot trajectories. However, poor feedback efficiency and poor exploration in the state and reward spaces make current preference-based RL algorithms perform poorly in complex interactive tasks. To improve the performance of preference-based RL, we incorporate prior knowledge of the task into preference-based RL. Specifically, we decouple the task from preference in human-robot interaction. We utilize a sketchy task reward derived from task priori to instruct robots to conduct more effective task exploration. Then a learned reward from preference-based RL is used to optimize the robot's policy to align with human preferences. In addition, these two parts are combined organically via reward shaping. The experimental results show that our method is a practical and effective solution for personalized human-robot interaction. Code is available at `https://github.com/Wenminggong/PbRL_for_PHRI`.**

I. INTRODUCTION

New frontiers in artificial intelligence and robotics hold the potential to realize human and robot symbiosis. In human social environments, robots aim to assist humans to live safer, easier, and more independent in various contents [1]. To achieve this goal, robots are developed to understand and interact with humans in long-term, real-world settings, which poses many challenges to learn from and for the diversity of humanity. Human learning, development, and care all follow nonlinear trajectories unique to each individual. Robots must have personalized skills to adapt to different users.

Interactive machine learning offers a solution to personalized human-robot interaction [2]. Reinforcement learning (RL) is one of the representative interactive machine learning methods. Benefit from the high-capacity function approximate ability of deep learning, deep reinforcement learning (DRL) has been applied in a range of challenging domains, from games (e.g., Go [3] and Atari [4]) to robotics (e.g., Legged Robots [5]). However, the success of these methods depends on hand-crafted reward functions. Unfortunately, many tasks involve goals that are complex and poorly defined, and an imprecise reward function will lead to reward hacking, that is the agent may maximize the defined reward without performing the intended goal. What is worse, the preferences of users can not be predicted in human-robot interaction.

An alternative to avoid hand-crafted reward is preference-based RL [6]–[8], which is a paradigm for learning from nonnumerical feedback in sequential domains [9]. Instead of maximizing long-term hand-crafted rewards, the agent uses qualitative feedback, usually in the form of human preferences between two robot trajectories, to learn the desired strategy that matches human preferences. Learning a reward function from human feedback and then optimizing that reward function is one of the representative approaches of preference-based RL [8], which has been scaled to off-policy RL to improve sample efficiency [10]. However, these preference-based RL algorithms are very inefficient since they attempt to learn a continuous reward function from binary human feedback. It is hard to obtain a good state-space coverage with random exploration guided by human preferences. Hence, it is intractable to train a robot to perform complex interactive tasks and conform to human preferences just by using preference-based RL.

In human-robot interaction, although the preferences of humans can not be predicted, it is practical to obtain some prior knowledge about the interactive task. For example, how much force should a robot use to shake hands with a human is uncertain, but we do know how to make the robot perform a handshake with a human. Inspired by this observation, we incorporate the prior knowledge of the task into the preference-based RL to implement personalized human-robot interaction. Specifically, we decouple the task from preference in human-robot interaction. We utilize a sketchy reward function derived from the prior knowledge of the task to instruct the robot to conduct more effective task exploration. Then a learned reward function from preference-based RL is used to optimize the robot strategy to align with human preferences. Our experiments demonstrate that the proposed method significantly improves the performance of preference-based RL methods (e.g., PrefPPO and PEBBLE [10]) on a complex human-robot interaction task. In addition, the proposed method is more robust to irrational human feedback. The main contributions of this paper are twofold:

- We decouple the task from preference and incorporate the prior knowledge of the task into preference-based RL to improve its performance.
- Sufficient experimental results show that the proposed method is an effective solution for personalized human-robot interaction.

This work was supported in part by the National Key Research and Development Program of China (No.2018AAA0101100) and the National Natural Science Foundation of China (No.62073160).

Mingjiang Liu and Chunlin Chen are with the Department of Control and Systems Engineering, School of Management and Engineering, Nanjing University, Nanjing, China.

Corresponding author: Chunlin Chen, `clchen@nju.edu.cn`

The rest of the paper is organized as follows. After discussing the related work in Section II, we systematically introduce the proposed method in Section III. In Section IV, the experiments involving human-robot interaction are implemented and the results demonstrate the success of the proposed method. Conclusions are given in Section V.

II. RELATED WORK

A. Personalized Human-Robot Interaction

Different individuals react differently to the same robot behavior which reflects the personalization of humans. To satisfy the personalized needs of humans, robots should take unique information from every individual as input. The unique information can be a current dynamic model of the environment, with which we can plan an optimal policy for robots. For example, personalized collaborative plans were implemented for robot-assisted dressing via optimization [11]. Considering personalized machine learning methods, we can cluster users according to their characteristics, and then train a separate machine learning model for each cluster [12] or utilize multitask learning techniques [13]. However, these methods require expert knowledge of the specific field. Recently, RL shows its potential for personalized human-robot interaction, which has been used to optimize parameters of the interaction model to learn personalized proxemics [14] and to help robots select an appropriate action for tea-making [15]. The reward function reflects individual preferences, which is the unique information of individuals for robots. A meticulous design of the reward function is the key to the success of this method. Instead of hand-crafted reward, learning from pairwise preferences of human was used to optimize personalized exoskeleton gait [16], [17].

B. Learning from Pairwise Human Feedback

Several works have successfully utilized pairwise preferences feedback from humans to train agents [6], [7]. Learning a reward function from human feedback and then optimizing that reward function is one of the potential methods [18]. Following this basic approach, preference-based reinforcement learning was scaled to more complex domains including Atari games and robotics tasks in MuJoCo by utilizing modern deep learning techniques [8]. In the real world, human feedback is very expensive, hence poor sample efficiency and feedback efficiency are the main problems in preference-based RL. Recently, an off-policy preference-based RL algorithm was proposed to improve both sample efficiency and feedback efficiency via relabeling history experience and unsupervised pre-training [10]. In addition, incorporating expert demonstrations and pairwise preferences has been proved to be an effective way to improve the efficiency of preference-based RL [19], [20]. An efficient exploration method was proposed by incorporating uncertainty from the reward function [21]. To reduce the need for human feedback without sacrificing performance, several works were presented, and they trained a preference predictor to provide pseudo preference labels [22]–[24].

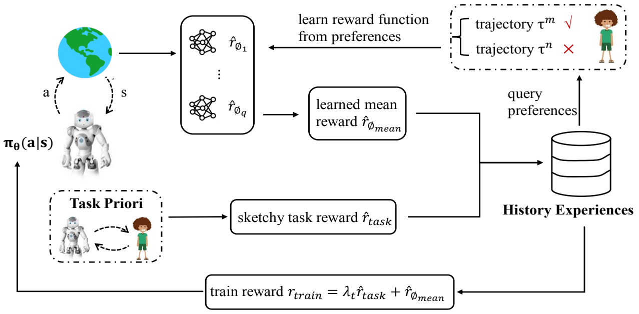
learn reward function from preferences
trajectory τ^m √
trajectory τ^n ×
$\hat{r}_{\emptyset_1}$
$\hat{r}_{\emptyset_q}$
a
s
$\pi_\theta(a|s)$
learned mean reward $\hat{r}_{\emptyset_{mean}}$
query preferences
History Experiences
Task Priori
sketchy task reward $\hat{r}_{task}$
train reward $r_{train} = \lambda_t \hat{r}_{task} + \hat{r}_{\emptyset_{mean}}$

Fig. 1. Illustration of our method. Guided by the sum of the sketchy task reward $\hat{r}_{task}$ and the learned mean reward $\hat{r}_{\phi_{mean}}$, the robot optimizes its interactive policy to align with human preferences.

III. OUR METHOD

A. Preference-based Reinforcement Learning

Reinforcement learning is a framework where an agent learns from interaction with environments [25]. At each timestep t, the agent observes a state s_t from the environment and chooses an action a_t based on its current policy $\pi(a_t|s_t)$. In conventional RL framework, the environment gives a numerical reward $r(s_t, a_t)$ and the goal of the agent is to maximize the discounted return $G_t = \sum_{k=0}^{T} \gamma^k r(s_{t+k}, a_{t+k})$.

However, the preferences of humans can not be predicted in human-robot interaction, there will be no hand-crafted numerical reward existence. Hence, we consider preference-based RL which replaces the numerical reward with the preferences between two robot behavior segments [9]. Formally, a behavior segment σ is a sequence of observations and actions $\{(s_k, a_k), (s_{k+1}, a_{k+1}), \dots, (s_{k+H}, a_{k+H})\}$. In human-robot interaction, the robot needs to perform a complete behavior to serve humans and then gets the preferences feedback from the human. Instead of using a short behavior segment σ, we consider using the whole behavior trajectory $\tau = \{(s_1, a_1), (s_2, a_2), \dots, (s_T, a_T)\}$ to query human preferences in this paper. The robot demonstrates a pair of behavior trajectories (τ^0, τ^1) to interact with human, human indicates which trajectory is preferred (i.e., $y = (\tau^0 \succ \tau^1)$ or $(\tau^1 \succ \tau^0)$), that the two trajectories are equally preferred $y = (\tau^0 = \tau^1)$, or that the two trajectories are incomparable (i.e., discarding this query). Each preference feedback is stored in a dataset $\mathcal{D}$ as a triple (τ^0, τ^1, y). To overcome the lack of numerical reward, preference-based RL utilizes human preference feedback to learn a reward function. After that, the learned reward function is used to guide the robot to optimize the policy [8], [10], [20], [21], [24].

The reward function $\hat{r}_\phi$ and the policy π_θ are both parametrized by deep neural networks. These networks are updated by three processes:

- *step 1*: The reward function $\hat{r}_\phi$ is optimized via supervised learning to fit the preference feedback received from humans.
- *step 2*: The policy π_θ interacts with the environment to collect a set of trajectories $\{\tau^1, \tau^2, \dots, \tau^i\}$ and it is updated via conventional RL algorithms to optimize the

sum of the learned reward $\hat{r}_\phi$.

- *step 3*: The robot selects pairs of trajectories (τ^m, τ^n) from the collected trajectory dataset, and performs them to query human preferences.

1) Reward learning from human preferences: Intuitively, the trajectories with more desirable behaviors should have higher cumulative rewards. The learned reward function needs to satisfy this criterion. Following the Bradley-Terry model [26], we model the preference predictor of a pair of trajectories based on the learned reward function $\hat{r}_\phi$ as

$$P_\phi[\tau^i \succ \tau^j] = \frac{\exp \sum_{t=0}^{T} \hat{r}_\phi(s_t^i, a_t^i)}{\sum_{k \in \{i,j\}} \exp \sum_{t=0}^{T} \hat{r}_\phi(s_t^k, a_t^k)}, \quad (1)$$

where $\tau^i \succ \tau^j$ denotes the event that the trajectory τ^i is preferable to the trajectory τ^j. To align preference predictor with the preference feedback received from human, preference-based RL algorithms translate updating reward function to a binary classification problem. Specifically, the reward function $\hat{r}_\phi$ parametrized by ϕ is updated to minimize the following cross-entropy loss:

$$\mathcal{L}_{Reward} = -\mathop{\mathbb{E}}_{(\tau^i, \tau^j, y) \sim \mathcal{D}} \Big[\mathbb{I}\{y = (\tau^i \succ \tau^j)\} \log P_\phi[\tau^i \succ \tau^j] + \mathbb{I}\{y = (\tau^j \succ \tau^i)\} \log P_\phi[\tau^j \succ \tau^i] \Big], \quad (2)$$

where $\mathcal{D}$ is the preference feedback dataset.

2) Optimizing the policy: Once the reward function $\hat{r}_\phi$ has been optimized from human preferences, there is left a conventional RL problem. Generally, we can train robots with any existing RL algorithms. Depending on whether the target policy is the same as the behavior policy, RL algorithms can be divided into two categories, i.e., on-policy algorithms (e.g., PPO [27]) and off-policy algorithms (e.g., SAC [28]). For on-policy algorithms, we just need to replace the hand-crafted reward function with the learned reward function $\hat{r}_\phi$ [8]. However, this method will not work well in off-policy algorithms. A caveat is that the reward function $\hat{r}_\phi$ may be non-stationary because we update it during training. In off-policy algorithms, previous experiences in the replay buffer are labeled with the previously learned reward function. As a result, the learning process of off-policy algorithms will be unstable. To handle this issue, we can relabel all of the robot's experience every time we update the reward function $\hat{r}_\phi$ [10]. Compared to on-policy algorithms, off-policy algorithms are more sample-efficiency via reusing past experiences. To validate our approach more fully, we construct our experiments on both on-policy and off-policy algorithms in this paper.

3) Selecting queries: The goal of preference-based RL is to train an agent to perform behaviors desirable to a human using as little preference feedback as possible. During training, all history trajectories are stored in an annotation buffer $\mathcal{B}$, and the robot should generate N_{query} pairs of trajectories to query human preferences at each feedback session. What query strategy should the robot take to reduce the burden on humans? Uniform sampling (i.e., picking N_{query} pairs of trajectories uniformly at random from the buffer $\mathcal{B}$) is the simplest method, while it is not efficient enough in complex domains. Ensemble-based sampling is an effective query strategy to solicit preferences to maximize the information received [8], [10], [29]. In this paper, we use the ensemble-based sampling strategy to select queries. We fit q reward function $\{\hat{r}_{\phi_1}, \hat{r}_{\phi_2}, \ldots, \hat{r}_{\phi_q}\}$ as an ensemble, and each reward function is trained on $|\mathcal{D}|$ triples sampled from preference feedback dataset $\mathcal{D}$ with replacement. We take their average as the ensemble result to support policy optimization. For selecting queries, the robot first generate the initial batch of N_{init} pairs of trajectories $\mathcal{G}_{init}$ uniformly at random from the buffer $\mathcal{B}$, then using each reward predictor in our ensemble to predict preferences $\{P_{\phi_1}[\tau^m \succ \tau^n], \ldots, P_{\phi_q}[\tau^m \succ \tau^n]\}$ from each pair (τ^m, τ^n). Finally, selecting N_{query} ($N_{query} \leqslant N_{init}$) pairs of trajectories for which the predictions have the highest variance across ensemble members (i.e., $Var\{P_{\phi_1}[\tau^m \succ \tau^n], \ldots, P_{\phi_q}[\tau^m \succ \tau^n]\}$) to query human preferences.

B. Decoupling Task from Preference

Generally, binary preference feedback is less informative than numerical rewards. Hence, preference-based RL algorithms are more inefficient than conventional RL algorithms with numerical rewards. The credit assignment of the reward function is a tough challenge in long-episode human-robot interaction. Besides, it is hard for preference-based RL to obtain good state- and action-space coverage with random exploration, especially in high-dimensional robotics tasks. To improve the performance of preference-based RL in human-robot interaction, we incorporate the prior knowledge of the task into preference-based RL. The key idea of our method is to decouple the task from preference in human-robot interaction. We define a sketchy reward function to communicate the desirable task behavior, and the sketchy reward function is used to instruct the robot to conduct more effective task exploration. After the robot has mastered the task knowledge, we utilize the learned reward function from preference-based RL to optimize the robot strategy to align human preferences in reduced strategy space. Specifically, we incorporate the defined sketchy task reward into preference-based via reward shaping. The framework of our method is shown in Fig. 1.

1) Hypothesizes of task and preference in personalized human-robot interaction: In personalized human-robot interaction, the robot not only needs to perform interactive behavior with human successfully but also needs to find a personalized strategy to align human preferences. For example, in a handshake task, the robot not only needs to perform a physical act of shaking hands with a human but also needs to decide to use how much force to execute it. In this paper, we call these two goals as the goal of task and the goal of preference. Furthermore, the goal of task and the goal of preference are often coupled in the real world. To achieve the goal of preference, the robot should utilize preference-based RL to learn personalized skills from interacting with humans. The reason conventional RL methods are not applicable is that the preferences of humans can not be predicted in advance. However, it is difficult for preference-based RL to train a robot to achieve both goals simultaneously. Although the preferences of humans can not be predicted, we can obtain some prior knowledge about the task. For example, we know that the robot's hand should move close to the hand of a human in a handshake task. Inspired by this, we decouple the task from preference in human-robot interaction. According to the prior knowledge of the task, we define a sketchy reward function $\hat{r}_{task}$ to communicate the desirable task behavior. In conclusion, our method is based on these two hypothesizes:

- *Hypothesis of preference:* the preferences of humans can not be predicted and the robot should learn personalized skills from interacting with humans.
- *Hypothesis of task:* we can obtain some prior knowledge of the task and we can define a sketchy reward function $\hat{r}_{task}$ according to the prior knowledge.

2) Decoupled preference-based RL with task priori: In preference-based RL, once the reward functions $\{\hat{r}_{\phi_l}\}_{l=1}^{q}$ are optimized from human preferences, the robot is usually trained with conventional RL algorithms guided by the learned mean reward function:

$$\hat{r}_{\phi_{mean}} = \frac{1}{q}\sum_{l=1}^{q}\hat{r}_{\phi_l}. \tag{3}$$

To incorporate the prior knowledge of the task into preference-based RL, we combine the defined sketchy task reward $\hat{r}_{task}$ with the learned mean reward $\hat{r}_{\phi_{mean}}$, and train a robot's policy to optimize the sum of these two reward functions:

$$r_{train}(s_t, a_t) = \lambda_t \hat{r}_{task}(s_t, a_t) + \hat{r}_{\phi_{mean}}(s_t, a_t), \tag{4}$$

where $\lambda_t \geqslant 0$ is the task reward rate that can be used to determine the trade-off between the goal of task and the goal of preference at current training timestep t. However, the hand-crafted reward function $\hat{r}_{task}$ is imprecise, it can be just viewed as an approximation of the ground truth task reward. As a result, if λ_t remains to be a large number throughout the whole training process, the introduced task reward may bias the desirable strategy. To avoid this situation, we use a reward rate that decreases over training time:

$$\lambda_t = \frac{T-t}{T}\lambda_0, \tag{5}$$

where T is the maximum training timestep, and λ_0 is the initial task reward rate which is also a hyperparameter in this paper. The full procedure of our method is summarized in *Algorithm* 1.

IV. Experiments

To evaluate the performance of the proposed method on personalized human-robot interaction, we design our experiments on a physical simulation environment [30], particularly answering the following questions:

- *Q1:* Can the proposed method improve the performance of state-of-the-art preference-based RL methods in personalized human-robot interaction via decoupling task from preference?
- *Q2:* Does the proposed method have high robustness to utilize imperfect human feedback?
- *Q3:* How does the reward rate λ_t influence the performance?

Algorithm 1: Decoupled Preference-based RL with Task Priori

```
Initialize the robot's policy $\pi_\theta$ and the reward
  functions $\{\hat{r}_{\phi_l}\}_{l=1}^{q}$
Define a sketchy task reward $\hat{r}_{task}$ according to the
  prior knowledge of the task
Initialize the frequency of human feedback $K$ and
  the preference feedback dataset $\mathcal{D} \leftarrow \emptyset$
for each training timestep $t$ do
    //INTERACTION WITH ENVIRONMENT
    Collect $s_{t+1}$ by excuting $a_t \sim \pi_\theta(a_t|s_t)$
    Store transition $\{s_t, a_t, s_{t+1}, \hat{r}_{task}(s_t, a_t)\}$ in $\mathcal{B}$
    // REWARD LEARNING
    if $t\%K == 0$ then
        Selecte $N_{query}$ pairs of trajectories from
          history experience buffer $\mathcal{B}$ using
          Ensemble-based sampling
        Query human preferences $y$ and store them
          $\mathcal{D} \leftarrow \mathcal{D} \cup \{(\tau^m, \tau^n, y)\}_{i=1}^{N_{query}}$
        Update reward functions $\{\hat{r}_{\phi_l}\}_{l=1}^{q}$ according
          to Equation (2)
    end
    // POLICY OPTIMIZATION
    for each gradient step do
        Sample minibatch from history buffer
          $\{s_j, a_j, \hat{r}_{task}(s_j, a_j)\}_{j=1}^{N}$
        Label rewards $\{\hat{r}_{\phi_{mean}}(s_j, a_j)\}_{j=1}^{N}$, update $\lambda_t$
          and compute $r_{train}$
        Train policy with reward $r_{train}$
    end
end
```

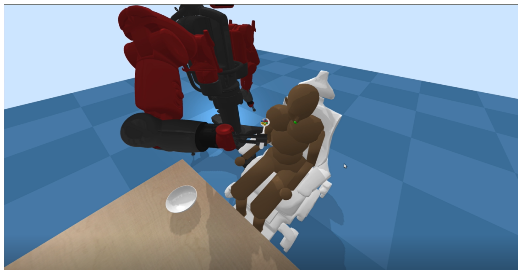

Fig. 2. The feeding task in Assistive-Gym.

A. Setups

We evaluate our method on a feeding task (as shown in Fig. 2) from Assistive-Gym, a physics simulation framework for assistive robotics [30]. In the simulation environment, a Baxter robot holds a spoon with small spheres representing food on the spoon and it must bring this food

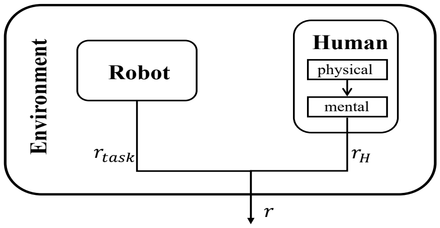

Fig. 3. An environmental model with human preferences. In Assistive-Gym, the environment can provide a human preference reward r_H via estimating the mental state of humans which is affected by the physical state of humans. The output reward r is a combination of human preference reward r_H and robot's task reward r_R.

to a human's mouth without spilling it. Furthermore, the behavior performed by the robot should align with human preferences. For example, the robot should not apply large forces to the human body, or the robot should take its actions slowly and interpretably. To model human preferences, a novel RL environmental model [30] was presented as shown in Fig. 3. At each time step, the environment computes a human preference reward r_H based on how well the robot is satisfying the human's preferences. Then combining this human preference reward r_H with the robot's task reward r_{task} to output an overall reward r. Specifically, the human preference reward r_H is defined as

$$\begin{aligned} r_H = \omega \odot [&C_d(s), C_e(s), C_v(s), C_f(s), \\ &C_{hf}(s), C_{fd}(s), C_{fdv}(s)], \end{aligned} \tag{6}$$

where ω represents a vector of weights for each preference, and $C_{\cdot}(s)$ represents the cost of deviating from human preference in the current state s. In this paper, we use $\omega = [0.5, 0.5, 0.25, 0.3, 0.1, 2.5, 10.0]$, and we define the penalty terms as

- $C_d(s)$: cost for long distance from robot's end effector to the target assistance location (e.g., human mouth in our paper).
- $C_e(s)$: reward for successfully feeding food to the human mouth.
- $C_v(s)$: cost for high robot's end effector velocities.
- $C_f(s)$: applying force away from the target assistance location.
- $C_{hf}(s)$: applying high forces near the target.
- $C_{fd}(s)$: spilling food on the human.
- $C_{fdv}(s)$: food entering mouth at high velocities.

To verify the efficacy of preference-based RL to learn from non-numerical feedback, we assume that the robot can not observe the ground truth reward r. Instead, similar to prior works [8], [10], [20], the robot learns to interact with humans by getting preference feedback from a scripted human teacher. The scripted human teacher can provide preferences between robot's trajectories according to the true, underlying reward r. Since the preferences of the scripted human teacher exactly reflect the ground truth reward of the environment, we can evaluate the performance of our method by measuring the true average return. Besides, we can get the underlying human preference reward r_H to evaluate whether our method aligns with human preferences.

For our method, we incorporate the prior knowledge of the task into preference-based RL. In the feeding task, the goal of task of the robot is to feed food to a human's mouth using a spoon. We define the sketchy task reward as

$$\hat{r}_{task} = -\|d\|_2, \tag{7}$$

where d is the distance from the spoon to the human mouth. We remark that our method can be combined with any preference-based RL algorithms by replacing the policy optimization procedure of its backbone method. To validate our method more comprehensively, we choose state-of-the-art on-policy algorithms (e.g., PrefPPO [10]) and off-policy algorithms (e.g., PEBBLE [10]) as our backbone algorithms in this paper.

B. Performance of our method

To evaluate the performance of our method, the comparison with various other methods is presented, including conventional RL methods (i.e., RL with true reward and RL with sketchy task reward) and state-of-the-art preference-based RL methods (i.e., PrefPPO and PEBBLE). The settings of all methods are listed as follows:

- *RL with true reward*: The robot can get ground truth reward r from the environment, and we utilize conventional RL algorithms (i.e., PPO and SAC) to train the robot to maximize the expected ground truth return.
- *RL with sketchy task reward*: The robot can only get the sketchy task reward $\hat{r}_{task}$ from the environment, and we utilize PPO and SAC to train the robot to maximize the sketchy task return.
- *Preference-based RL*: The robot learns a reward function according to the preference feedback from the scripted human teacher, and then uses the learned reward function to optimize its policy. Specifically, we implement PrefPPO and PEBBLE in this paper.
- *Decoupled prefrence-based RL (our method)*: We define a sketchy task reward $\hat{r}_{task}$ and combine it with the learned reward function from preference-based RL. Choosing PrefPPO and PEBBLE as our backbone algorithms, we propose two decoupled preference-based RL algorithms respectively, called Decoupled PrefPPO and Decoupled PEBBLE.

We compare our method with the baseline of RL with true reward, our aim here is not to show that our method can outperform RL with the true reward, but rather to do nearly as well. For all preference-based algorithms, we set $q = 3$ and train an ensemble of reward functions according to Equation 2. 50000 queries are generated during the training process. In particular, we consider $\lambda_0 = 1.0$ and $\lambda_0 = 15.0$ for Decoupled PrefPPO and Decoupled PEBBLE. The experimental results with mean and standard deviation across 3 runs are reported in Fig. 4 and Fig. 5.

As shown in Fig. 4 and Fig. 5, our methods can achieve the performance almost as well as RL with the true reward in the feeding task from Assistive-Gym, although our methods

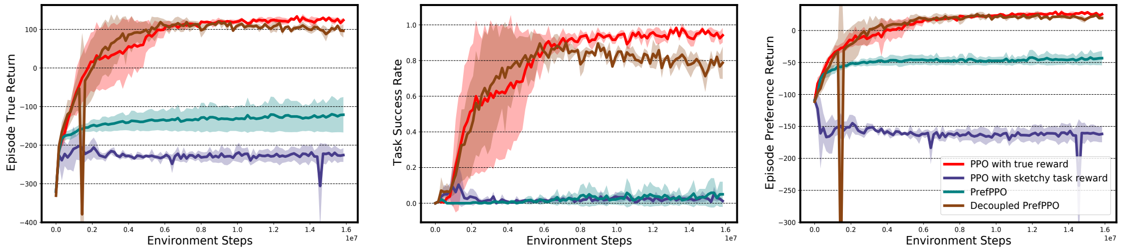

Fig. 4. The learning curves of on-policy case on feeding task. The experimental results are measured on the ground truth return, task success rate, and ground truth preference return. The solid line and shaded regions represent the mean and standard deviation, respectively across 3 runs.

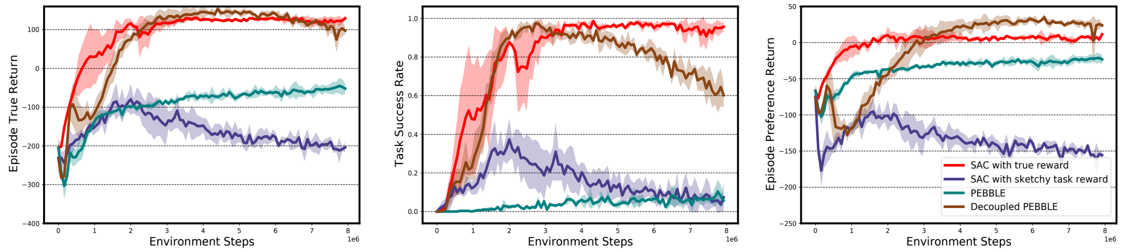

Fig. 5. The learning curves of off-policy case on feeding task. The experimental results are measured on the ground truth return, task success rate, and ground truth preference return. The solid line and shaded regions represent the mean and standard deviation, respectively across 3 runs.

have no access to numerical reward information. Compared to preference-based RL algorithms, we decouple task from preference in personalized human-robot interaction, and utilize a sketchy task reward and a learned reward function to guide the robot to explore well for the task and human preferences respectively. The results show that our methods can improve the performance of preference-based RL on complex interactive tasks and have been proved to be a successful attempt for personalized human-robot interaction. Besides, although a sketchy task reward seems useless for conventional RL, it provides great help in our methods.

C. Robustness of our method

In the real world, many possible irrationalities are affecting a human's preferences feedback. Hence, it is unrealistic to evaluate our method using an ideal scripted human teacher. We consider more realistic models of scripted human teachers designed by [29]:

- *Stochastic preference model*: A stochastic model is defined to support noisy preferences from human:

$$P[\tau^m \succ \tau^n; \beta, \gamma_{my}] = \exp(\beta \sum_{t=1}^{T} \gamma_{my}^{T-t} r(s_t^m, a_t^m)) / \big(\exp(\beta \sum_{t=1}^{T} \gamma_{my}^{T-t} r(s_t^m, a_t^m)) + \exp(\beta \sum_{t=1}^{T} \gamma_{my}^{T-t} r(s_t^n, a_t^n))\big) \tag{8}$$

  where $\gamma_{my} \in (0, 1]$ is a discount factor to model myopic behavior, β is a rationality constant, and $P[\tau^m \succ \tau^n]$ denotes the probability of preferring trajectory m than trajectory n.
- *Myopic behavior*: Humans are sometimes myopic, hence a human teacher may remember and focus on the behavior at the end of the trajectory he watched. The myopic behavior is modeled by introducing a weighted sum of rewards with a discount factor γ_{my} in Equation 8.
- *Skipping queries*: If both trajectories are not desired behaviors, a human would like to mark them as incomparable and discard this query. This behavior is modeled by skipping a query if the sums of rewards over both trajectories are smaller than skip threshold, i.e., $\max_{k \in \{m,n\}} (\sum_t r(s_t^k, a_t^k) - R_{min}) < (R_{avg}(\pi_t) - R_{min}) \delta_{skip}$, where $R_{avg}(\pi_t)$ is the average return of current policy π_t, and R_{min} is the minimum return.
- *Equally preferable*: If two trajectories are equally good, a human would like to mark them as equally preferable. Hence, if two trajectories have similar sum of rewards (e.g., $|\sum_t r(s_t^m, a_t^m) - \sum_t r(s_t^n, a_t^n)| < (R_{avg}(\pi_t) - R_{min}) \delta_{equal}$), an uniform distribution $(0.5, 0.5)$ should be provided.
- *Making a mistake*: Humans may make errors sometimes. To reflect this, the preferences are flipped with the probability of ϵ.

To evaluate the robustness, we implement our method using various realistic scripted human teachers, their properties are listed in Table. I. We consider one modification to the oracle scripted human teacher at a time. The experimental results are shown in Fig. 6 and Fig. 7. In both on-policy and off-policy cases, our methods can still achieve good performance, although the scripted human teachers are imperfect. This shows that our methods are robust to irrational human

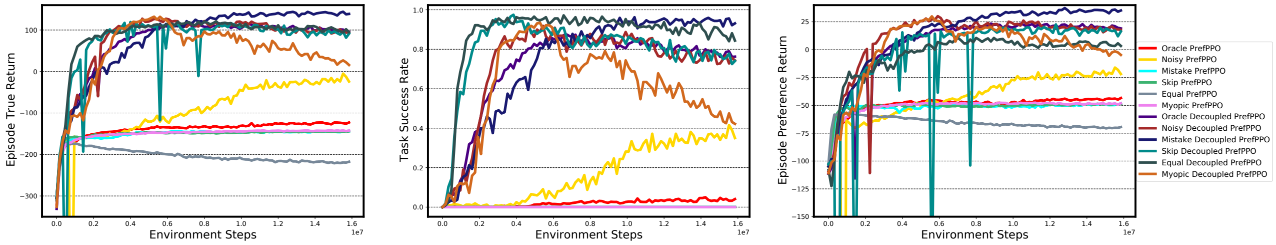

Fig. 6. The learning curves of PrefPPO and Decoupled PrefPPO using various scripted human teachers. The experimental results are measured on the ground truth return, task success rate, and ground truth preference return.

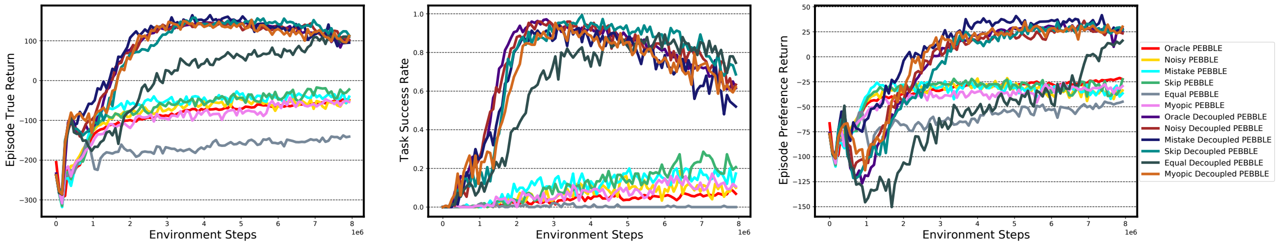

Fig. 7. The learning curves of PEBBLE and Decoupled PEBBLE using various scripted human teachers. The experimental results are measured on the ground truth return, task success rate, and ground truth preference return.

feedback and have great potential to be scaled to interacting with real humans in the physical world.

TABLE I
REALISTIC SCRIPTED HUMAN TEACHERS USED IN THIS PAPER.

Type	β	γ_{my}	ϵ	δ_{skip}	δ_{equal}
Oracle	∞	1	0	0	0
Noisy	1	1	0	0	0
Mistake	∞	1	0.1	0	0
Skip	∞	1	0	0.1	0
Equal	∞	1	0	0	0.1
Myopic	∞	0.99	0	0	0

D. Influences of the reward rate λ_t

λ_t is an important parameter, and is used to determine the trade-off between the sketchy task reward and the learned reward. To investigate the influences of λ_t, we design two decay strategies of λ_t in our experiments, e.g., linear strategy and non-linear strategy.

- *Linear strategy*: λ_t decreases linearly over training time. As shown in Equation 5, λ_0 is the only hyperparameter needs to be determined.
- *Non-linear strategy*: Similar to [21], λ_t decreases by an exponential decay schedule of $\lambda_t = (1-\rho)^t\lambda_0$, where ρ is a decay rate.

For linear strategy, we consider using $\lambda_0 \in \{1.0, 5.0, 10.0\}$ in Decoupled PrefPPO and using $\lambda_0 \in \{10.0, 15.0, 20.0\}$ in Decoupled PEBBLE. For non-linear strategy, we consider using $\rho = 0.00001, \lambda_0 \in \{5.0, 10.0, 30.0\}$ and $\rho = 0.000001, \lambda_0 \in \{5.0, 10.0, 30.0\}$ in both Decoupled PrefPPO and Decoupled PEBBLE. We report the experimental results in Fig. 8 and Fig. 9. To help robot explore task more effectively and acquire task skills more quickly, we should make sure that the sketchy task reward $\hat{r}_{task}$ works over a long period of time, e.g., using linear strategy or non-linear strategy with $\rho = 0.000001$. However, it is still necessary for us to adjust λ_0 carefully.

V. CONCLUSIONS

Our goal is to develop a robot that has personalized skills to adapt to different users in human-robot interaction. In this paper, we present decoupled preference-based RL, a novel preference-based RL method with prior knowledge of the task. We decouple the task from preference in personalized human-robot interaction. We utilize a sketchy task reward derived from the prior knowledge of the task to help the robot explore the task more effectively, and use a learned reward from preference-based RL to optimize the robot's policy to align with human preferences. We combine them organically via reward shaping. The experimental results show that the proposed method can achieve good performance on a complex interactive task and is an effective solution for personalized human-robot interaction.

REFERENCES

[1] M. Liu, C. Xiao, and C. Chen, "Perspective-corrected spatial referring expression generation for human-robot interaction," *IEEE Trans. Syst. Man Cybern. -Syst., Doi:10.1109/TSMC.2022.3161588*, 2022.
[2] C. Clabaugh and M. Matarić, "Robots for the people, by the people: personalizing human-machine interaction," *Sci. Robot.*, vol. 3, no. 21, p. eaat7451, 2018.
[3] D. Silver, A. Huang, C. J. Maddison *et al.*, "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, 2016.

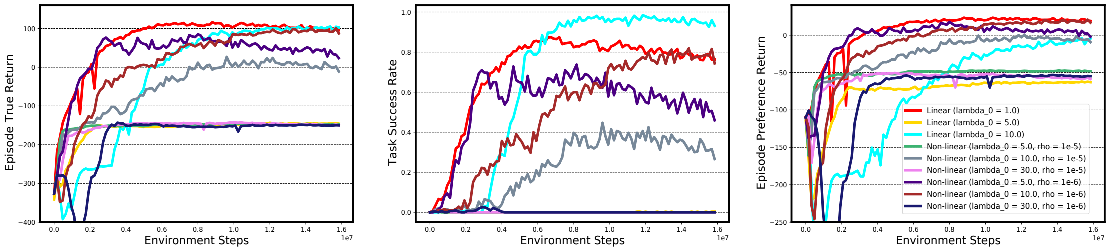

Fig. 8. The learning curves of Decoupled PrefPPO with different decay strategies. The experimental results are measured on the ground truth return, task success rate, and ground truth preference return.

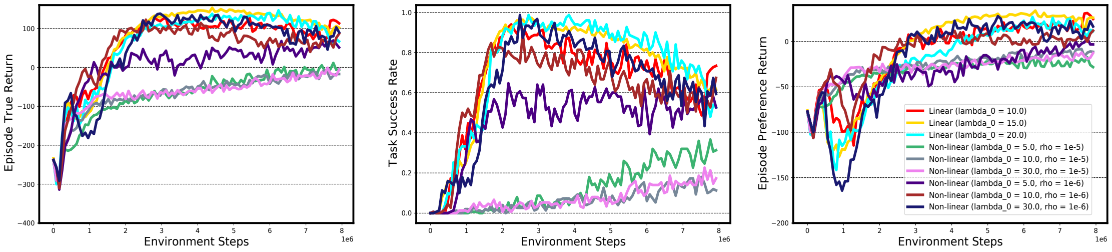

Fig. 9. The learning curves of Decoupled PEBBLE with different decay strategies. The experimental results are measured on the ground truth return, task success rate, and ground truth preference return.

[4] V. Mnih, K. Kavukcuoglu, D. Silver *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[5] J. Hwangbo, J. Lee, A. Dosovitskiy *et al.*, "Learning agile and dynamic motor skills for legged robots," *Sci. Robot.*, vol. 4, no. 26, p. eaau5872, 2019.

[6] A. Wilson, A. Fern, and P. Tadepalli, "A bayesian approach for policy learning from trajectory preference queries," *Adv. Neural Information Processing Systems*, vol. 25, 2012.

[7] J. Fürnkranz, E. Hüllermeier, W. Cheng *et al.*, "Preference-based reinforcement learning: a formal framework and a policy iteration algorithm," *Mach. Learn.*, vol. 89, no. 1, pp. 123–156, 2012.

[8] P. F. Christiano, J. Leike, T. Brown *et al.*, "Deep reinforcement learning from human preferences," *Adv. Neural Information Processing Systems*, vol. 30, 2017.

[9] C. Wirth, R. Akrour, G. Neumann *et al.*, "A survey of preference-based reinforcement learning methods," *J. Mach. Learn. Res.*, vol. 18, no. 136, pp. 1–46, 2017.

[10] K. Lee, L. M. Smith, and P. Abbeel, "Pebble: Feedback-efficient interactive reinforcement learning via relabeling experience and unsupervised pre-training," in *Int. Conf. Machine Learning*, 2021, pp. 6152–6163.

[11] A. Kapusta, Z. Erickson, H. M. Clever *et al.*, "Personalized collaborative plans for robot-assisted dressing via optimization and simulation," *Auton. Robot.*, vol. 43, no. 8, pp. 2183–2207, 2019.

[12] O. Rudovic, J. Lee, M. Dai *et al.*, "Personalized machine learning for robot perception of affect and engagement in autism therapy," *Sci. Robot.*, vol. 3, no. 19, p. eaao6760, 2018.

[13] S. Taylor, N. Jaques, E. Nosakhare *et al.*, "Personalized multitask learning for predicting tomorrow's mood, stress, and health," *IEEE Trans. Affect. Comput.*, vol. 11, no. 2, pp. 200–213, 2017.

[14] P. Patompak, S. Jeong, I. Nilkhamhang *et al.*, "Learning proxemics for personalized human–robot social interaction," *Int. J. Social Robotics*, vol. 12, no. 1, pp. 267–280, 2020.

[15] C. Moro, G. Nejat, and A. Mihailidis, "Learning and personalizing socially assistive robot behaviors to aid with activities of daily living," *ACM Trans. Hum.-Rob Interact.*, vol. 7, no. 2, pp. 1–25, 2018.

[16] M. Tucker, E. Novoseller, C. Kann *et al.*, "Preference-based learning for exoskeleton gait optimization," in *IEEE Int. Conf. Robotics and Automation*, 2020, pp. 2351–2357.

[17] M. Tucker, M. Cheng, E. Novoseller *et al.*, "Human preference-based learning for high-dimensional optimization of exoskeleton walking gaits," in *IEEE/RSJ Int. Conf. Intelligent Robots and Systems*, 2020, pp. 3423–3430.

[18] R. Akrour, M. Schoenauer, and M. Sebag, "April: Active preference learning-based reinforcement learning," in *Joint European conf. Machine Learning and Knowledge Discovery in Databases*, 2012, pp. 116–131.

[19] M. Palan, G. Shevchuk, N. Charles Landolfi *et al.*, "Learning reward functions by integrating human demonstrations and preferences," in *Robotics: Science and Systems*, 2019.

[20] B. Ibarz, J. Leike, T. Pohlen *et al.*, "Reward learning from human preferences and demonstrations in atari," *Adv. Neural Information Processing Systems*, vol. 31, 2018.

[21] X. Liang, K. Shu, K. Lee *et al.*, "Reward uncertainty for exploration in preference-based reinforcement learning," in *Deep RL Workshop NeurIPS*, 2021.

[22] H. Zhan, F. Tao, and Y. Cao, "Human-guided robot behavior learning: A gan-assisted preference-based reinforcement learning approach," *IEEE Robot. Autom. Lett.*, vol. 6, no. 2, pp. 3545–3552, 2021.

[23] Z. Cao, K. Wong, and C.-T. Lin, "Weak human preference supervision for deep reinforcement learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 12, pp. 5369–5378, 2021.

[24] J. Park, Y. Seo, J. Shin *et al.*, "Surf: Semi-supervised reward learning with data augmentation for feedback-efficient preference-based reinforcement learning," in *Deep RL Workshop NeurIPS*, 2021.

[25] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[26] R. A. Bradley and M. E. Terry, "Rank analysis of incomplete block designs: I. the method of paired comparisons," *Biometrika*, vol. 39, no. 3/4, pp. 324–345, 1952.

[27] J. Schulman, F. Wolski, P. Dhariwal *et al.*, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[28] T. Haarnoja, A. Zhou, P. Abbeel *et al.*, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Int. conf. Machine Learning*, 2018, pp. 1861–1870.

[29] K. Lee, L. Smith, A. Dragan *et al.*, "B-pref: Benchmarking preference-based reinforcement learning," *Adv. Neural Information Processing Systems*, 2021.

[30] Z. Erickson, V. Gangaram, A. Kapusta *et al.*, "Assistive gym: A physics simulation framework for assistive robotics," in *IEEE Int. Conf. Robotics and Automation*, 2020, pp. 10 169–10 176.