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

# Robustifying a Policy in Multi-Agent RL with Diverse Cooperative Behaviors and Adversarial Style Sampling for Assistive Tasks

Takayuki Osa[1,2], Tatsuya Harada[1,2]

***Abstract*— Autonomous assistance of people with motor impairments is one of the most promising applications of autonomous robotic systems. Recent studies have reported encouraging results using deep reinforcement learning (RL) in the healthcare domain. Previous studies showed that assistive tasks can be formulated as multi-agent RL, wherein there are two agents: a caregiver and a care-receiver. However, policies trained in multi-agent RL are often sensitive to the policies of other agents. In such a case, a trained caregiver's policy may not work for different care-receivers. To alleviate this issue, we propose a framework that learns a robust caregiver's policy by training it for diverse care-receiver responses. In our framework, diverse care-receiver responses are autonomously learned through trials and errors. In addition, to robustify the care-giver's policy, we propose a strategy for sampling a care-receiver's response in an adversarial manner during the training. We evaluated the proposed method using tasks in an Assistive Gym. We demonstrate that policies trained with a popular deep RL method are vulnerable to changes in policies of other agents and that the proposed framework improves the robustness against such changes.**

## I. Introduction

In the United states, it was reported that approximately 26% of adults have some type of disability, and 3.7% of the 26% that have a form of disability have difficulty in self-care, including behavior such as dressing and bathing [1]. To assist such people with motor impairments, assistive robotics systems have been investigated for decades [2]. Recent advances in machine learning and robotics have developed quickly, and recent studies have demonstrated impressive results for various applications. Reinforcement learning (RL) [3], which is an approach for learning the optimal behavior through autonomous trials and errors, has been applied to a diverse range of applications, including robotic control [4] and autonomous driving [5]. However, although many advancements have been made in RL research, there still exist many challenges in general [6], and assistive robots in particular.

Previous studies in [7], [8] developed a simulator for assistive robots and showed that assistive tasks can be formulated as multi-agent RL, wherein there are two agents: a caregiver and a care-receiver. In their framework, a deep RL method was applied to both agents, and each agent learned natural behavior through autonomous trial and error during the simulation. This framework was referred to as co-optimization [7], [8]. Alternatively, a limitation of this framework is that the performance of the caregiver's policy depends on the care-receiver's policy, and therefore the caregiver's policy may not work if the care-receiver's policy is changed. This vulnerability of a policy is a common problem in multi-agent RL. In the literature of multi-agent RL, it is reported that the performance of the learned policies is often highly sensitive to the policies of other agents [9]. In practice, the behavior of the care-receiver is unknown and diverse in the real world. When we transfer a caregiver's policy trained in simulation to the real world, the behavior of a care-receiver must be different from that obtained with co-optimization in simulation. Therefore, it is essential to consider the robustness of the policy against the change in the care-receiver's policy.

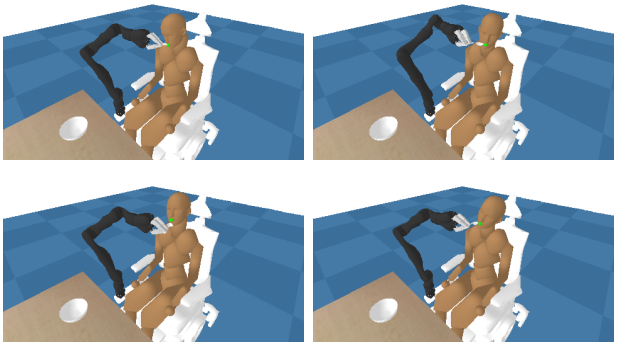


Fig. 1. Diverse care receiver's responses for the feeding task in Assistive Gym [7]. Our framework autonomously learns diverse care receiver's responses and robustifies the caregiver's policy in an adversarial-training fashion.

To address this issue, we propose a framework that robustifies the caregiver's policy by learning diverse behaviors of the care-receiver in co-optimization. Our contribution is to propose a practical algorithm for learning a caregiver's policy that is robust against the change in the care-receiver's behavior for assistive tasks. By training the caregiver's policy for diverse care-receiver's responses, we can ensure that the caregiver's policy is robust against the changes in the care-receiver's behavior. In prior work, diverse care-receiver's responses were obtained by manually designing various reward functions for the care-receiver [10]. By contrast, our framework does not require such reward engineering because diverse care-receiver's responses is autonomously obtained by maximizing the mutual information. Furthermore, we propose to sample the care-receiver's behavior style in an adversarial-training fashion during training. Even if we train the caregiver's policy for diverse care-receiver's responses,

[1]T. Osa and T. Harada are with Graduate School of Information Science and Technology, the University of Tokyo, Tokyo, Japan {osa, harada}@mi.t.u-tokyo.ac.jp
[2]T. Osa and T. Harada are also with RIKEN Center for Advanced Intelligence Project (AIP), Tokyo, Japan.

we observed that uniformly sampling the diverse care-receiver's response did not lead to satisfactory performance of the caregiver's policy. Our strategy for sampling the caregiver's response style can be considered as adversarial training that improves the worst case performance. We evaluated the proposed method using tasks in Assistive Gym [7]. The experimental results shows that caregiver's policies obtained by a standard co-optimization are actually vulnerable to the change in the care-receiver's policy. Our results demonstrate that the caregiver's policy obtained by the proposed framework is more robust against changes in the care-receiver's policy compared with that obtained by co-optimization using a widely-used deep RL method.

## II. Related Work

Assistive robots have been investigated as a promising approach for empowering people with motor impairments [2]. Previous studies addressed the application of assistive robots to tasks such as dressing [8], [11] and feeding [12]–[14]. Within the context of learning assistive tasks, imitation learning [15] and reinforcement learning [3] are two of the most popular approaches. However, a limitation of imitation learning is that collecting expert demonstrations is time-consuming and thousands of demonstrations are often necessary to obtain satisfactory generalization performance [15]. In contrast, the optimal policy is learned through autonomous trials and errors in RL [3]. RL has been successful in various domains, including robotic manipulation [6], board games [16], and autonomous driving [17]. Unlike imitation learning, it is not necessary to provide human demonstrations, and the performance of RL agents often outperforms human experts [18]. However, performing trials and errors in real robotic systems can be costly and risky in practice. To address this issue, Erickson et al. recently developed a simulator, Assistive Gym, which is designed for assistive tasks to accelerate the research in this field [7]. Assistive Gym encompasses various tasks, including dressing, drinking, and feeding. Our work on multi-agent RL for assistive tasks is built on top of Assistive Gym.

Based on tasks in Assistive Gym, Clegg et al. showed that natural assistive motions can be obtained by jointly optimizing policies for a caregiver and care-receiver [8]. However, in their framework, the resulting policy of the caregiver is dependent on the care-receiver's policy and thus may not work for a different care-receiver who have a different behavior style. Recent work by He et al. addressed this issue from a meta-learning approach [10]. They prepared diverse care-receiver's response by manually engineering the reward function for the care-receiver in co-optimization. Subsequently, the latent space of the care-receiver's responses are learned. The caregiver's policy was adapted to the care-receiver's response by estimating the corresponding latent variable through the interaction. Although the approach proposed in [10] is promising, it requires manual engineering of the reward function and running co-optimization multiple times to obtain diverse care-receiver's responses. In contrast, our framework does not require such reward function engineering to obtain diverse care-receiver's response styles because we leverage an algorithm based on mutual information maximization [19]. Furthermore, diverse care-receiver's responses are learned by performing co-optimization just one time in our framework.

## III. Background

### A. Assistive Tasks as Multi-agent Reinforcement Learning

In RL, we consider a Markov decision process (MDP) that consists of a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma, d)$ where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{P}(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t, \boldsymbol{a}_t)$ is the transition probability density, $r(\boldsymbol{s}, \boldsymbol{a})$ is the reward function, $\gamma$ is the discount factor, and $d(\boldsymbol{s}_0)$ is the probability density of the initial state. A policy $\pi(\boldsymbol{a}|\boldsymbol{s}) : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$ is defined as the conditional probability density over actions given the states. The aim of RL is to learn a policy that maximizes the expected return $\mathbb{E}[R_0|\pi]$ where $R_t = \sum_{k=t}^{T} \gamma^{k-t} r(\boldsymbol{s}_k, \boldsymbol{a}_k)$.

In this work, we are particularly interested in a setting in which there are two agents, namely a caregiver and care-receiver, in assistive tasks. Multi-agent RL is often formulated based on Markov games, which is an extension of MDPs [20]. A Markov game for $N$ agents is defined by a state space $\mathcal{S}$, a set of action spaces for $N$ agents $\mathcal{A}_1, \ldots, \mathcal{A}_N$, and a set of observation spaces for $N$ agents $\mathcal{O}_1, \ldots, \mathcal{O}_N$. In our problem setting, the agent $i$ cannot observe the action taken by another agent $j$ for $j \neq i$. This problem setting is referred to as a partially observable Markov game [20], [21]. In multi-agent RL, each agent $i$ tries to maximize the its own expected returns $\mathbb{E}[R_0^i|\pi^i]$.

In our problem setting, the two agents share the same reward function and aim to achieve successful and comfortable care. Therefore, the two agents *cooperate* to achieve the same goal [21], [22]. In this study, we propose a framework for learning diverse behaviors in a cooperative multi-agent RL setting. We denote the caregiver's action and state by $\boldsymbol{a}_{\mathrm{g}} \in \mathcal{A}_{\mathrm{g}}$ and $\boldsymbol{s}_{\mathrm{g}} \in \mathcal{S}_{\mathrm{g}}$, respectively. Similarly, we denote the care-receiver's action and state by $\boldsymbol{a}_{\mathrm{r}} \in \mathcal{A}_{\mathrm{r}}$ and $\boldsymbol{s}_{\mathrm{r}} \in \mathcal{S}_{\mathrm{r}}$, respectively. We also denote the policies of the caregiver and care-receiver by $\pi^{\mathrm{g}}$ and $\pi^{\mathrm{r}}$, respectively.

### B. Latent-conditioned policies for modeling diverse behaviors

Previous studies [23]–[25] have showed that diverse behaviors can be represented by a policy conditioned on a latent variable $\pi(\boldsymbol{a}|\boldsymbol{s}, \boldsymbol{z})$, where $\boldsymbol{z}$ is the latent variable. They proposed methods for training the latent-conditioned policy such that it changes the output according to the value of the latent variable $\boldsymbol{z}$. When we have a latent-conditioned policy such that it changes the output according to the value of the latent variable $\boldsymbol{z}$, sampling a policy from a distribution over policies can be approximated with sampling a value of latent variable $\boldsymbol{z}$. The value of the latent variable is sampled at the beginning of an episode and fixed until the end of the episode. In this framework, latent-conditioned policy $\pi(\boldsymbol{a}|\boldsymbol{s}, \boldsymbol{z})$ is evaluated based on the state function conditioned on the latent variable defined by

$$V^{\pi}(\boldsymbol{s}, \boldsymbol{z}) = \mathbb{E}_{\pi}[R|\boldsymbol{s}, \boldsymbol{z}], \tag{1}$$

which represents the expected return when starting in state $\boldsymbol{s}$ and following policy $\pi$ given the latent variable $\boldsymbol{z}$.

## IV. Robusfitying Policies in Assistive Tasks

In this section, we first consider how to train the caregiver's policy for diverse care-receiver's response. Subsequently, we present how to obtain diverse care-receiver's responses during training. Then, we describe the adversarial style sampling to robustify the caregiver's policy.

### A. Training Caregiver's Policy for Diverse Care-Receiver's responses

In our framework, we aim to obtain a robust caregiver's policy by training it with diverse care-receiver's behavior. Assuming that the distribution of the care-receiver's policy $\pi^{\mathrm{r}}$ is given by $p(\pi^{\mathrm{r}})$, the following expected return should be maximized to perform the task appropriately:

$$\max_{\pi^g, \pi^r} \mathbb{E}_{\pi^{\mathrm{r}} \sim p(\pi^{\mathrm{r}})} \left[ \mathbb{E}_{\boldsymbol{s}' \sim \mathcal{P}} [R | \pi^{\mathrm{g}}, \pi^{\mathrm{r}}] \right]. \tag{2}$$

We approximate this problem using the latent-conditioned policy, which we introduced in the previous section. As we can specify the type of the behavior of the latent-conditioned policy by setting the value of the latent variable, we can approximate the expected return in (2) with

$$\max_{\pi^g, \pi^r} \mathbb{E}_{\boldsymbol{z}_{\mathrm{r}} \sim p(\boldsymbol{z}_{\mathrm{r}})} \left[ \mathbb{E}_{\boldsymbol{s} \sim \mathcal{P}} [R | \pi^{\mathrm{g}}(\boldsymbol{a}_{\mathrm{g}} | \boldsymbol{s}_{\mathrm{g}}), \pi^{\mathrm{r}}(\boldsymbol{a}_{\mathrm{r}} | \boldsymbol{s}_{\mathrm{r}}, \boldsymbol{z}_{\mathrm{r}})] \right], \tag{3}$$

where $\boldsymbol{z}_{\mathrm{r}}$ is the latent variable that specifies the behavior of the care-receiver, respectively, and $p(\boldsymbol{z}_{\mathrm{r}})$ is the prior distributions of $\boldsymbol{z}_{\mathrm{r}}$. The expectation in (3) can be approximated using samples stored in the replay buffer in an off-the-shelf RL algorithm. In this study, we use PPO as a base RL algorithm [26], although other RL algorithms can also be used. We denote by by $\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{s}_{\mathrm{g}})$ the caregiver's deterministic policy parameterized with a vector $\boldsymbol{\theta}$. Once we collect samples through the interaction between the caregiver's policy and diverse care-receiver's responses, we update the caregiver's policy to maximize the expected return.

### B. Learning diverse care-receiver's responses

To learn diverse behaviors of the care-receiver, we extend PPO [26] to a method that trains a latent-conditioned policy by maximizing the mutual information between the latent variable and the state-action pairs. In the context of multi-agent RL, it is reported that training each agent using PPO often leads the performance better than state-of-the-art multi-agent RL methods [27]. To train a latent-conditioned policy, we consider the problem of maximizing the following objective function, which can be obtained by extending the one in [26].

$$\mathcal{L}_{\mathrm{adv}}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{z}, \boldsymbol{s}_{\mathrm{r}}, \boldsymbol{a}_{\mathrm{r}} \sim p(\boldsymbol{z}), d^{\pi_{\mathrm{old}}}, \pi_{\mathrm{old}}} \left[ L_{\mathrm{clip}}(\boldsymbol{s}_{\mathrm{r}}, \boldsymbol{a}_{\mathrm{r}}, \boldsymbol{z}) \right], \tag{4}$$

subject to

$$\mathbb{E}_{\boldsymbol{s} \sim d^{\pi}} \left[ D_{\mathrm{KL}} \left( \pi^{\mathrm{r}}_{\mathrm{old}}(\boldsymbol{a} | \boldsymbol{s}, \boldsymbol{z}) || \pi^{\mathrm{r}}_{\boldsymbol{\theta}}(\boldsymbol{a} | \boldsymbol{s}, \boldsymbol{z}) \right) \right] < \eta, \tag{5}$$

where

$$L_{\mathrm{clip}}(\boldsymbol{s}_{\mathrm{r}}, \boldsymbol{a}_{\mathrm{r}}, \boldsymbol{z}) = \min \left( r(\boldsymbol{\theta}) A_t, \tilde{r}(\boldsymbol{\theta}) A_{\mathrm{r}} \right), \tag{6}$$

$$r(\boldsymbol{\theta}) = \frac{\pi^{\mathrm{r}}_{\boldsymbol{\theta}}(\boldsymbol{a} | \boldsymbol{s}, \boldsymbol{z})}{\pi^{\mathrm{r}}_{\mathrm{old}}(\boldsymbol{a} | \boldsymbol{s}, \boldsymbol{z})}, \tag{7}$$

$$\tilde{r}(\boldsymbol{\theta}) = \mathrm{clip}(r(\boldsymbol{\theta}), 1 - c, 1 + c), \tag{8}$$

$c$ is a constant, and $A_{\mathrm{r}}$ is the advantage function for the care-receiver's policy. The constraint in (5) constrains the change of the policy conditioned on the latent variable stabilizes the learning process. While solving the above problem leads to obtain a latent-conditioned policy that maximizes the expected return, the diversity of the behaviors encoded in the policy is not encouraged.

To encourage the diversity of the behaviors, we maximize the lower bound of the mutual information between the latent variable and the state-action pairs induced by policy $\pi$, which we denote by $I_{\pi}(\boldsymbol{z}; \boldsymbol{s}, \boldsymbol{a})$. As shown in [19], the variational lower bound of $I_{\pi}(\boldsymbol{z}; \boldsymbol{s}, \boldsymbol{a})$ is given by

$$I_{\pi}(\boldsymbol{z}; \boldsymbol{s}, \boldsymbol{a}) \geq \mathbb{E}_{(\boldsymbol{s}, \boldsymbol{a}) \sim \pi, \mathcal{P}} [\log q_{\boldsymbol{\phi}}(\boldsymbol{z} | \boldsymbol{s}, \boldsymbol{a})], \tag{9}$$

where $q_{\boldsymbol{\phi}}(\boldsymbol{z} | \boldsymbol{s}, \boldsymbol{a})$ is an approximated posterior distribution parameterized with a vector $\boldsymbol{\phi}$. As the right-hand side of (9) is a simple log-likelihood, maximizing this lower bound is tractable in practice. Based on (9), we train the care-receiver's policy maximizing the following objective function:

$$\mathcal{L}_{\mathrm{LPPO}}(\boldsymbol{\theta}) = \mathcal{L}_{\mathrm{adv}}(\boldsymbol{\theta}) + \alpha \mathbb{E}_{\boldsymbol{s}_{\mathrm{r}}, \boldsymbol{a}_{\mathrm{r}} \sim \pi^{\mathrm{r}}, \mathcal{P}} [\log q_{\boldsymbol{\phi}}(\boldsymbol{z}_{\mathrm{r}} | \boldsymbol{s}_{\mathrm{r}}, \boldsymbol{a}_{\mathrm{r}})] \tag{10}$$

subject to

$$\mathbb{E}_{\boldsymbol{s} \sim d^{\pi}} \left[ D_{\mathrm{KL}} \left( \pi_{\mathrm{old}}(\boldsymbol{a} | \boldsymbol{s}, \boldsymbol{z}) || \pi_{\boldsymbol{\theta}}(\boldsymbol{a} | \boldsymbol{s}, \boldsymbol{z}) \right) \right] < \eta, \tag{11}$$

where $\alpha$ is a constant that balance the weight on the expected return and mutual information terms. We refer to the resulting algorithm as Latent-conditioned Proximal Policy Optimization (LPPO).

### C. Adversarial Style Sampling

When we design a practical algorithm to solve the problem in (3), the choice of $p(\boldsymbol{z}_{\mathrm{r}})$ is crucial to obtain a satisfactory performance of the caregiver's policy. In the context of learning diver solutions in RL, a popular choice is to use the uniform distribution as $p(\boldsymbol{z}_{\mathrm{r}})$ [24]. If we use the uniform distribution as $p(\boldsymbol{z}_{\mathrm{r}})$, the average performance over the diverse care-receiver's responses is maximized. Although maximizing the average performance is reasonable, the worst-case performance is not considered in this case. As indicated in the literature of risk-aware RL [28], it is often necessary to improve the worst-case performance to obtain a robust policy.

Based on the above consideration, we propose to solve the following max-min problem:

$$\max_{\pi^g, \pi^r} \min_{\tilde{p}(\boldsymbol{z}_{\mathrm{r}})} \mathbb{E}_{\boldsymbol{z}_{\mathrm{r}} \sim \tilde{p}(\boldsymbol{z}_{\mathrm{r}})} \left[ \mathbb{E}_{\boldsymbol{s} \sim \mathcal{P}} [R | \pi^{\mathrm{g}}(\boldsymbol{a}_{\mathrm{g}} | \boldsymbol{s}_{\mathrm{g}}), \pi^{\mathrm{r}}(\boldsymbol{a}_{\mathrm{r}} | \boldsymbol{s}_{\mathrm{r}}, \boldsymbol{z}_{\mathrm{r}})] \right], \tag{12}$$

which can be viewed as a type of adversarial training in which the caregiver's and care-receiver's policies $\pi^g$, $\pi^r$ are

updated so as to maximize the expected rewards, and the behavior style sampler $\tilde{p}(\boldsymbol{z}_\mathrm{r})$ is set to minimize the expected reward during training. In this problem formulation, the agent attempts to maximize the worst-case performance, leading to improve the robustness of the caregiver's policy.

To solve the problem in (12), we sample the latent variable as follows during training:

$$\boldsymbol{z}_\mathrm{r} = \arg\min_{\tilde{\boldsymbol{z}}_\mathrm{r}} \mathbb{E}_{\boldsymbol{s}\sim\mathcal{P}}[R|\pi^\mathrm{g}(\boldsymbol{a}_\mathrm{g}|\boldsymbol{s}_\mathrm{g}), \pi^\mathrm{r}(\boldsymbol{a}_\mathrm{r}|\boldsymbol{s}_\mathrm{r}, \tilde{\boldsymbol{z}}_\mathrm{r})]. \tag{13}$$

As the value of the latent variable specifies the behavior style of the care-receiver, this approach enable us to select the behavior style of the care-receiver that lead to the worst performance during the training. In our framework, we consider cooperative tasks where the reward is shared between the caregiver and care-receiver. Therefore, the approximated latent-conditioned state value, $V^\mathrm{r}_{\boldsymbol{w}}(\boldsymbol{s}_\mathrm{r}, \boldsymbol{z}_\mathrm{r})$, indicates the expected return when the care-receiver takes the response corresponding to the value of $\boldsymbol{z}_\mathrm{r}$ under state $\boldsymbol{s}_\mathrm{r}$. In our implementation, the value of the latent variable $\boldsymbol{z}_\mathrm{r}$ is sampled as follows:

$$\boldsymbol{z}_\mathrm{r} = \arg\min_{\boldsymbol{z}_\mathrm{r}} \frac{1}{N} \sum_{(\boldsymbol{s}_\mathrm{r}, \boldsymbol{z}_\mathrm{r})\in\mathcal{B}} V^\mathrm{r}_{\boldsymbol{w}}(\boldsymbol{s}_\mathrm{r}, \boldsymbol{z}_\mathrm{r}). \tag{14}$$

When latent variable $\boldsymbol{z}_\mathrm{r}$ is continuous, it is not feasible to analytically perform the minimization in (14). In practice, we generate $M$ samples of latent variable $\boldsymbol{z}_\mathrm{r}$ from the uniform distribution $U(-1, 1)$, and use the sample with the lowest value as the minimizer. Meanwhile, to encourage the diversity of the care-receiver's response, it is also necessary to sample a wide range of values of $\boldsymbol{z}_\mathrm{r}$ during training. Thus, in practice, we sample the value of $\boldsymbol{z}_\mathrm{r}$ in a $\epsilon$-greedy-like fashion; with probability $\epsilon$, the value of $\boldsymbol{z}_\mathrm{r}$ is determined by (14). Otherwise, the value of $\boldsymbol{z}_\mathrm{r}$ is sampled from the uniform distribution $U(-1, 1)$.

### D. Practical algorithm

Our algorithm is summarized in Algorithm 1. The caregiver and care-receiver retain separate replay buffers for each. Latent variable $\boldsymbol{z}_\mathrm{r}$ that specifies the care-receiver's behavior style is stored in the care-receiver's replay buffer. The value of the latent variable $\boldsymbol{z}_\mathrm{r}$ is sampled at the beginning of an episode and fixed until the end of the episode. We set $\epsilon = 0.5$ for the adversarial style sampling in our implementation. When training the care-receiver's policy, we set $\alpha = 0.2$ in (10).

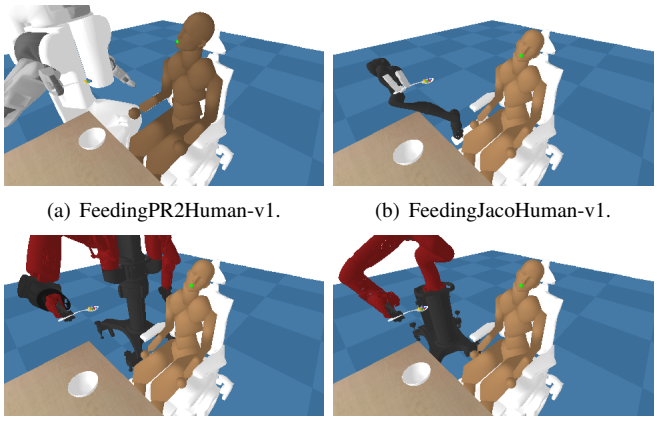
## V. Evaluation

In the experiment, we investigated the following points: 1) sample-efficiency of the propose method in the training, and 2) robustness of the caregiver's policy against the change in the care-receiver's policy.

We evaluated the proposed method using tasks implemented in Assistive Gym [7], which is based on the PyBullet physics engine [29]. We used FeedingPR2Human-v1, FeedingJacoHuman-v1, FeedingBaxterHuman-v1, and FeedingSawyerHuman-v1 in our evaluation, as shown in Fig. 2. Each episode consists of 200 time steps. Observations

**Algorithm 1** Robustifying the caregiver's policy with diverse care-receiver's response and adversarial style sampling

**Input:** Dimension of latent variable $\boldsymbol{z}_\mathrm{r}$ for the care-receiver's policy, $\epsilon$ for adversarial style sampling
Initialize policies $\pi_\mathrm{g}$, $\pi_\mathrm{r}$ and buffers $\mathcal{D}_\mathrm{g}$, $\mathcal{D}_\mathrm{r}$
**repeat**
  **while** the data size in the buffers is not sufficient **do**
    sample $x_\mathrm{rng}$ from the uniform distribution $U(0, 1)$
    **if** $x_\mathrm{rng} < \epsilon$ **then**
      Set the latent variable $\boldsymbol{z}_\mathrm{r}$ with (14)
    **else**
      Sample $\boldsymbol{z}_\mathrm{r}$ from the uniform distribution $U(-1, 1)$
    **end if**
    **for** $t = 0$ **to** $T$ **do**
      Select actions with exploration noise for each agent
      Observe reward $r$ and new state $\boldsymbol{s}'_\mathrm{g}$ and $\boldsymbol{s}'_\mathrm{r}$
      Store tuple $(\boldsymbol{s}_\mathrm{g}, \boldsymbol{a}_\mathrm{g}, \boldsymbol{s}'_\mathrm{g}, r)$ in $\mathcal{D}_\mathrm{g}$
      Store tuple $(\boldsymbol{s}_\mathrm{r}, \boldsymbol{a}_\mathrm{r}, \boldsymbol{s}'_\mathrm{r}, r, \boldsymbol{z}_\mathrm{r})$ in $\mathcal{D}_\mathrm{r}$
    **end for**
  **end while**
  Update the care-receiver's policy by maximizing $\mathcal{L}_\mathrm{LPPO}(\boldsymbol{\phi})$ in (10)
  Update the caregiver's policy with PPO
  Empty the buffers $\mathcal{D}_\mathrm{g}$ and $\mathcal{D}_\mathrm{r}$
**until** $\pi_\mathrm{g}$ and $\pi_\mathrm{r}$ are optimized



(a) FeedingPR2Human-v1. (b) FeedingJacoHuman-v1. (c) FeedingBaxterHuman-v1. (d) FeedingSawyerHuman-v1.

Fig. 2. Tasks in Assistive Gym used in the evaluation.

of the caregiver include the position and orientation of the spoon and head of the care-receiver and the joint angles of the caregiver. Observations of the care-receiver include the position and orientation of the spoon and head of the care-receiver and the joint angles of the care-receiver. The reward function is computed based on the distance between the spoon and the mouth, the state of the food (e.g., spilled or not), and the velocity of the end-effector. For more details on Assistive Gym, please refer to [7].

To investigate the effect of learning diverse care-receiver's response and the adversarial style sampling, we evaluated the following methods. As a baseline, we evaluate policy performance in the case where both caregiver's and care-receiver's policies were trained with PPO. We refer to this baseline as *PPO-PPO*. Similarly, we refer to the case where

TABLE I

HYPERPARAMETERS FOR PPO/LPPO

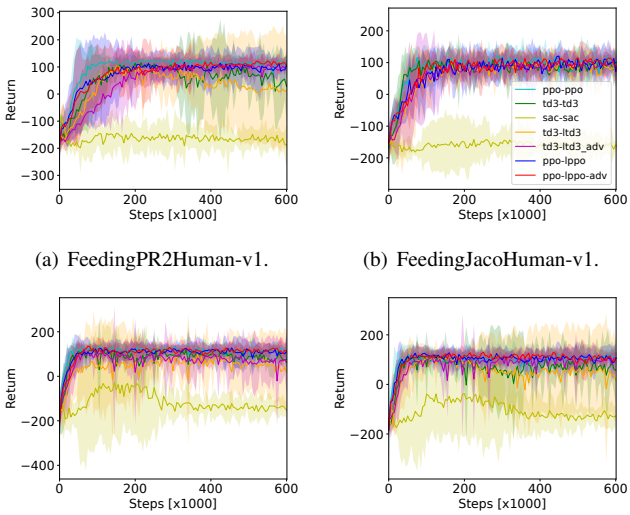
| Parameter | Value | Method |
|---|---|---|
| Optimizer | Adam | PPO/LPPO |
| Policy learning rate | $3 \cdot 10^{-4}$ | PPO/LPPO |
| Critic learning rate | $1 \cdot 10^{-3}$ | PPO/LPPO |
| Discount factor $\gamma$ | 0.99 | PPO/LPPO |
| Steps per epoch | 4000 | PPO/LPPO |
| Number of hidden layers | 2 | PPO/LPPO |
| Number of hidden units | (64, 64) | PPO |
| Number of hidden units | (128, 64) | LPPO |
| Activation function | tanh | PPO/LPPO |
| Coefficient for GAE $\lambda$ | 0.95 | PPO/LPPO |
| Clip ratio $c$ | 0.2 | PPO/LPPO |
| Target KL | 0.01 | PPO/LPPO |
| $\alpha$ | 0.2 | LPPO |

both caregiver's and care-receiver's policies were trained with TD3 [30] as *TD3-TD3*. TD3-TD3 and PPO-PPO can be considered as baselines based on a standard co-optimization proposed in [7]. To investigate the effect of learning diverse care-receiver's responses, we evaluated policy performance in the case where the caregiver's policy was trained with PPO and the care-receiver's policy with LPPO as described in Section IV-B. we refer to this variant of the proposed method as *PPO-LPPO*. The differences in policy performance between PPO-PPO and PPO-LPPO indicates the effect of learning diverse care-receiver's responses during training. Finally, we evaluated the proposed method that incorporates the adversarial style sampling with PPO-LPPO, which is referred to as *PPO-LPPO-adv*. To investigate the effect of the algorithm for learning diverse behaviors, we also evaluated variants using LTD3 [19], which is an existing method for learning diverse behaviors in RL. TD3-LTD3 refers to the method where the caregiver's policy was trained with TD3 and the care-receiver's policy with LTD3. Similarly, TD3-LTD3-adv refers to the method that incorporates the proposed adversarial style sampling with TD3-LTD3. For LPPO and LTD3, the latent variable of the care-receiver's policy was two-dimensional. In LPPO and LTD3, we used the uniform distribution $U(-1, 1)$ as the prior distribution of the latent variable $p(\boldsymbol{z})$. The implementation of PPO and TD3 were adapted from spinningup [31]. Hyperparameters of PPO and LPPO are summarized in Table I.

### A. Learning Curve

The learning curves during the traing are shown in Fig. 3. Regarding FeedingPR2Human-v1, while the performance of TD3-TD3 and TD3-LTD3 often dropped after 4 million steps, the performances of PPO-LPPO-adv, PPO-LPPO and PPO-PPO were stable during training. The difference between LTD3-based methods and LPPO-based methods implies that LPPO-based methods are more stable in cooperative multi-agent RL. Interestingly, there was no significant difference in the performance and sample-efficiency among PPO-LPPO-adv, PPO-LPPO and PPO-PPO. This result indicates that learning diverse care-receiver's responses in PPO-LPPO, and PPO-LPPO-adv does not have a significant effect on the sample-efficiency of the training in these tasks, although LPPO learns diverse behaviors of the care-receiver.

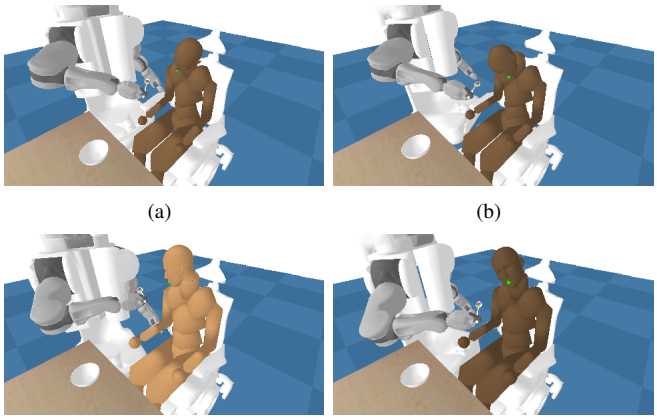
We visualized the behavior of policies obtained in PPO-




(a) FeedingPR2Human-v1. (b) FeedingJacoHuman-v1.

(c) FeedingBaxterHuman-v1. (d) FeedingSawyerHuman-v1.

Fig. 3. Learning curves of the proposed and baseline methods.



(a) (b)

(c) (d)

Fig. 4. Diverse behaviors of the care-receiver obtained for the FeedingPR2Human-v1 task. The orientation of the care-receiver's head changes according to the value of the latent variable. (a)-(d) correspond to $\boldsymbol{z}_r = [0.9, 0.9], [-0.9, 0.9], [0.9, -0.9], [-0.9, -0.9]$, respectively. Human size and color are randomly set.

LPPO-adv after training with 4 million time steps. We obtained diverse behaviors of the care-receiver in PPO-LPPO-adv, as shown in Fig. 4. As shown in Fig. 4, the care-receiver cooperated with the caregiver by moving the head towards the spoon in different ways. This result implies that there can be diverse behaviors of the care-receiver and that the caregiver's policy should be robust against the diversity of the care-receiver's behavior. It is worth noting that it is not trivial to hard-code the reward function to obtain diverse care-receiver's responses shown in Fig. 4. The diversity of the care-receiver's responses shown in Fig. 4 supports the validity of our framework based on mutual information maximization for learning diverse behaviors.

### B. Robustness Against the Changes in the Care-Receiver's Policy

To investigate the robustness of the caregiver's policy against the change in the care-receiver's policy, we evaluated

TABLE II

RETURNS IN COLLABORATION WITH THE CARE-RECEIVER TRAINED SEPARATELY WITH TD3

| Methods | FeedingPR2Human-v1 | | FeedingJacoHuman-v1 | | FeedingBaxterHuman-v1 | | FeedingSawyerHuman-v1 | |
|---|---|---|---|---|---|---|---|---|
| | training | test | training | test | training | test | training | test |
| PPO-LPPO-adv (ours) | **114.3±36.1** | **77.1±67.8** | **101.0±71.0** | **89.1±82.4** | **120.5±40.7** | **104.7±47.9** | **111.7±54.6** | 87.8±68.3 |
| PPO-LPPO (ours) | 98.4±57.4 | **87.8±62.4** | **97.5±80.5** | **84.6±92.2** | 103.9±50.7 | **103.0±63.4** | **115.9±44.3** | 79.1±79.1 |
| PPO-PPO | **114.9±33.4** | 47.0±81.8 | 107.7±71.6 | 81.7±84.8 | 105.9±48.4 | 60.3±73.1 | 107.1±51.9 | 49.1±95.9 |
| TD3-LTD3-adv (ours) | **112.8±29.4** | 74.1±67.6 | 77.4±69.6 | **91.6±60.3** | 68.7±70.4 | 41.4±100.0 | 92.9±54.0 | **100.0±49.8** |
| TD3-LTD3 | 88.8±63.4 | 52.8±88.8 | 72.8±84.7 | 65.1±82.8 | 44.7±98.8 | 20.8±98.0 | **113.8±26.3** | 87.1±62.5 |
| TD3-TD3 | 104.0±59.4 | 10.3±97.2 | 66.6±91.4 | 61.1±85.2 | 92.8±59.1 | 72.7±76.2 | 87.4±67.0 | 47.8±100.8 |

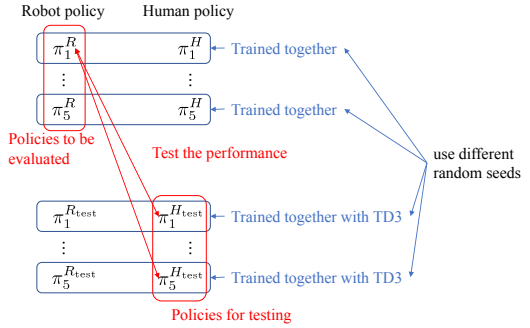



Fig. 5. Procedure for evaluating the robustness of the caregiver's policy. A caregiver's policy was evaluated using a care-receiver's policy, which was trained separately.

the performance of the caregiver's policy when it was used with the care-receiver's policy that was separately trained with TD3. The evaluation procedure is summarized in Fig. 5. For PPO-LPPO-adv, PPO-LPPO and PPO-PPO, policies after training with 6 million steps were evaluated. For TD3-TD3 and TD3-LTD3, policies after training with 2 million steps were evaluated because the policy performance got worse after 2 million steps. We first train the caregiver's and care-receiver's policies using five different seeds with the proposed and baseline methods. In the second step, we prepare another set of the caregiver's and care-receiver's policies trained using five different seeds with TD3-TD3. We then evaluate the performance of the caregiver's policy trained in the first step when it is used with the care-receiver's policy trained in the second step. For comparison, we also show the performance of the caregiver's policy when it is used with the care-receiver's policy, which was trained together with it in the first step. We performed 10 test episodes for each combination of policies, and report the average and standard deviation across five random seeds.

The results are shown in Table II. The column of "train" shows the policy performance when working with an agent trained together, and the column of "test" shows the performance when a caregiver's policy was used with a care-receiver's policy that was separately trained with TD3. The bold text shows the best results in each task. We examined the statistical difference based on unpaired t-test. When there are bold and non-bold numbers, it indicates that the there is a statistically significant difference between them, and p-value is less than 0.05. In baseline methods such as TD3-TD3 and PPO-PPO, the test performance was significantly lower than the training performance. This result demonstrate that caregiver's policies obtained by a standard co-optimization are actually vulnerable to the change in the care-receiver's policies. In contrast, the proposed method, PPO-LPPO-adv, clearly outperformed PPO-PPO in terms of the test performance, and the difference between the training and test performance was small in PPO-LPPO-adv. This result demonstrate that the proposed method significantly improved the robustness of the caregiver's policy. The difference between PPO-PPO and PPO-LPPO indicates the learning diverse behaviors improves the robustness of the caregiver's policy. Furthermore, the difference between PPO-LPPO-adv and PPO-LPPO indicates the adversarial style sampling improves the robustness of the caregiver's policy. The comparison between TD3-TD3, TD3-LTD3, and TD3-LTD3-adv aligns with this observation. TD3-LTD3-adv outperformed TD3-LTD3, indicating the effectiveness of the proposed adversarial style sampling. The learning of diverse care-receiver's behavior and adversarial style sampling improved the robustness of the caregiver's policy in TD3-based methods, whereas the PPO-based methods outperformed the TD3-based methods in our evaluation.

## VI. CONCLUSIONS

We presented a framework for robustifying a cooperative policy in multi-agent RL for assistive tasks. In our framework, diverse care-receiver's responses are learned autonomously by maximizing the mutual information, and the caregiver's policy is robustified by generating care-receiver's responses in an adversarial manner during the training. The proposed algorithm was evaluated in robotic assistive tasks implemented in Assistive Gym. The experimental results showed that caregiver's policies obtained by standard co-optimization are vulnerable to the change in the care-receiver's policy. The results also demonstrate that a caregiver's policy obtained by the proposed framework is more robust against changes in the care-receiver's policy. In future work, we will address challenges to be resolved to deploy a caregiver's policy in the real world. Additionally, we plan to study other types of robustness in the future, such as those reported in [10], [32].

## ACKNOWLEDGMENT

This work was partially supported by JST AIP Acceleration Research JPMJCR20U3, Moonshot R&D Grant Number JPMJPS2011, CREST Grant Number JPMJCR2015, JSPS KAKENHI Grant Number JP19H01115, JP23K18476 and Basic Research Grant (Super AI) of Institute for AI and Beyond of the University of Tokyo.

## REFERENCES


[1] C. A. Okoro, N. D. Hollis, A. C. Cyrus, and S. Griffin-Blake, "Prevalence of disabilities and health care access by disability status and type among adults – united states, 2016," *Morbidity and Mortality Weekly Report (MMWR)*, vol. 67, pp. 882–887, 2018.

[2] 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, H. Nguyen, A. Paepcke, C. Pantofaru, W. D. Smart, and L. Takayama, "Robots for humanity: using assistive robotics to empower people with disabilities," *IEEE Robotics & Automation Magazine*, vol. 20, no. 1, pp. 30–39, 2013.

[3] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. MIT Press, 2018.

[4] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-end training of deep visuomotor policies," *Journal of Machine Learning Research*, vol. 17, no. 39, pp. 1–40, 2016.

[5] A. Kendall, J. Hawke, D. Janz, P. Mazur, D. Reda, J.-M. Allen, V.-D. Lam, A. Bewley, and A. Shah, "Learning to drive in a day," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2019.

[6] J. Ibarz, J. Tan, C. Finn, M. Kalakrishnan, P. Pastor, and S. Levine, "How to train your robot with deep reinforcement learning: lessons we have learned," *The International Journal of Robotics Research*, vol. 40, no. 4–5, pp. 698–721, 2021.

[7] Z. Erickson, V. Gangaram, A. Kapusta, C. K. Liu, and C. C. Kemp, "Assistive gym: A physics simulation framework for assistive robotics," *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2020.

[8] 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.

[9] Y. Li, J. Song, and S. Ermon, "InfoGAIL: Interpretable imitation learning fromvisual demonstrations," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[10] J. Z.-Y. He, A. Raghunathan, D. S. Brown, Z. Erickson, and A. D. Dragan, "Learning representations that enable generalization in assistive tasks," in *Proceedings of Conference on Robot Learning (CoRL)*, 2022.

[11] F. Zhang, A. Cully, and Y. Demiris, "Probabilistic real-time user posture tracking for personalized robot-assisted dressing," *IEEE Transactions on Robotics*, vol. 35, no. 4, pp. 873–888, 2019.

[12] T. Rhodes and M. Veloso, "Robot-driven trajectory improvementfor feeding tasks," in *Proceedings of the IEEE/RSJ International Conference onIntelligent Robots and Systems (IROS)*, 2018.

[13] D. Park, Y. Hoshi, and C. C. Kemp, "A multimodal anomaly detectorfor 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, "Transferdepends on acquisition: Analyzing manipulation strategies for robotic feeding," in *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2019.

[15] T. Osa, J. Pajarinen, G. Neumann, J. A. Bagnell, P. Abbeel, and J. Peters, "An algorithmic perspective on imitation learning," *Foundations and Trends®in Robotics*, vol. 7, no. 1-2, pp. 1–179, 2018.

[16] J. Schrittwieser, I. Antonoglou, T. Hubert, K. Simonyan, L. Sifre, S. Schmitt, A. Guez, E. Lockhart, D. Hassabis, T. Graepel, T. Lillicrap, and D. Silver, "Mastering atari, go, chess and shogi by planning with a learned model," *Nature*, 2020.

[17] B. R. Kiran, I. Sobh, V. Talpaert, P. Mannion, A. A. A. Sallab, S. Yogamani, and P. Pérez, "Deep reinforcement learning for autonomous driving: A survey," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–18, 2021.

[18] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. van den Driessche, T. Graepel, and D. Hassabis, "Mastering the game of go without human knowledge," *Nature*, 2017.

[19] 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.

[20] M. L. Littman, "Markov games as a framework for multi-agent reinforcement learning." in *Proceedings of the International Conference on Machine Learning (ICML)*, 1994.

[21] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[22] J. Ackermann, V. Gabler, T. Osa, and M. Sugiyama, "Reducing overestimation bias in multi-agent domains using double centralized critics," in *the Deep Reinforcement Learning Workshop at NeurIPS 2019*, 2019.

[23] B. Eysenbach, A. Gupta, J. Ibarz, and S. Levine, "Diversity is all you need: Learning skills without a reward function," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

[24] S. Kumar, A. Kumar, S. Levine, and C. Finn, "One solution is not all you need:few-shot extrapolation via structured maxent rl," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[25] A. Sharma, S. Gu, S. Levine, V. Kumar, and K. Hausman, "Dynamics-aware unsupervised discovery of skills," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

[26] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," in *arXiv*, 2017.

[27] C. Yu, A. Velu, E. Vinitsky, J. Gao, Y. Wang, A. Bayen, and Y. Wu, "The surprising effectiveness of ppo in cooperativemulti-agent games," in *ArXiv*, 2021.

[28] C. Bodnar, A. Li, K. Hausman, P. Pastor, and M. Kalakrishnan, "Quantile qt-opt for risk-awarevision-based robotic grasping," in *Robotics and Science and Systems*, 2020.

[29] E. Coumans and Y. Bai, "PyBullet, a python module for physics simulation for games, robotics and machine learning," http://pybullet.org, 2016–2021.

[30] S. Fujimoto, H. van Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *Proceedings of the International Conference on Machine Learning (ICML)*, vol. 80, 2018, pp. 1587–1596.

[31] J. Achiam, "Spinning Up in Deep Reinforcement Learning," 2018.

[32] A. Bukharin, Y. Li, Y. Yu, Q. Zhang, Z. Chen, S. Zuo, C. Zhang, S. Zhang, and T. Zhao, "Robust multi-agent reinforcement learning viaadversarial regularization: Theoretical foundationand stable algorithms," in *Advances in Neural Information Processing Systems*, 2023.