

# Learning Human-Aware Robot Policies for Adaptive Assistance

Jason Qin<sup>1</sup>, Shikun Ban<sup>2</sup>, Wentao Zhu<sup>2†</sup>, Yizhou Wang<sup>2</sup>, and Dimitris Samaras<sup>1</sup>

**Abstract**—Developing robots that can assist humans efficiently, safely, and adaptively is crucial for real-world applications such as healthcare. While previous work often assumes a centralized system for co-optimizing human-robot interactions, we argue that real-world scenarios are much more complicated, as humans have individual preferences regarding how tasks are performed. Robots typically lack direct access to these implicit preferences. However, to provide effective assistance, robots must still be able to recognize and adapt to the individual needs and preferences of different users. To address these challenges, we propose a novel framework in which robots infer human intentions and reason about human utilities through interaction. Our approach features two critical modules: the *anticipation module* is a motion predictor that captures the spatial-temporal relationship between the robot agent and user agent, which contributes to predicting human behavior; the *utility module* infers the underlying human utility functions through progressive task demonstration sampling. Extensive experiments across various robot types and assistive tasks demonstrate that the proposed framework not only enhances task success and efficiency but also significantly improves user satisfaction, paving the way for more personalized and adaptive assistive robotic systems. Code and demos are available at <https://asonin.github.io/Human-Aware-Assistance/>.

## I. INTRODUCTION

Developing robots that understand and assist humans is a critical long-term goal in Artificial Intelligence (AI) research. Beyond merely completing tasks, these systems must interact with humans in a friendly manner, dynamically coordinating and adapting to their needs. Such capabilities are essential for applications in home environments, assistive technologies, and healthcare. However, significant challenges remain, including ensuring safety, addressing the opacity of human goals, and navigating the diversity of human preferences. For example, when assisting disabled or elderly individuals with feeding, users may prefer robots not only to deliver food to the mouth but also to handle it gently and avoid spills. These tasks, though seemingly simple, are inherently complex, requiring precision, safety, comfort, and adaptability to individual preferences.

To achieve this goal, previous research has explored human-robot collaboration in assistive scenarios [1], [2]. Some studies [1], [3], [4] propose modeling humans and robots as a unified system, using co-optimization to learn assistive policies where both agents are jointly trained with the

<sup>1</sup>First Author and Fifth Author are with Department of Computer Science, Stony Brook University, USA [jaqin@cs.stonybrook.edu](mailto:jaqin@cs.stonybrook.edu), [samaras@cs.stonybrook.edu](mailto:samaras@cs.stonybrook.edu),

<sup>2</sup>Second Author, Third Author, and Forth Author are with Center on Frontiers of Computing Studies, Peking University, China [bansk@stu.pku.edu.cn](mailto:bansk@stu.pku.edu.cn), [wtzhu@pku.edu.cn](mailto:wtzhu@pku.edu.cn), [yizhou.wang@pku.edu.cn](mailto:yizhou.wang@pku.edu.cn),

† Corresponding author.

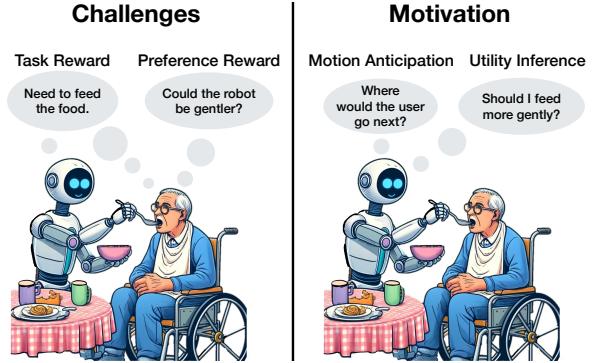


Fig. 1. In our task scenario (demonstrated with the feeding example in the figure), the robot’s initial objective is to achieve the basic goal of “feeding the food,” represented by the *Task Reward*. However, as shown in **Left**, the human user also has more nuanced *Preference Reward*, which are unknown to the robot, leading to a misalignment between the human and robot reward functions. To address this, we propose a novel framework, as depicted in **Right**. Beyond learning to fulfill the basic task requirements, we introduce two additional modules to better model human behavior and preferences: a *Motion Anticipation* module for predicting the human’s future motion and a *Utility Inference* module for estimating user preferences.

same reward signal. However, this approach does not fully capture real-world dynamics. In practice, robots cannot fully access the human mind, leading to a gap or misalignment between human and robot reward functions. For the robot, the most straightforward *task reward* is completing the task, such as whether the food is delivered to the intended location. For humans, there is an additional layer of *preference reward* — preferences regarding how the task is performed, such as the speed or force used during food delivery [1], [5]. These human rewards are often difficult to articulate and vary across individuals, making them unknown to the robot. Therefore, robots must gather information during interactions, inferring human intentions from behavioral cues and reasoning about the underlying human preferences [6]–[8]. By adjusting their policies accordingly, robots can more effectively collaborate with humans to complete tasks while better aligning with human preferences and needs, thereby enhancing safety, robustness, and user satisfaction.

In light of this, we propose a more realistic formulation for assistive robotics and introduce a novel policy learning approach that first infers human intentions and utility functions, and then provides assistance accordingly. We first define an assistive task setting in which humans and robots cooperate under different reward signals. Specifically, robots are guided solely by clearly defined task rewards, while human behavior is driven by both task rewards and human preference rewards. This setting more accurately reflects real-world conditions and presents significant challenges. To facilitate effective

human-robot collaboration, we introduce a policy learning framework which comprises two key components, as shown in Figure 1. Specifically, we propose an *anticipation module* which enables the robot to predict human actions before determining its own actions. Furthermore, our *utility module* estimates human utility function weight progressively from online interactions through distribution sampling, further improving the robot's adaptation ability to user preferences without the need of any preference queries.

We conduct comprehensive experiments across multiple tasks, preference settings, and robot types. Experimental results show that these modules not only improve task success and efficiency but also significantly enhance human rewards, resulting in a safer and more user-friendly experience. Our contributions are summarized as follows: (1) We propose a more realistic formulation in which human and robot rewards are not fully aligned. (2) We introduce an anticipation module, enabling the robot to predict human behavior before making its own action decisions. (3) We develop a utility module that infers the human utility function on the fly, without relying on explicit preference queries or human feedback.

## II. RELATED WORK

### A. Assistive Robotics

Assistive robots have been explored as a viable solution to help disabled and elderly individuals perform daily tasks [1]–[5], [9]–[16]. Some works have focused on designing and developing virtual environments to facilitate the development and comparison of assistive algorithms [1], [2], [17]. For example, Assistive Gym [1] is a simulation framework for physical human-robot interaction and robotic assistance, providing several tasks crucial for daily living and introducing the co-optimization paradigm. Similarly, RCareWorld [2] proposes a framework for developing and evaluating socially-aware robotic assistance. Researchers have also designed various approaches to enhance assistive performance [3]–[5], [14], [18]–[24]. For example, Müller *et al.* [18] focus on integrating human-robot collaboration to enhance ergonomics, flexibility, and quality in the assembly line, Unhelkar *et al.* [19] introduce a human-aware robotic assistant system that combines human motion prediction with time-optimal path planning. Gallenberger *et al.* [14] examine manipulation strategies for robotic feeding, focusing on how different bite acquisition and transfer methods impact the success and ease of feeding for users. Jakhotiya *et al.* [21] explores using recurrent neural networks to enhance policy learning for assistive tasks like feeding and dressing. Osa *et al.* [22] introduce a solution by maximizing mutual information between state-action pairs and latent variables, allowing for a wider range of adaptive behaviors in tasks requiring robustness. He *et al.* [3] aims to efficiently transfer knowledge gained from one task to another by learning robust representations, enhancing the effectiveness and versatility of assistive robots. Osa *et al.* [4] suggests employing a latent variable to model the behavior of an opposing agent.

### B. Human-aware Policy Learning

To develop collaborative robots better suited to work alongside or for humans, some works have proposed incorporating human feelings and experiences into consideration [5], [25]–[30]. Sadigh *et al.* [26] propose an interactive framework that efficiently learns complex reward landscapes from minimal user input, making the learning process both efficient and aligned with human preferences while the expressiveness of such minimal user input remains limited. Chen *et al.* [27] present a human-in-the-loop approach for tele-operation in assistive robotics. While leveraging a control loop, their method enhances the robot's capability to deliver real-time assistance in dynamic environments. Palan *et al.* [28] enables systems to understand and predict what users value without them having to specify it explicitly each time. Another line of works uses preference-based reinforcement learning (PbRL) to train agents with pairwise human feedback [31]–[33]. Recent advances aim to alleviate the reliance on explicit human feedback by introducing off-policy algorithms that relabel history experience and conduct unsupervised pre-training [34], and training a preference predictor to provide pseudo preference labels [35]–[37]. Specifically, Liu *et al.* [5] enhances the efficiency and robustness of preference-based reinforcement learning in complex interactive tasks by decoupling task objectives from preferences and integrating task prior knowledge into the PbRL framework. Their combination of the sketchy task reward shaping with an automated scripted teacher, reduces reliance on real human feedback during the training process. Although these methods make valuable contributions to incorporating human factors into policy learning, they still often rely on human demonstrations or explicit preference queries, which limits their ability to autonomously learn and adapt to the complexity of human behavior.

## III. METHOD

In this section, we first introduce the problem setup and formulate the reward misalignment problem between human and robot. Then, we describe our proposed method step-by-step. The overview of the proposed human-aware policy learning framework is demonstrated in Figure 2.

### A. Problem Formulation

The assistive task is modeled as a two-agent, finite-horizon decentralized partially observable Markov decision process (Dec-POMDP), defined by the tuple  $(\mathcal{S}, \mathcal{O}, \mathcal{A}, T, r)$ . Here,  $\mathcal{S}$  is the state space,  $\mathcal{A} = \mathcal{A}_H \times \mathcal{A}_R$  is the joint action space of the human ( $\mathcal{A}_H$ ) and robot ( $\mathcal{A}_R$ ), and  $T(\mathbf{s}^{t+1}|\mathbf{s}^t, \mathbf{a}^t)$  defines the transition dynamics.  $\mathcal{O} = \mathcal{O}_H \times \mathcal{O}_R$  is the joint observation space, where the agents' observations are derived from the previous state and actions. The policies  $\pi_H$  and  $\pi_R$  generate actions based on the agents' observations,  $\pi(a^t|o^t)$ . The reward function  $r(\mathbf{s}, \mathbf{a})$  assigns a scalar reward based on the current state and actions. A trajectory  $\xi = (\mathbf{s}^t, o_H^t, o_R^t, a_H^t, a_R^t)_{t=0}^N$  represents the sequence of states, observations, and actions over  $N$  timesteps.

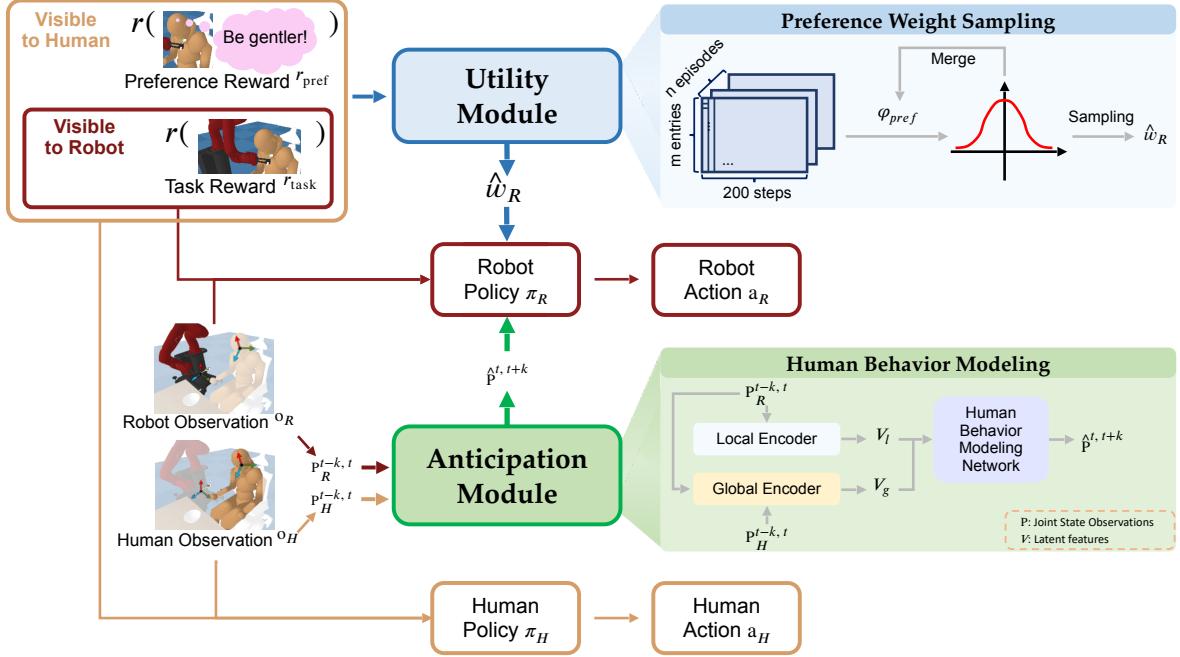


Fig. 2. Overview of the proposed framework. Each agent in the system is receiving an observation about its own information and some critical information about the other agent. The human agent is controlled by an independent policy powered by RL algorithms. The robot agent consists of three parts: an RL backbone, an anticipation module III-B, and a utility module III-C. The *anticipation module* predicts future human motion by taking past  $k$  frames of joint information  $p_R^{t-k,t}$  and  $p_H^{t-k,t}$  from both agents, and predicting an anticipated  $k$  steps future human joint information  $\hat{p}_H^{t,t+k}$ . The *utility module* leverages the interaction histories to estimate a robot preference reward weight  $\hat{w}_R$ , which is used to compute an estimated preference reward  $\hat{r}_{\text{pref}}$  to further guide robot learning.

Specifically, for a given trajectory  $\xi$ , we define two types of rewards: the task success reward and the human preference reward. The task reward is defined as:

$$r_{\text{task}} = \phi_{\text{task}}(\xi), \quad (1)$$

where  $\phi_{\text{task}}$  is a function that maps the agents' trajectories to a reward based on task completion and performance. Meanwhile, the human preference reward is defined as:

$$r_{\text{pref}} = \phi_{\text{pref}}(\xi; w), \quad (2)$$

where  $\phi_{\text{pref}}(\xi; w)$  is a feature function parameterized by the weight parameter  $w$ , mapping the trajectory  $\xi$  to a scalar reward representing human satisfaction. The parameter  $w$  reflects the weight assigned to individual human preferences. In this work, we use a fixed set of human preference items and various preference settings. Each setting assigns different weights to the preference items, shaping the overall human preference reward.

We note that a fundamental misalignment often arises because robots cannot fully access or understand human preferences. These task performance preferences are subjective, vary among individuals, and are often opaque to robots. Specifically, the human reward, combining both task and preference components, is:

$$r_H(\xi) = r_{\text{pref}} + r_{\text{task}} \quad (3)$$

In contrast, robots generally have access only to the task reward, with their objective focused solely on task success,  $r_R(\xi) = r_{\text{task}}$ . This often leads to suboptimal outcomes from

the human's perspective, highlighting the challenge of designing robots that can better incorporate human preferences for more satisfying interactions.

### B. Anticipation Module

To create robots that effectively integrate human preferences for more engaging interactions, an intuitive approach is to predict human future actions before making its own decisions [38]. By using this predictive information to guide robot policies, the robot can generate movements that align more closely with human motion trends, reducing discomfort and enhancing task performance. Given that agents' behavior patterns can be directly learned through observation [39] within the limited observability, we attempt to establish an anticipation module  $\mathcal{M}$  to predict human's joint positions  $\hat{p}_H$  in future  $k$  steps  $\hat{p}_H^{t,t+k}$ , as shown on the bottom of Figure 2. Specifically,

$$\hat{p}_H^{t,t+k} = \mathcal{M}(p_R^{t-k,t}, p_H^{t-k,t}), \quad (4)$$

where  $p_H^{t-k,t}$  and  $p_R^{t-k,t}$  represent joint information extracted from human and robot observations in the past  $k$  timesteps. In practice, we utilize different length  $k$  of anticipated future human joint angle at different stages, refer to Appendix B for explanation. We optimize the anticipation module parameters  $\theta$  by minimizing the prediction error:

$$\theta^* = \arg \min_{\theta} \sum_{\xi \in \Xi} \sum_{t=k}^{N-k} \mathcal{L} \left( \mathcal{M}_{\theta}(p_R^{t-k,t}, p_H^{t-k,t}), p_H^{t,t+k} \right), \quad (5)$$

where  $\xi \in \Xi$  are buffered trajectories, and  $\mathcal{L}$  is L2 Loss.

Therefore, we derive a human-aware policy for the robot, which generates an action that more closely aligns with the user's motion tendencies based on the anticipated future joint information of the human user:

$$a_R^t = \pi_R(o_R^t, \hat{p}_H^{t,t+k}). \quad (6)$$

In practice, our anticipation module utilizes data from the PPO buffer to predict outcomes, which, when combined with robot observations, guide the robot policy in action decisions. To maintain efficiency, the anticipation module is updated every  $e_k$  epochs within the PPO policy update cycle. Importantly, the human joint information used to train the anticipation module is not shared with the PPO update algorithm, ensuring independence of information between the robot's PPO policy and the human agent.

---

**Algorithm 1** Framework Pseudocode

---

```

1: Initialize policy  $\pi_R$ ,  $\pi_H$  and estimated human utility
   function weight  $\hat{w}_H$ 
2: for each epoch do
3:   for each episode do
4:     Collect trajectories  $\{\xi_i\}$ 
5:      $p_R^t \leftarrow o_R^t$  and  $p_H^t \leftarrow o_H^t$ 
6:     Anticipation module generates predicted future joint
       information  $\hat{p}_H^{t,t+k}$  with Equation (4)
7:     Policies generate actions  $a_H$  and  $a_R$  with Equation
       (6) and  $a_H^t = \pi_H(o_H^t)$ 
8:   end for
9:   Calculate robot reward  $r_R(\xi)$  with Equation (11) and
   human reward  $r_H(\xi)$  with Equation (3)
10:  if epoch mod  $e_k = 0$  then
11:    Update Anticipation module  $\mathcal{M}$  with Equation (5)
12:    Update estimated human utility function weight  $\hat{w}_H$ 
       with Equation (9):
13:  end if
14:  Update  $\pi_R$  through PPO algorithm with  $(\xi^D, r_R(\xi^D))$ 
15:  Update  $\pi_H$  through PPO algorithm with  $(\xi^D, r_H(\xi^D))$ 
16: end for

```

---

### C. Utility Module

In addition to directly anticipating human behavior, it is even more crucial to recognize that human behavior is guided by an underlying utility function that reflects individual values and preferences. Therefore, we propose a utility module that enables the robot agent to infer an individual's utility function weight. The utility module design is displayed in Figure 2.

Based on [28], the preference weight distribution can be directly obtained from expert demonstrations in the environment through a sampling process. Acknowledging the challenge in determining utility function weights through limited demonstrations, it's noted that as the learning progresses, the policy narrows from a broad to a precise spectrum, and the consistency of demonstrated behaviors increases [40], [41]. This alignment facilitates both a reduction in the uncertainty

of preference distributions and an improved ability for robots to understand human preferences as the policy converges. Specifically, the probability of generating a specific trajectory given the utility function weights  $w$  and the current policy  $\pi$  is given by:

$$P(\xi^D | w, \pi) = \prod_{i=1}^n P(\xi_i^D | w, \pi), \quad (7)$$

Consequently, we can derive the global update formula, which forms the complete utility module :

$$P(\hat{w}_H | \xi^D, t) = \exp \left( \sum_{j=1}^n \phi_{\text{pref}}(\xi_j)_{\pi_t}^D \right), \quad (8)$$

In reinforcement learning, the uncertainty of a policy decreases as training progresses, resulting in more stable demonstrations, which aids in estimating the utility function weights. In our approach, we aim to iteratively refine the estimated utility function weight distribution. Rather than discarding earlier estimates, we combine them linearly with the current estimates using a weighting parameter  $\gamma$ . This allows us to formulate the final update rule for the utility function weight distribution as follows:

$$\begin{aligned} P(\hat{w}_H | \xi^D) &= \lim_{t \rightarrow \infty} \left( \gamma \exp \left( \sum_{j=1}^n \phi_{\text{pref}}(\xi_j)_{\pi_t}^D \right) \right. \\ &\quad \left. + (1 - \gamma) P(\hat{w}_H | \xi^D, t) \right). \end{aligned} \quad (9)$$

We implement utility module by initializing the utility function weights space  $\mathcal{X}$  onto a unit ball, and then sample through  $N$  demonstrations  $\xi^D$  under the current policy. For simplicity, we first map the high-dimensional trajectory data to a set  $\phi_{\text{pref}}(\xi^D) \in \mathbb{R}^{n*m}$  where  $n$  is number of demonstrations and  $m$  is the number of preference entries. Then, we sample the weight distribution via the Markov Chain Monte Carlo (MCMC) method [26]. We follow [28] and further give a specific update formula:

$$f(\mathcal{X}) = -\log \left( \sum e^{\mathcal{X}} + (\phi_{\text{pref}}(\xi^D) \mathcal{X}) \right). \quad (10)$$

During training, the sampled human utility weight  $\hat{w}_H$  is used to generate the estimated preference reward  $\hat{r}_{\text{pref}} = \phi_{\text{pref}}(\xi; \hat{w}_H)$ , which, combined with the task reward, forms the overall robot reward shaping the robot's subsequent actions:

$$r_R(\xi) = \hat{r}_{\text{pref}} + r_{\text{task}}, \quad (11)$$

In practice, the update frequency of this module is synchronized with the anticipation module. The process does not rely on human-provided demonstrations or queries but instead utilizes PPO buffer data, similar to the anticipation module. This enables the entire process to operate autonomously without requiring human feedback.

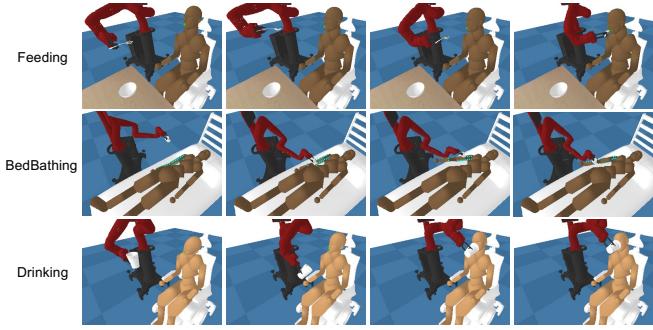


Fig. 3. Successful episodes of our method for each task scenario. The key frames are arranged sequentially from left to right, following the progression of the episode.

#### IV. EXPERIMENTS

In this section, we evaluate our approach in collaborative human-robot interaction environments of varying robot models and assistive tasks. In particular, we seek to answer the following research questions. RQ1: What is the influence of human and robot reward misalignment? RQ2: How does our proposed method adapt to different human preferences? RQ3: How does our proposed method adapt to different assistive tasks? RQ4: How does our proposed method adapt to different robot types? RQ5: How does different merging ratio in preference weight estimation affect agent’s performance?

##### A. Setup

*a) Tasks:* We explore multiple task scenarios, each presenting unique operational challenges and distinct action patterns. In the *Feeding* task, the robot is required to precisely deliver food particles to the human user’s mouth using a robotic arm. The *Drinking* task is similar but includes an additional complexity of arm rotation, which introduces operational difficulty. Conversely, the *Bed Bathing* task differs significantly in its action pattern; it demands that the robotic arm wipe along the user’s arm, maintaining constant contact and incorporating rotational movements throughout the process. In Figure 3, we visualize one successful episode of our method for each task. For a complete qualitative comparison with the baseline methods, please refer to the attached video.

*b) Human Preference Settings:* We focus on several aspects of human preference following [1]: *Hit*, which refers to penalties incurred when particles touch points outside the target area, indicating whether the robot has delivered particles to incorrect locations or caused spillage; *Force*, referring to the penalties for additional forces applied by the robot or tool outside the target point, representing penalties for accidental contact with non-target areas; *High Force*, which refers to penalties for excessive force applied by the robot or tool on the human body, indicative of whether the robot can adequately control the service force, typically set at 10N. In our experiments, we evaluate our approach with simulated users, each prioritizing different preferences through varying weight combinations.

TABLE I  
FEEDING TASK PERFORMANCE WITH SAWYER ROBOT. A “/” DENOTES INVALID VALUE DUE TO TASK INCOMPLETION.

Method	Human Reward	Hit Penalty	Force Penalty	High Force Penalty	Success Rate
PPO ( <i>co-opt.</i> )	115.992	-0.406	-0.756	-2.102	100.0%
PPO	109.063	-2.668	-1.421	-4.645	98.8%
TD3	87.069	-2.341	-2.149	-12.146	90.0%
MADDPG	-125.921	-9.086	-8.382	-0.537	2.5%
Ours	<b>118.656</b>	<b>-0.654</b>	<b>-0.470</b>	<b>-0.319</b>	<b>100.0%</b>

*c) Metrics and Baselines:* We compute the following metrics to comprehensively evaluate the performance of assistive policies: (1) *Human Reward*  $R_H$ , which includes both task success and human preference rewards per episode; (2) Decomposed human preference reward terms, capturing different aspects including *Hit*, *Force*, and *High Force* penalties per episode; and (3) Average *Success Rate* across episodes. We implement our framework on top of PPO [42]. We compare it with PPO [42], TD3 [43], and a multi-agent RL method MADDPG [44]. By default, we consider the scenario where human and robot rewards are misaligned. For comparison, we also experiment with a setting where the two rewards are co-optimized (*co-opt.*), following [1].

##### B. Evaluation and Comparison

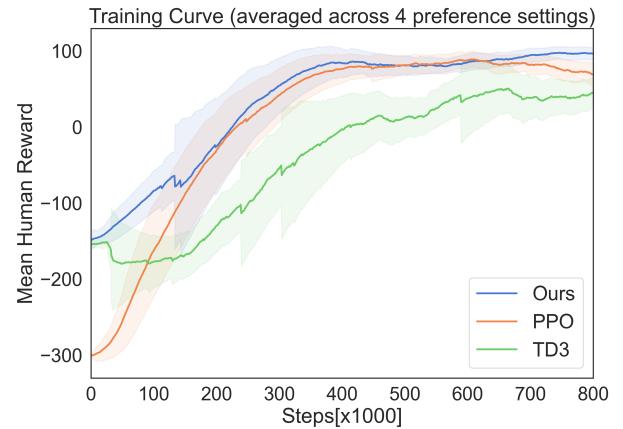


Fig. 4. Training curves of baseline PPO, TD3 and our method in 4 different human preference settings (with robot Sawyer conducting the feeding task). Our method delivers generally superior performance throughout the training process.

*1) Varying Human Preferences:* In this section, we aim to address two pivotal questions: Firstly, we explore the influence of misalignment between human and robot rewards (RQ1). Secondly, we evaluate the performance of our proposed method in comparison to others among users with varying preferences (RQ2). These inquiries are fundamental to understanding the dynamics of reward systems and their impact on user satisfaction and task effectiveness in robotic applications. We evaluate the feeding task under four

different human preference settings, with results averaged across preferences shown in Table I. The results show that reward misalignment increases task difficulty, reducing both success rate and human reward, as seen in the first two lines. Meanwhile, our approach steadily improves both task performance and human rewards. It not only outperforms the baseline under the same reward misalignment condition but also achieves performance comparable to or even better than the co-optimization scenario. We believe this is due to our method’s ability to explicitly model the behavior of human users more effectively, enabling a better understanding of human demands.

As a reference, we present the training curves for the two primary baseline methods, PPO and TD3, alongside our own method in Figure 4.

TABLE II  
PERFORMANCE ON DIFFERENT ASSISTIVE TASKS. A “/” DENOTES INVALID VALUE DUE TO TASK INCOMPLETION.

Method	Human Reward	Hit Penalty	Force Penalty	High Force Penalty	Success Rate
<b>Feeding</b>					
PPO	99.273	-5.500	<b>-0.054</b>	-6.135	95%
TD3	102.300	<b>-2.000</b>	-0.324	-6.900	<b>100%</b>
MADDPG	-213.700	/	/	/	0%
Ours	<b>108.102</b>	<b>-2.000</b>	-0.133	<b>-1.245</b>	<b>100%</b>
<b>Bed Bathing</b>					
PPO	123.585	<b>0.000</b>	-2.886	-2.630	15%
TD3	52.372	/	/	/	0%
MADDPG	11.536	/	/	/	0%
Ours	<b>148.558</b>	<b>0.000</b>	<b>-0.978</b>	<b>-1.204</b>	<b>60%</b>
<b>Drinking</b>					
PPO	344.189	-53.000	-2.359	-0.597	90%
TD3	-68.404	/	/	/	0%
MADDPG	-141.585	/	/	/	0%
Ours	<b>415.390</b>	<b>-16.500</b>	<b>-0.387</b>	<b>-0.312</b>	<b>100%</b>

2) *Varying Assistive Tasks:* We aim to evaluate the generalization performance of our proposed method across tasks with varying operational modes. To this end, we test it on two additional complex tasks to assess whether our approach can be broadly applied to a diverse range of assistive tasks. This exploration is critical for validating the wide applicability and effectiveness of our method in various daily assistive scenarios. The results, shown in Table II, demonstrate strong generalization performance. For visualized comparisons, please refer to the attached video. For example, in the Bed Bathing task, TD3 fails to complete the task, and the success rate of the PPO baseline drops significantly, while our method achieves a 60% success rate. In the Drinking task, while maintaining a 100% task completion rate, our method reduces the Hit penalty to less than half of its original level compared to PPO.

3) *Varying Robot Types:* We also investigate whether the proposed method maintains its generalization ability and robustness across different robotic embodiments, a critical question for deploying to various robotic platforms in practical scenarios. Experiments are conducted using four robots—Sawyer, Baxter, Jaco, and PR2—on the feeding task, with details provided in Table III. These findings confirm

TABLE III  
PERFORMANCE ON DIFFERENT ROBOTS. A “/” DENOTES INVALID VALUE DUE TO TASK INCOMPLETION.

Method	Human Reward	Hit Penalty	Force Penalty	High Force Penalty	Success Rate
Jaco					
PPO	84.157	-9.000	-0.654	-3.402	85%
TD3	98.581	-8.000	-0.240	-1.989	<b>100%</b>
MADDPG	-137.214	/	/	/	0%
Ours	<b>111.074</b>	<b>-1.000</b>	<b>-0.169</b>	<b>-0.985</b>	<b>100%</b>
PR2					
PPO	88.850	-2.500	<b>-0.179</b>	-1.828	90%
TD3	-23.22	-21.500	-0.803	-11.385	40%
MADDPG	-97.440	/	/	/	0%
Ours	<b>91.982</b>	<b>-0.000</b>	-0.292	<b>-1.201</b>	<b>100%</b>
Baxter					
PPO	104.759	-1.500	-0.228	-4.623	<b>100%</b>
TD3	100.350	-3.500	-0.268	-5.602	<b>100%</b>
MADDPG	-139.963	/	/	/	0%
Ours	<b>115.401</b>	<b>-0.000</b>	<b>-0.095</b>	<b>-2.074</b>	<b>100%</b>
Sawyer					
PPO	99.273	-5.500	<b>-0.054</b>	-6.135	95%
TD3	102.3	<b>-2.000</b>	-0.324	-6.900	<b>100%</b>
MADDPG	-213.7	/	/	-2.100	0%
Ours	<b>108.102</b>	<b>-2.000</b>	-0.133	<b>-1.245</b>	<b>100%</b>

TABLE IV  
ABLATION RESULTS ON DIFFERENT MODULES.

Method	Human Reward	Hit Penalty	Force Penalty	High Force Penalty	Success Rate
PPO	109.063	-2.668	-1.421	-18.581	98.8%
Ours	118.211	-2.901	-0.689	-4.642	98.8%
(w/o Utility)					
Ours (full)	<b>118.656</b>	<b>-0.654</b>	<b>-0.470</b>	<b>-0.319</b>	<b>100.0%</b>

the proposed method’s strong generalization ability and robustness across diverse robotic embodiments and operational scenarios.

### C. Ablation Studies

a) *Module Contribution:* To validate the contributions of the proposed modules to task performance, we conduct a comparative analysis on the feeding task under various preference settings. As shown in Table IV, incorporating the Anticipation module alone improves overall human reward, while integrating both modules further enhances preference satisfaction and success rates, particularly in specific preference categories. This demonstrates the complementary roles and individual contributions of the two modules.

b) *Distribution Merging Ratio:* In Section III-C, we describe the progressive merging process for estimating preference value combinations. To investigate the impact of different distribution merging ratios, we evaluate them in the Sawyer robot Feeding experiment. As shown in Table V, a merge ratio of 0.0 represents the complete exclusion of previously sampled distributions. Overall, the proposed merging mechanism improves human satisfaction.

TABLE V  
ABLATION RESULTS ON MERGE RATIO.

Merge Ratio	Human Reward	Hit Penalty	Force Penalty	High Force Penalty	Success Rate
0.0	103.668	-2.500	-0.198	-2.696	<b>100%</b>
0.1	110.978	-6.500	<b>-0.083</b>	-1.427	<b>100%</b>
0.3	108.102	-2.000	-0.133	<b>-1.245</b>	<b>100%</b>
0.5	<b>113.711</b>	<b>-0.500</b>	-0.185	-3.608	<b>100%</b>

Overall, the proposed merging mechanism improves human satisfaction.

## V. CONCLUSION

In this work, we propose a framework that enables robots to infer human intentions and adapt to individual preferences through an anticipation module for behavior prediction and a utility module for modeling human utilities. Experiments across various robots and tasks show that the framework improves task success, efficiency, and user satisfaction. One limitation of this work is that, due to the complexity of human subjects, experiments are primarily conducted in simulation. Future work could explore real-world deployment and incorporate explicit communication mechanisms for better human-robot collaboration.

## REFERENCES

- [1] Z. Erickson, V. Gangaram, A. Kapusta, C. K. Liu, and C. C. Kemp, “Assistive gym: A physics simulation framework for assistive robotics,” in *ICRA 2020*. IEEE, 2020, pp. 10 169–10 176.
- [2] R. Ye, W. Xu, H. Fu, R. K. Jenamani, V. Nguyen, C. Lu, K. Dimitropoulou, and T. Bhattacharjee, “Rcareworld: A human-centric simulation world for caregiving robots,” *IROS*, 2022.
- [3] J. Z.-Y. He, Z. Erickson, D. S. Brown, A. Raghunathan, and A. Dragan, “Learning representations that enable generalization in assistive tasks,” in *Conference on Robot Learning*. PMLR, 2023, pp. 2105–2114.
- [4] T. Osa and T. Harada, “Robustifying a policy in multi-agent rl with diverse cooperative behavior and adversarial style sampling for assistive tasks,” *arXiv preprint arXiv:2403.00344*, 2024.
- [5] M. Liu and C. Chen, “Task decoupling in preference-based reinforcement learning for personalized human-robot interaction,” in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 848–855.
- [6] G. Hoffman, T. Bhattacharjee, and S. Nikolaidis, “Inferring human intent and predicting human action in human-robot collaboration,” *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 7, 2024.
- [7] W. Zhu, Z. Zhang, and Y. Wang, “Language models represent beliefs of self and others,” in *Forty-first International Conference on Machine Learning*, 2024.
- [8] S. Jain and B. Argall, “Recursive bayesian human intent recognition in shared-control robotics,” in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 3905–3912.
- [9] T. L. Chen, M. Ciocarlie, S. Cousins, P. M. Grice, K. Hawkins, K. Hsiao, C. C. Kemp, C.-H. King, D. A. Lazewatsky, A. E. Leeper *et al.*, “Robots for humanity: using assistive robotics to empower people with disabilities,” *IEEE Robotics & Automation Magazine*, vol. 20, no. 1, pp. 30–39, 2013.
- [10] T. Rhodes and M. Veloso, “Robot-driven trajectory improvement for feeding tasks,” in *IROS 2018*. IEEE.
- [11] A. Nanavati, P. Alves-Oliveira, T. Schrenk, E. K. Gordon, M. Cakmak, and S. S. Srinivasa, “Design principles for robot-assisted feeding in social contexts,” in *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, 2023, pp. 24–33.
- [12] T. Bhattacharjee, M. E. Cabrera, A. Caspi, M. Cakmak, and S. S. Srinivasa, “A community-centered design framework for robot-assisted feeding systems,” in *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*, 2019, pp. 482–494.
- [13] D. Park, Y. Hoshi, and C. C. Kemp, “A multimodal anomaly detector for robot-assisted feeding using an lstm-based variational autoencoder,” *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 1544–1551, 2018.
- [14] D. Gallenberger, T. Bhattacharjee, Y. Kim, and S. S. Srinivasa, “Transfer depends on acquisition: Analyzing manipulation strategies for robotic feeding,” in *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2019, pp. 267–276.
- [15] J. Thumm, F. Trost, and M. Althoff, “Human-robot gym: Benchmarking reinforcement learning in human-robot collaboration,” in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 7405–7411.
- [16] B. Zhou, H. Zhou, T. Liang, Q. Yu, S. Zhao, Y. Zeng, J. Lv, S. Luo, Q. Wang, X. Yu *et al.*, “Clothesnet: An information-rich 3d garment model repository with simulated clothes environment,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 20 428–20 438.
- [17] B. Shen, F. Xia, C. Li, R. Martín-Martín, L. Fan, G. Wang, C. Pérez-D’Arpino, S. Buch, S. Srivastava, L. Tchapmi *et al.*, “igibson 1.0: A simulation environment for interactive tasks in large realistic scenes,” in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 7520–7527.
- [18] R. Müller, M. Vette, and M. Scholer, “Robot workmate: a trustworthy coworker for the continuous automotive assembly line and its implementation,” *Procedia CIRP*, vol. 44, pp. 263–268, 2016.
- [19] V. V. Unhelkar, P. A. Lasota, Q. Tyroller, R.-D. Buhai, L. Marceau, B. Deml, and J. A. Shah, “Human-aware robotic assistant for collaborative assembly: Integrating human motion prediction with planning in time,” *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 2394–2401, 2018.
- [20] F. Zhang and Y. Demiris, “Learning garment manipulation policies toward robot-assisted dressing,” *Science robotics*, vol. 7, no. 65, p. eabm6010, 2022.
- [21] Y. Jakhotiya and I. Haque, “Improving assistive robotics with deep reinforcement learning,” *arXiv preprint arXiv:2209.02160*, 2022.
- [22] T. Osa, V. Tangkaratt, and M. Sugiyama, “Discovering diverse solutions in deep reinforcement learning by maximizing state-action-based mutual information,” *Neural Networks*, vol. 152, pp. 90–104, 2022.
- [23] R. Madan, R. K. Jenamani, V. T. Nguyen, A. Moustafa, X. Hu, K. Dimitropoulou, and T. Bhattacharjee, “Sparses: Structuring physically assistive robotics for caregiving with stakeholders-in-the-loop,” in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 641–648.
- [24] K. Puthuveetil, C. C. Kemp, and Z. Erickson, “Bodies uncovered: Learning to manipulate real blankets around people via physics simulations,” *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 1984–1991, 2022.
- [25] S. Christen, W. Yang, C. Pérez-D’Arpino, O. Hilliges, D. Fox, and Y.-W. Chao, “Learning human-to-robot handovers from point clouds,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 9654–9664.
- [26] D. Sadigh, A. Dragan, S. Sastry, and S. Seshia, *Active preference-based learning of reward functions*, 2017.
- [27] S. Chen, J. Gao, S. Reddy, G. Berseth, A. D. Dragan, and S. Levine, “Asha: Assistive teleoperation via human-in-the-loop reinforcement learning,” in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 7505–7512.
- [28] M. Palan, N. C. Landolfi, G. Shevchuk, and D. Sadigh, “Learning reward functions by integrating human demonstrations and preferences,” *arXiv preprint arXiv:1906.08928*, 2019.
- [29] A. Clegg, Z. Erickson, P. Grady, G. Turk, C. C. Kemp, and C. K. Liu, “Learning to collaborate from simulation for robot-assisted dressing,” *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 2746–2753, 2020.
- [30] Y.-W. Chao, C. Paxton, Y. Xiang, W. Yang, B. Sundaralingam, T. Chen, A. Murali, M. Cakmak, and D. Fox, “Handoversim: A simulation framework and benchmark for human-to-robot object handovers,” in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 6941–6947.

TABLE VI  
PREFERENCE WEIGHTS FOR SETTINGS UTILIZED IN EXPERIMENTS.

Weight Type	Setting 0	Setting 1	Setting 2	Setting 3	Setting 4
Force Non-Target Weight	0.01	0.01	0.10	0.001	0.10
High Forces Weight	0.05	0.50	5.00	0.005	0.005
Particle Hit Weight	1.00	10.0	1.00	0.10	10.0

- [31] A. Wilson, A. Fern, and P. Tadepalli, “A bayesian approach for policy learning from trajectory preference queries,” in *Advances in Neural Information Processing Systems*, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds., vol. 25. Curran Associates, Inc., 2012.
- [32] C. Wirth and J. Fürnkranz, “Preference-based reinforcement learning: A policy iteration algorithm for learning from preference-based feedback,” in *Advances in Intelligent Data Analysis XII*, A. Tucker, F. Höppner, A. Siebes, and S. Swift, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 427–437.
- [33] C. Wirth, R. Akrou, G. Neumann, and J. Fürnkranz, “A survey of preference-based reinforcement learning methods,” *Journal of Machine Learning Research*, vol. 18, no. 136, pp. 1–46, 2017.
- [34] K. Lee, L. M. Smith, and P. Abbeel, “Pebble: Feedback-efficient interactive reinforcement learning via relabeling experience and unsupervised pre-training,” in *International Conference on Machine Learning*, 2021.
- [35] H. Zhan, F. Tao, and Y. Cao, “Human-guided robot behavior learning: A gan-assisted preference-based reinforcement learning approach,” *IEEE Robotics and Automation Letters*, vol. 6, pp. 3545–3552, 2020.
- [36] Z. Cao, K. Wong, and C.-T. Lin, “Weak human preference supervision for deep reinforcement learning,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, pp. 5369–5378, 2021.
- [37] J. Park, Y. Seo, J. Shin, H. Lee, P. Abbeel, and K. Lee, “Surf: Semi-supervised reward learning with data augmentation for feedback-efficient preference-based reinforcement learning,” *ArXiv*, vol. abs/2203.10050, 2022.
- [38] W. Zhu, J. Qin, Y. Lou, H. Ye, X. Ma, H. Ci, and Y. Wang, “Social motion prediction with cognitive hierarchies,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [39] D. A. Pomerleau, “Efficient training of artificial neural networks for autonomous navigation,” *Neural computation*, vol. 3, no. 1, pp. 88–97, 1991.
- [40] M. Kearns and S. Singh, “Near-optimal reinforcement learning in polynomial time,” *Machine learning*, vol. 49, pp. 209–232, 2002.
- [41] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot *et al.*, “Mastering the game of go with deep neural networks and tree search,” *nature*, vol. 529, no. 7587, pp. 484–489, 2016.
- [42] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Proximal policy optimization algorithms,” *arXiv preprint arXiv:1707.06347*, 2017.
- [43] S. Fujimoto, H. Hoof, and D. Meger, “Addressing function approximation error in actor-critic methods,” in *International conference on machine learning*. PMLR, 2018, pp. 1587–1596.
- [44] R. Lowe, Y. I. Wu, A. Tamar, J. Harb, O. Pieter Abbeel, and I. Mordatch, “Multi-agent actor-critic for mixed cooperative-competitive environments,” *Advances in neural information processing systems*, vol. 30, 2017.

## APPENDIX

### A. Implementation Details

We train the proposed approach on NVIDIA GeForce RTX 2080 Ti GPUs. Each experiment takes about 12 hours for Feeding task, 24 hours for bed bathing and feeding tasks.

In the experimental section IV, we employ four distinct preference settings VI. *Setting 1* features High Forces Weight and Food Hit Weight, reflecting user preferences against

excessive force and food spillage. *Setting 2* intensifies both weights associated with the robot’s force application, especially emphasizing the High Forces Weight to prevent any force on the body. *Setting 3*, acting as a control group, represents users with minimal preference demands. *Setting 4*, an extreme version of *Setting 1*, significantly values Food Hit Weight while being less concerned with other metrics, suited for users insensitive to force but averse to food spillage.

### B. Anticipation Module



Fig. 5. An example procedure of feeding task. *Left*: Early stage when Robot arm is far from human body. *Mid*: Middle stage when Robot arm is closer to human body. *Right*: Late stage when Robot arm touches human body, completing feeding.

*a) Dynamic Future Mechanism.:* In each task scenario, Assistive Gym sets the duration to 200 time steps. We demonstrate a typical task flow in the feeding task as shown in Figure 5. In the initial stage, as depicted in the left panel of Figure 5, the robot starts at a distance from the target point and needs to move the robotic arm swiftly to a position near the target. This requires anticipation of the human agent’s potential movement over an extended period. Mid-task, as shown in the middle panel of Figure 5, the robot approaches the human agent while preparing to adjust its angle for delivering food, necessitating a shorter foresight than in the initial phase, yet still reliant on future predictions for accurate delivery. In the final stage, illustrated in the right panel of Figure 5, the robot’s arm is close to or in contact with the target point, where long-term predictions are no longer necessary, and only immediate adjustments based on the last 1-2 time steps are required.

This example highlights that different task stages require varying lengths of future predictions. Accordingly, we adaptively adjust the foresight length used by the anticipation module based on the task phase, with specific adjustments as follows: during the first 50 steps of an episode, a foresight length of 10 steps is applied; between steps 50 and 100, the foresight length is increased to 8 steps; and after 100 steps, it is further extended to 5 steps.

### C. Utility Module

*a) Probability Gate Design:* During our experiments, the policy learning objective prioritizes task completion first, followed by meeting human preference demands. To address this, we propose adapting the estimated preference reward for the robot agent only after achieving a certain success rate in evaluation. Specifically, we re-weight the preference estimated reward based on the latest task success rate  $\eta$ .

### D. Training Curves

We provide additional training curves to compare the performance of different methods across various scenarios.

Figure 6 shows the training curves for different assistive tasks, where our method achieves superior performance throughout training. Figure 7 presents the training curves for different robots performing the feeding task, demonstrating the stability and consistent superiority of our method across the training process.

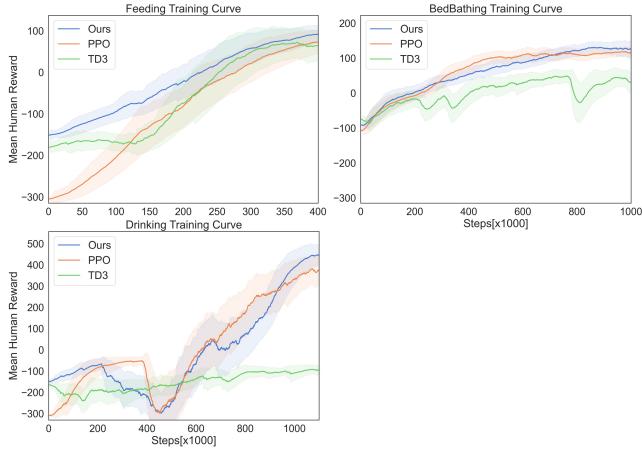


Fig. 6. Training curves of baseline PPO, TD3 and our method in the feeding, bed bathing and drinking task (with robot Sawyer). Our method delivers generally superior performance throughout the training process.

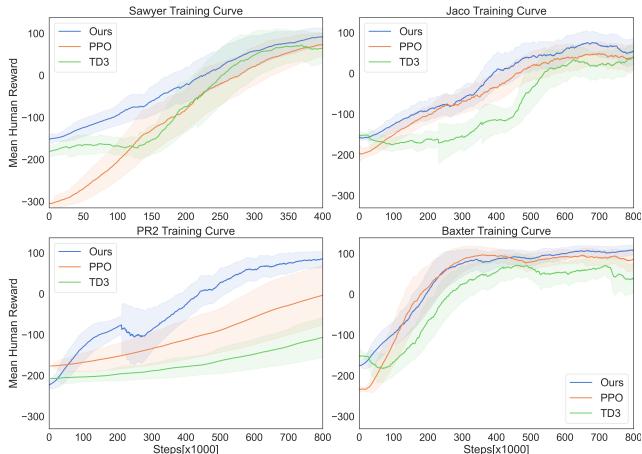


Fig. 7. Training curves of baseline PPO, TD3 and our method in 4 different robot scenarios in the feeding task. Our method consistently delivers superior performance throughout the entire training process.