LEARNING HIERARCHICAL MULTI-AGENT COOPERATION WITH LONG SHORT-TERM INTENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Communication is a significant method to relieve partial observable and non-stationary problems in Multi-Agent System. However, most of existing work needs to persistently communicate by exchanging partial observation or latent embeddings (intention) which is unrealistic in real-world settings. To overcome this problem, we propose learning hierarchical multi-agent cooperation with long short-term intention (HLSI), a hierarchical multi-agent cooperation framework. In our work, each agent communicates by sharing high-level policy's latent embeddings (long-term intention) which keeps contant until macro action change. To make the communication messages contain more useful content, we maximize mutual information between agent's macro action and agent's future trajectory conditioned on historical trajectory. Agent integrates these messages through the attention mechanism. Then, long short-term intention fusion module will fuse the long-term intention received from other agents and short-term intention inferred by a behaivor inference network to approximate other agents' real short-term intention, which helps agent better understand others' next behavior. We provide comprehensive evaluation and ablations studies in multi-agent cooperative settings. The results show that our method achieves better performance than other multi-agent communication and hierarchical multi-agent reinforcement learning baselines.

1 INTRODUCTION

Multi-Agent Reinforcement Learning (MARL) receives more and more attention in many real-world scenes (game playing(P. Peng & Wang (2017));traffic light control (Wiering (2000));auto driving (Shalev-Shwartz S (2016)). Differing from single agent control problems, Multi-agent control tasks suffer from partially observable and non-stationary problems which make it harder than single agent reinforcement learning. To overcome these problems, centralized training and decentralized execution (CTDE) paradigm is proposed to solve these problems. Nevertheless, many baselines such as MADDPG (Lowe et al. (2017)), COMA (Foerster et al. (2018)), VDN (Sunehag et al. (2017)) based on CTDE paradigm are still struggled in coping with these problems due to the partially observable environments during execution stage. Communication is one of the significant methods to really overcome these problems. Through persistently sharing information with other agents help agents achieve better coordination. Most of work on communication for MARL persistently broadcast own information such as current observation or agent's current intention(short-term intention) which is latent embeddings of policy function. Agent's current intention often contains more valid information than partial observation. (Sukhbaatar et al. (2016)) achieves superior performance by broadcasting own intention. However, the intention is large, and continual broadcasting its own intention could cause large communication overhead which makes it is difficult to be applied for real-world problems.

Recently, some work focuses on reducing communication overhead such as TMC (Zhang et al. (2020)). However, TMC just simply reducing the communication frequency could impair agent's cooperation. H-comm (Tang et al. (2018)) proposes communicating with high-level policy's intention. This intention is a long-term intention which keep constant for a while. Thus it achieves more sparse communication. But directly communicating with this intention could be lack of agent's current information and enough future information which could hinder agent's cooperation. Considering a real-world task playing football. In this sport, athletes don't directly communicate with other team member rather cooperating with each other through previous tacit understanding. There are two key factors in their cooperation. First of all, everyone in the team are with own roles and different roles undertake different subtasks which is analogue to long-term intention in MARL communication. If there is no clear division of labor, team members'coordination is harder coming true. Then their accumulated tacit understanding during daily training are equally important in cooperation. Tacit understanding is similar to inferred intention. According to these information, athletes can make correct choices. These two important factors help the team to continuously win the games. Therefore, we argue that clear role allocation and inferring their behavior based on historical experience are useful ways to finish teamwork.

In this paper, we propose learning hierarchical multi-agent ooperation with long-short term intention (HLSI), a succinct but efficient method which enables agent to achieve better coordination . Firstly, each agent maximizes mutual information between agent's macro action and agent's future trajectory condition on historical trajectory to make long-term intentions contain more useful long-term information. Each agent exchanges their long-term intention which is similar to the roles in playing football just once until the long-term intention changes and realizes by sharing high-level policy's latent embeddings of hierarchical multi-agent reinforcement learning. Then, after recieving these messages, unlike existing work h-comm that only shares long-term intention and averages them to represent other's intention, our agents integrate these long-term intention through the attention mechanism to greater leverage these messages. To compensate the missing current information. We learn a behavior inference network which can roughly infer other's short-term intentions. Agent infers other agent current intention (short intention) via historical trajectory. Behavior inference network is learned according to other agents' real actions during training stage. Then long-short term intention fushion module fuses real other's long-term intentions and inferrd other's short-term intentions to help agent more precisely understand others' next intentions.

HLSI communicates based on extremely sparse hierarchical communication pattern and it can achieve same performance compared with sharing low-level policy's latent embeddings, which needs to real-time communicate. We evaluate HLSI in three different multi-agnet cooperative environment:level-baed foraging (LBF) (Papoudakis et al. (2020)), and a modified version of cooperative navigation (CN) and predator-prey proposed by (Ding et al. (2020)). Our results indicate that fusing long-term and short-term intention is useful in team cooperation. Based on these information, agent can make better decisions under partially observable environment and achieve better performance than other baselines. We also provide complete evaluations and ablations studies to verify the effect of each module.

2 Related Work

Hierarchical Reinforcement Learning is a framework which decomposes long-horizon task into multiple simple subtasks. Most of existing work is based on two level policy. The high-level policy outputs an abstract action which is represented as a macro action or a subtask which would last for multiple timesteps. The low-level policy makes decisions according to high-level's abstract action. Recently hierarchical reinforcement learning is gradually applied to multi-agent reinforcement learning. FMH (Ahilan & Dayan (2019)) proposes a manager-work structure, manager set subgoals for works and workers complete own tasks to together obtain maximun team reward. HSD (Yang et al. (2019)) introduces a skill decoder as an extra intrinsic reward to make low-level policy learn useful and distinct skills. HAVEN (Xu et al. (2021)) introduces a dual coordination mechanism to address instabilities when concurrently optimizing high-level and low-level policies.

Multi-Agent Reinforcement Learning often suffers non-stationary and partially observable problems. Communication is one of effective methods to solve these problems. RIAL (Foerster et al. (2016)) and DIAL (Foerster et al. (2016)) are first proposed to communicate with others by discrete messages. To overcome the limitation of discrete messages channel. CommNet (Sukhbaatar et al. (2016)) averages the received hidden layer to achieve more effective collaboration. The hidden layers are the output of recurrent neural network which encodes the past and current information. Differing from above work which only considers current information, IS(Kim et al. (2020)) encodes imagined trajectory to capture agent's intention and combines the current information and future information based on attention mechanism. However, the aforementioned methods communicate through broadcast mechanism which could lead to information redundancy. Based on CommNet, IC3Net (Singh et al. (2018)) also shares hidden layer with remaining agents but it proposes to add a gate mechanism to make agents choose if communicating with other agents to reduce unnecessary information. ATOC (Jiang & Lu (2018)) communicates with agnet's intention which is latent embeddings of policy function and uses attention module to implement a more complex gate mechanism to decide whether communicating with adjacent agents. To really abandon broadcast mechanism which spread large redundant information. I2C (Ding et al. (2020)) trains a prior network according to the causal effect between agents and communicate with those who could bring large influence to it based on this prior network . And they design correlation regularization to help the agent correlates other agent's observation. TMC (Zhang et al. (2020)) designs a new communication protocol that avoids real-time communication via retransmission timeout (RTO) to achieve succinct and robust communication. H-comm (Tang et al. (2018)) proposes communicating with long-term intention which is latent embeddings of high-level policy which keeps constant until choosing new high-level action and averages the received messages. It can achieve more sparse communication but doesn't achieve effective cooperation.

Inspired by h-comm (Tang et al. (2018)), we propose a new hierarchical multi-agent reinforcement Learning cooperation framework that achieves sparse and sufficient coordination. H-comm only considers a simply average received message. We not only use extra mutual information reward to make the message contain more useful content, but also leverage the attention module to more effectively combine received messages which is similar to MAIC(Yuan et al. (2022)). It uses attention module to combine the representation of tailored incentive messages. In addition, we use inferred others' intentions to help agent understand others' next behavior.

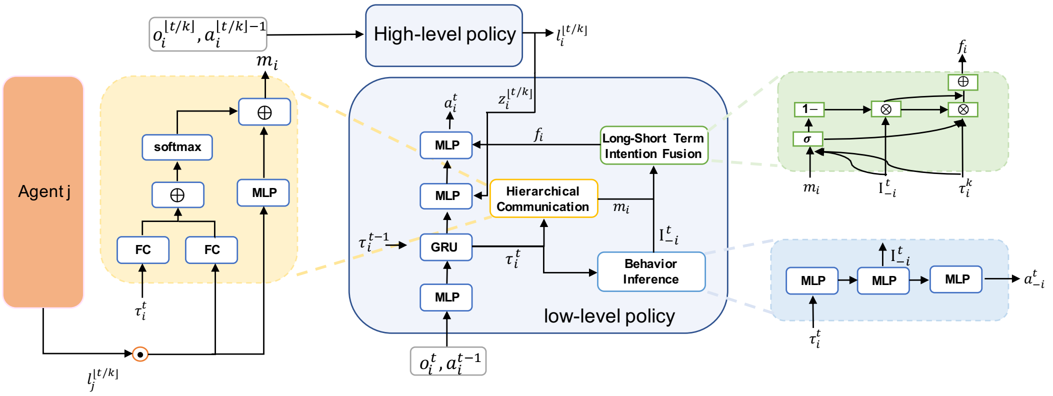


Figure 1: Overall structure of the HISI. (left) hierarchical attention communication module. (middle) network of agent i. High-level policy produces a long-term intention $l_i^{\lfloor t/k \rfloor}$ which also can be considered as its role. It is penultimate layer output of high-level policy. $z_i^{\lfloor t/k \rfloor}$ is the final output of high-level policy which is a one-hot variable. (right) long short-term intention fusion module and behavior inference module.

3 Method

In this section, we will describe the detailed design of our HLSI. HLSI can be instantiated by any hierarchical CTDE framework. As shown in Figure 1, our HLSI consists of a high-level policy, a low-level policy, an attention communication module, a behavior inference module, and a long short-term intention fusion module. We consider that agents work together to solve cooperative tasks under partially observable environment. At every k timesteps, high-level policy leverages observation $o_i^{\lfloor t/k \rfloor}$, last action $z_i^{\lfloor t/k \rfloor - 1}$ and hidden state $h_i^{\lfloor t/k \rfloor - 1}$ to get new hidden state $h_i^{\lfloor t/k \rfloor}$ encoded by GRU. Then agent obtains macro action $z_i^{\lfloor t/k \rfloor} \sim \pi_i^{high}(z_i^{\lfloor t/k \rfloor} | h^{\lfloor t/k \rfloor})$ and this action will keep k steps constant. Differing from high-level policy, low-level policy real-time interacts with the environment. At each timestep, agent i receives observation o_i^t. Then agent infers the other's next intention I_{-i}^t depending on the behavior inference module and communicates with observed agents if agent j doesn't buffer its' long-term intention which is the latent embeddings of high-level policy. Attention Communication network would combine received messages based on their importance under own historical information and output a variable m_i^t. Depending on the aforementioned messages, long short-term intention fusion module to fuse them to obtain f_i^t

to approximate others' real short-term intentions. Finally, low-level policy selects an action $a_i^t \sim \pi_i^{low}(a_i^t|\tau_i^t, f_i^t, z_i^{\lfloor t/k \rfloor})$ based on current observation o_i^t, macro action $z_i^{\lfloor t/k \rfloor}$, and fused intention f_i^t.

3.1 HIERARCHICAL ATTENTION COMMUNICATION(HAC)

Our hierarchical multi-agent reinforcement framework contains two levels. The agent has a high-level policy and a low-level policy. The high-level policy chooses action $z_i^{\lfloor t/k \rfloor}$ every k steps and lasts for k steps. This action is named macro action $z_i^{\lfloor t/k \rfloor}$. We define our macro action as one-hot variables. The low-level policy makes decisions based on historical trajectory information τ_i^t which is the history trajectory$(o_i^1, a_i^1, \dots, o_i^{t-1}, a_i^{t-1}, o_i^t)$ encoded by GRU and corresponding macro action $z_i^{\lfloor t/k \rfloor}$. Like the general reinforcement algorithm, low-level would continuously interact with the environment and get feedback from it. Similar to h-comm, we also share high-level policy latent embeddings (long-term intention) with others. To make latent embeddings contain more future information, we optimize the high-level policy by maximizing expected return and mutual information (MI) between the next trajectory τ_{t+k} and the random variable z conditioned on historical trajectory τ_t,

$$I(\tau_{t+k}; z|\tau_t) = H(\tau_{t+k}|\tau_t) - H(\tau_{t+k}|\tau_t, z). \tag{1}$$

Maximizing MI represents the uncertainty of τ_{t+k} conditioned on τ_t can be reduced after obtaining random variable z. Then maximizing Eq.1 can help z contain more own future trajectory information which can acquire a powerful intention representation. MI can be written in the following form

$$I(\tau_{t+k}; z|\tau_t) = \int p(z, \tau_t, \tau_{t+k}) log \frac{p(z|\tau_{t+k}, \tau_t)}{p(z|\tau_t)} dz d\tau_{t+k} d\tau_t \tag{2}$$

Inspired by (Sharma et al. (2019)), we use variational inference to solve this problem. Using $q_\phi(z|\tau_{t+k}, \tau_t)$ to approximate distriion $p(z|\tau_{t+k}, \tau_t)$. And we approximately assume the high-level policy $\pi^{high}(z|\tau_t)$ is equivalent to $p(z|\tau_t)$. Then we can obtain the lower bound of Eq.2 (a detailed version in Appendix A).

$$I(\tau_{t+k}; z|\tau_t) \geq E_{z,\tau_{t+k},\tau_t \sim p}[log \frac{q_\phi(z|\tau_{t+k}, \tau_t)}{\pi^{high}(z|h_{\lfloor t/k \rfloor})}] \tag{3}$$

To maximize the lower bound, we leverage it as an extra reward. Then the high-level policy's reward is $r = r_{env}\sigma + (1-\sigma) r_{mi}$ and $r_{mi} = log \frac{q_\phi(z|\tau_{t+k}, \tau_t)}{\pi^{high}(z|h_{\lfloor t/k \rfloor})}$.

Attention Communication. Others' intentions can help the agents make better decisions. However, there is some invalid information in these messages. To effectively utilize received messages from other agents, we apply the attention mechanism (Bahdanau et al. (2014)) to learn the importance of other agents' messages according to own historical trajectory τ_i. Then the received messages can be combined based on this weight to get useful information. The query q_i is agent historical information τ_i and key k_{ij} is the long-term intention from seen agents. The value v_{ij} is also the long-term intention l_{ji}, which is a feature vector of the output of the agent j's high-level policy's penultimate layer network, from others. In practice, we find that communicating with this feature vector is more effective than communicating with one-hot variables. Both key and query are a full connection layer and value is a full connection layer with a relu activation function. The weight is calculated by the dot product of key and value, then normalized to get the weight to [0,1] by softmax,

$$\alpha^i = softmax[\frac{q_i^T k_1}{\sqrt{d_k}}, \frac{q_i^T k_2}{\sqrt{d_k}}, \dots, \frac{q_i^T k_W}{\sqrt{d_k}}]. \tag{4}$$

This weight means which agent's messages could be helpful to the current agent The final output is

$$m_i = \sum_{n=1}^{W} \alpha_n^i v_n. \tag{5}$$

The network of key, query, and value are updated via gradient from minimax temporal difference error of IQL. W is the number of other agents.

3.2 HIERARCHICAL LONG SHORT-TERM INTENTION FUSION (HLSI)

Behavior Inference Network. Communication with hierarchical reinforcement learning is sparse, however, it is lack of some agent's current information which could impair coordination. Even though the initial state is the same, the following process still has different possibilities. It is important to obtain others' current intentions. However, it is difficult to predict other's next action under an unconstrained partially observable environment. Thus we merely hope to learn a feature to represent others' current intentions to some extent, but accurately predict other action. In some computer vision areas such as object detection and object segmentation, to obtain a backbone work that can effectively extract image features, researchers often train it in the imageNet dataset. Thus to get such a feature, we design a behavior inference network that enables the agent to infer others' current intentions according to historical information τ_t^i. τ_t^i mean the encoded trajectory o_t^i, a_t^i of agent i. The τ_t is the output of anget's GRU layer. It is represented as $f_i(\tau_t^i) = a_t^{-1}$ which hopes to predict others actions based historical information. After training this network, the embeddings I_{-i}^t of this network will contain enough information about others' current intentions which can be represented as others' short intentions.

Long Short-Term Intention Fusion. Sharing high-level intention help agent understands others' long-term intentions. The behavior inference network infers the rough current intention. Only one of these two elements is not enough to understand the others' next behavior. The long-term intention lacks enough current information and the short-term intention is not accurate. Thus we hope to combine the accurate long-term intention and coarse short-term intention to approximate the real others' next intentions to choose a better action. To fuse long-term and short-term intetntions of other agents, we design a Long short-term intention fusion module which is inspired by LSTM (Graves (2012)) and GRU (Chung et al. (2014)). This module only includes a fushion gate

$$p_i = sigmoid(W^1 m_i + W^2 \tau_i + W^3 I_{-i}) \tag{6}$$

which balances the importance between long-term and short-term intention in the current time and the final output gate is computed by

$$f_i = (1 - p_i) \cdot m_i + p_i \cdot I_{-i}. \tag{7}$$

which output a feature to approximate others' real short-term intentions.

3.3 Training

To implement our HLSI, we use MADDPG (Papoudakis et al. (2020)) which contains a critic and an actor as high-level policy. It needs to train centralized critic Q-function $Q^{\pi_h}(o, z)$ which uses global observation and macro action as input. It can help agent optimize individual policy actor $\pi(z_i|h_i)$. They respectively parameter by $\theta_{Q_h}^i$ and $\theta_{\pi_h}^i$ Thus the high-level critic and actor are updated as

$$L(\theta_{Q_h}^i) = E_{o,a,r,o'}[(Q^{\pi_h}(o, z) - y)^2] \tag{8}$$

$$y = r + \gamma Q_h^\pi(o', z')|z' \sim \pi(z_i|h_i) \tag{9}$$

$$\nabla_{\theta_{\pi_h}^i} J(\theta_{\pi_h}^i) = E_{o,z}[\nabla_{\theta_{\pi_h}^i} \pi(z_i|h_i) Q^{\pi_h}(o, z)] \tag{10}$$

Our low-level policy is implemented based on IQL. Besides a critic network, our HLSI also contains an attention communication network, and a short-term and long-term intention fusion module. For agent i, they respectively parameter by a set of trainable parameter $\theta_{low} = \left(\theta_{Q_l}^i, \theta^i = \left(\theta_Q^i, \theta_K^i, \theta_V^i\right), \theta_l^i\right)$. We can train critic network as

$$L(\theta_{Q_{low}}^i) = E_{o,a,r,o'}[(Q^{\pi_l}(\tau_i, a_i, z_i) - y)^2] \tag{11}$$

$$y = r + \gamma Q_l^\pi(\tau_i', a_i', z_i)|a' = argmax_a Q_l^\pi(a_i, \tau_i', z_i) \tag{12}$$

Based on chain roles, gradient can flow from Q-fuction to update each module. The behavior infer network parameterized by θ_{b_i} which can be learned by mean-squre loss of real others' actions. Then the loss is

$$L(\theta_{b_i}) = E[(f_{\theta_{b_i}}^i(\tau^i) - a^{-i})^2]. \tag{13}$$

4 Experiments

We evaluate our HLSI in three multi-agent cooperative tasks: Cooperative Navigation, Predator Prey, and Level-Based Foraging. For this three different tasks, we implement HLSI based MADDPG (Papoudakis et al. (2020)) as high-level policy and IQL (Tan (1993);Mnih et al. (2015)) as low-level policy. Our HLSI is a hierarchical communication multi-agent reinforcement algorithm built on MADDPG and IQL. Thus we respectively choose MADDPG, COMA(Foerster et al. (2018)) and

IQL which are baselines of multi-agent reinforcement algorithm, IS(Kim et al. (2020)) which is an intention sharing state-of-the-art baseline of multi-agent communication, I2C (Ding et al. (2020)) which is a state-of-the-art baseline of multi-agent efficient communication, and HSD (Yang et al. (2019)) which is a state-of-the-art baseline of hierarchical multi-agent reinforcement learning. To make sure the fairness of our results, we evaluate all the algorithms with the same basic hyperparameters.

4.1 Multi-Agent Environments

Cooperative Navigation(CN): The agents in this environment are aimed to occupy all the landmarks. Agents can move with a certain velocity and landmarks keep still. In our environment, the agent only observes nearby three agents and landmarks and is rewarded by the Euclidean distance of the nearest landmark. However, if the agent collide with others, it will receive a negative reward as a penalty. The team reward is the sum of the Euclidean distance between the agent and the nearest landmark and all the collision penalties. In this setting, the agent only communicates with nearby agents and infers their next behavior. At the start of each episode, agents and landmarks are initialized in a random position. Then agents should coordinate to move to cover as many as possible landmarks and avoid collisions to gain the greatest team rewards. In this experiment, the number of agents and landmarks is seven. The penalty reward is $r_{penalty}$=-1. The length of each episode is 30 timesteps. The agent's action space is [left, down, right, up, stop]. The agent's observation is own velocity, position and the position of nearby three agents and landmarks.

Predator Prey(PP): We modify the Predator-prey introduced in (Ding et al. (2020)). In this environment, there are N predators and M preys. Both the predators and preys can move at every timesteps and preys are far from the nearest predator. Preys move faster than predators. Predators observe the closest three predators and preys. And they are allowed to share intention with these predators. Predators are rewarded by the distance of the closest prey and are penalty by colliding with other predators. The collision penalty reward is $r_{penalty}$=-1. Thus our goal is to control this agent to cooperate to capture as more as possible preys to get a higher team reward. Due to the prey velocity is faster than predator, it is difficult to capture these preys when they don't mutually cooperate. We train our algorithm in the setting of N=7 and M=5 and the length of the episode is 30. The agent's action space is [left, down, right, up, stop]. The agent's observation is own velocity, position and the position of nearby three agents and preys.

Level-Based Foraging (LBF): The task of Level-based foraging is introduced in (Papoudakis et al. (2020)). In our experimental setting, agents are located in a 15×15 grid world. They need to cooperate to collect as many as possible foods and they can load the food only when the summation of food's adjacent agents' level is greater than or equal to the level of the food. If they collect a food, agents will be rewarded correlated to the level of the food. The goal of this task is to collect as many as possible foods to gain maximum team. This environment is also partially observable and the agent perceives the agents and foods of three grid cells. The number of agents is four and the length of the episode is 30. We assume that agent can communicate with all agents, otherwise the communication in this problem is too sparse. The agent's action space is [left, down, right, up, stop, load]. The agent's observation is own position, level and the position of foods and other agents within vision

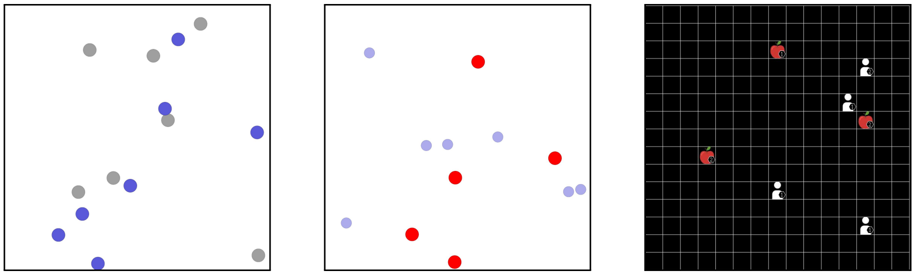


Figure 2: The three multi-agent environment: (left) Cooperative Navigation (CN), (middle) Predator Prey (PP) , (right) Level-Based Foraging (LBF).

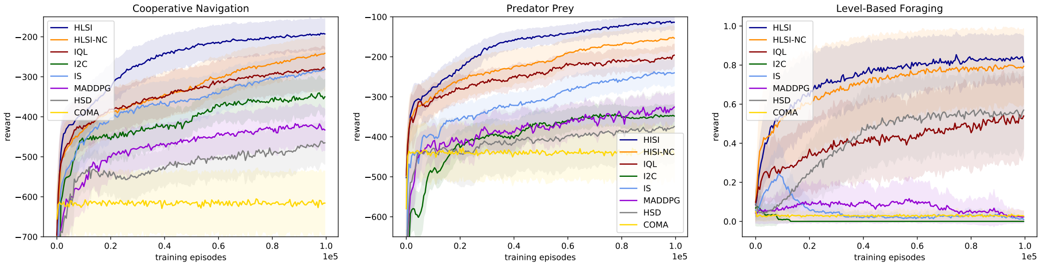


Figure 3: episodic reward of five evaluated methods in three multi-agent cooperative settings

4.2 QUANTITATIVE RESULTS

Figure 3 shows the learning curve of our HLSI and other baselines during training in three different settings. The rewards of every method are average performance over five random seeds and the shadow is the standard deviation. Appendix D shows more detailed indexes of three environments.

First of all, we observe that HLSI converges to the highest reward against other baselines in all three environments and HLSI-NC, which is HLSI without communication and behavior inference module, is closest to HLSI's final rewards. Compared with HLSI-NC, HLSI effectively leverages real long-term intention and inferred short-term intention to fuse to approximate others' real current intentions. This indicates real others' short-term intentions can help agent make better decisions. And the performance of HLSI-NC slightly surpasses IQL since the hierarchical structure played a certain role. HSD perform worst in these two environment. The possible reason is that HSD can't decompose the team reward. Differing from MADDPG, I2C communicates by sharing agent's observation and IS communicates by sharing own intention. They help agents achieve better coordination in cn environment through communicating. Further, sharing intention can achieve better performance than current observation, since it contain more valid information. Due to the terrible performance of the basic framework MADDPG, it doesn't behave very well and agent's partial observation includes less information compared with the policy's latent embeddings. In the LBF environment, the discrepancy between MADDPG and IQL is bigger. MADDPG hardly learns a valid policy. The possible reason is that training a larger network over the joint observation-action space, as required for these algorithms, demands sufficient training signals. However, this environment's reward is sparse. And the discrepancy between HISI-NC and IQL is also large. The performance improvement of our HLSI is weak and the performance of i2c even falls can be explained by the fact that the agent doesn't always observe agents for communication which make communication and behavior inference less effective. In this environment, hierarchical reinforcement learning can decompose the comparatively complex problems into some subproblems which make them perform better than other baselines. COMA hardly learns a valid policy as all results show.

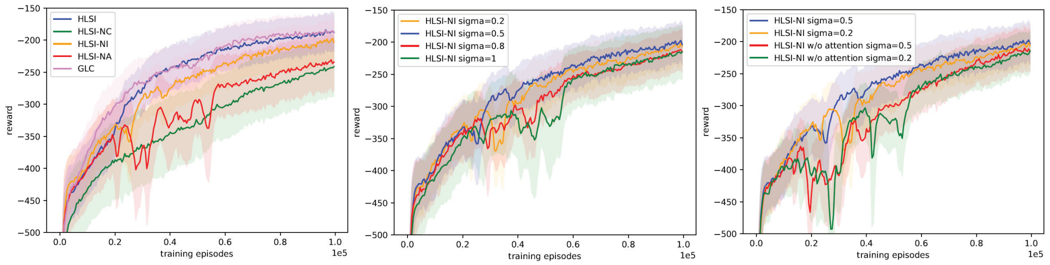


Figure 4: (left) performance for HLSI compared with different module ablation baselines. (middle)performance for HLSI-NI comared with different σ during training. (right) performace for HLSI-NI with and without attention module.

4.3 Ablation Study

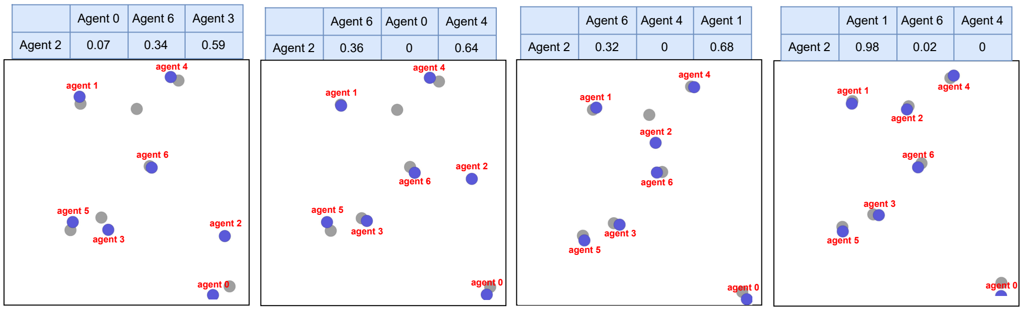
In this section, we perform some ablations studies to justify the performance of our algorithm. We design and evaluate three different ablations in cooperative navigation setting. Our HLSI composes of hierarchical communication module, behavior inference module, and long short-term intention fusion module. Firstly, we investigate the necessity of these three models. We evaluate the effect of real long-term intention and inferred short-term intention. Thus we compared our HLSI with the agent without inferring others' behavior model, denoted as HLSI-NI , the agent without hierarchical attention communication model, denoted as HLSI-NA, the agent communicates with general low-level policy's latent embeddings, denoted as GLC and the agent without communicating with others, denoted as HLSI-NC. Then, we verify the importance of our high-level's extra mutual information maximum reward and compare four different combined ratio σ to evaluate how it affect the performance of agent's cooperation. Finally, we compare that agent leverages the received long-term intention with and without attention module.

The left in figure 4 exhibits the reward of HLSI and the related ablations baselines in Cooperative Navigation setting. We observe that agent with hierarchical communication module performs better than agent without hierarchical communication module. The reason is that this environment is partially observable and agent could be influenced by the others behavior. Communicating with long-term intention can make agents understand others' next intention which lets them make more correct decisions. We also observe that either without hierarchical communication module (HLSI-NC) or without behavior module (HLSI-NI) results in lower rewards. This demonstrates that long-term and short-term intentions are necessary in multi-agent cooperative task. And the final reward of HLSI and GLC is equal, which indicates that fusing long-term and short-term intention can approximate the real agent current intention.

Second, it can be found that extra mutual information reward help agent capture more future information which make them coordinate better. As the middle in figure 4 shows, when σ is 0.5, it behaves better than σ is 0.8 and 0.2. The possible reason is that too larger σ could impair high-level policy to obtain optimal strategy and too little σ provide less future intention. In practice, we need to determine the optimal value of σ by the trial-and-error method according to specific task.

Finally, we investigate the effect of attention communication module. We replace the learnable weight α by averaging the received message to vertify the importance of weight learnt by attention module. We compare two different σ with or without attention module in the right figure 4. We can see that the method with attention module can more effectively combine the received messages to achieve better coordination .

4.4 Model Evaluation



	Agent 0	Agent 6	Agent 3
Agent 2	0.07	0.34	0.59

	Agent 6	Agent 0	Agent 4
Agent 2	0.36	0	0.64

	Agent 6	Agent 4	Agent 1
Agent 2	0.32	0	0.68

	Agent 1	Agent 6	Agent 4
Agent 2	0.98	0.02	0

Figure 5: Visualizing the relationship between attention weight and agent's long-term behavior.

We display some frames of an episode and the importance weight at that moment. The key is the other agents' long-term intentions and the query is the agent's own historical trajectory. Figure 5 exhibits partial frames of a test episode. In the first two frames, agent two is moving up. In the long term, agent four or six could bring greater impact than agent zero, thus their importance of weight is higher corresponding. In the next two frames, agent two moves to the left and then finds an unoccupied landmark. It doesn't move toward agent four, so its weight is lower compared with the other two agents. We speculate that attention communication module can capture possible future influence based on others‘ long-term intentions.

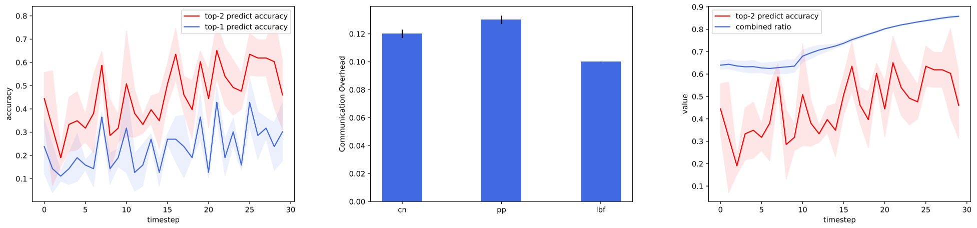


Figure 6: (left) Top-1 and top-2 actions predict accuracy of all agents as the episode progresses time. (middle) The communication overhead ratio between our pattern and normal patternt in three different environment. (right) The relationship between combination ratio and predict accuracy.

Figure 6(a) shows the top-1 and top-2 average actions predict the accuracy of all agents in the cooperative navigation environment through time. At each timestep t, we use the embedding z_t^i, which is the historical trajectory of agent i, to predict the adjacent agent's next actions. We observe that top-1 and top-2 accuracy is higher and higher and top-2 prediction accuracy is closer to 0.6 after time steps. This shows that at the later stage, the behavior inference network can output a good representation of others' current intentions.

Based on hierarchical communication pattern, each pair of agents communicates only once until the high-level policy change. We compare our pattern's number of communications with normal partten. Figure 6(b) illustrates the communication overhead ratio of all the overheads of our pattern and normal pattern in the total process. We find that our communication pattern's overhead is far less than normal communication pattern. Thus hierarchical communication pattern can be used to high costs real-world problems.

Finally, we explore the relationship between combined ratio and predict accuracy to investigate how long short-term intention fusion module Leverages the short-term and long-term intentions. In figure 6(c), we observe that at the beginning of the episode, the agent can't correctly infer other's short-term intention, so it combines more long-term intentions. After certainty interacting with the environment, the agent can more accurately infer other's short-term intentions, thus it will use more inferred short-term intention. Through appropriately combining real long-term and inferred short-term intention, we can approximate other's real short-term intentions which helps agent can better understand others' next behavior.

5 Conclusion

In this paper, we propose HLSI, a hierarchical multi-agent cooperation framework to improve agents' coordination with long short-term intention. First of all, we maximize mutual information to make latent embeddings contain more future information. Next, agents integrate long-term intentions come from other's high policy's long-term intention through the attention mechanism. Then to compensate the lack of current information, behavior inference network infer other's intentions according to own historical trajectory information. Finally, long short-term intention fushion module fushs long-term and short-term intention to approximate others' real short-term intentions, which helps agent better understand other's next intentions. Empirically, it is demonstrated that HLSI outperforms existing multi-agent communication and hierarchical multi-agent reinforcement baselines in a variety of cooperative multi-agent scenarios. The main limitation of our HLSI is that behavior inference module and communication module are less effective when the vision of agents is narrow. Due to the limitation of vision, agent couldn't observe others for a long time such as LBF environment. Thus, it is difficult to infer others' next behavior based on own historic trajectory. And communicating with agents, which is out of sight, doesn't help much since it doesn't directly interact with controlled agent. To overcome these, we think that there is some prior knowledge supplied to agent. To alleviate the difficult of behavior prediction under partially observable environment, we can set some limits on agent policy selection(Papoudakis et al. (2021)). In the future, we would like to extend our method to competitive and mixed environment and apply it to large-scale multi-agent tasks.

References

Sanjeevan Ahilan and Peter Dayan. Feudal multi-agent hierarchies for cooperative reinforcement learning. *arXiv preprint arXiv:1901.08492*, 2019.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

Ziluo Ding, Tiejun Huang, and Zongqing Lu. Learning individually inferred communication for multi-agent cooperation. *Advances in Neural Information Processing Systems*, 33:22069–22079, 2020.

Jakob Foerster, Ioannis Alexandros Assael, Nando De Freitas, and Shimon Whiteson. Learning to communicate with deep multi-agent reinforcement learning. *Advances in neural information processing systems*, 29, 2016.

Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Alex Graves. Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, pp. 37–45, 2012.

Jiechuan Jiang and Zongqing Lu. Learning attentional communication for multi-agent cooperation. *Advances in neural information processing systems*, 31, 2018.

Woojun Kim, Jongeui Park, and Youngchul Sung. Communication in multi-agent reinforcement learning: Intention sharing. In *International Conference on Learning Representations*, 2020.

Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems*, 30, 2017.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

Y. Yang Q. Yuan Z. Tang H. Long P. Peng, Y. Wen and J. Wang. Multiagent bidirectionally- coordinated nets: Emergence of human-level coordination in learning to play starcraft combat games. *arXiv preprint arXiv*, 1703.10069, 2017.

Georgios Papoudakis, Filippos Christianos, Lukas Schäfer, and Stefano V Albrecht. Benchmarking multi-agent deep reinforcement learning algorithms in cooperative tasks. *arXiv preprint arXiv:2006.07869*, 2020.

Georgios Papoudakis, Filippos Christianos, and Stefano Albrecht. Agent modelling under partial observability for deep reinforcement learning. *Advances in Neural Information Processing Systems*, 34:19210–19222, 2021.

Shashua A Shalev-Shwartz S, Shammah S. Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv preprint arXiv*, 1610.03295, 2016.

Archit Sharma, Shixiang Gu, Sergey Levine, Vikash Kumar, and Karol Hausman. Dynamics-aware unsupervised discovery of skills. *arXiv preprint arXiv:1907.01657*, 2019.

Amanpreet Singh, Tushar Jain, and Sainbayar Sukhbaatar. Learning when to communicate at scale in multiagent cooperative and competitive tasks. *arXiv preprint arXiv:1812.09755*, 2018.

Sainbayar Sukhbaatar, Rob Fergus, et al. Learning multiagent communication with backpropagation. *Advances in neural information processing systems*, 29, 2016.

Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z Leibo, Karl Tuyls, et al. Value-decomposition networks for cooperative multi-agent learning. *arXiv preprint arXiv:1706.05296*, 2017.

Ming Tan. Multi-agent reinforcement learning: Independent vs. cooperative agents. In *Proceedings of the tenth international conference on machine learning*, pp. 330–337, 1993.

Hongyao Tang, Jianye Hao, Tangjie Lv, Yingfeng Chen, Zongzhang Zhang, Hangtian Jia, Chunxu Ren, Yan Zheng, Zhaopeng Meng, Changjie Fan, et al. Hierarchical deep multiagent reinforcement learning with temporal abstraction. *arXiv preprint arXiv:1809.09332*, 2018.

M. A Wiering. Multi-agent reinforcement learning for traffic light control. *In Machine Learning: Proceedings of the Seventeenth International Conference(ICML*, 1151-1158, 2000.

Zhiwei Xu, Yunpeng Bai, Bin Zhang, Dapeng Li, and Guoliang Fan. Haven: Hierarchical cooperative multi-agent reinforcement learning with dual coordination mechanism. *arXiv preprint arXiv:2110.07246*, 2021.

Jiachen Yang, Igor Borovikov, and Hongyuan Zha. Hierarchical cooperative multi-agent reinforcement learning with skill discovery. *arXiv preprint arXiv:1912.03558*, 2019.

Lei Yuan, Jianhao Wang, Fuxiang Zhang, Chenghe Wang, Zongzhang Zhang, Yang Yu, and Chongjie Zhang. Multi-agent incentive communication via decentralized teammate modeling. 2022.

Sai Qian Zhang, Qi Zhang, and Jieyu Lin. Succinct and robust multi-agent communication with temporal message control. *Advances in Neural Information Processing Systems*, 33:17271–17282, 2020.

A Lower Bound on Mutual Information

To enable the long term intention contain more future information, we propose to maximize the mutual information between z_t and τ_{t+k} condition on τ_t. We borrow ideas from (Sharma et al. (2019)) to obtain a lower bound of this mutual information.

Theorm 1. *A lower bound of $I(\tau_{t+k}; z|\tau_t)$ is*

$$E_{z,\tau_{t+k},\tau_t \sim p}[log \frac{q_\phi(z|\tau_{t+k}, \tau_t)}{\pi^h(z|\tau_t)}], \tag{14}$$

where τ_t is historical trajactory, τ_{t+k} is agent's trajectory from t to t+k, z is agent macro action and $q_\phi(z|\tau_{t+k}, \tau_t)$ is variational distribution with parameter ϕ.

$Proof.$

$$I(\tau_{t+k}; z|\tau_t) = \int p(z, \tau_t, \tau_{t+k}) log \frac{p(z|\tau_{t+k}, \tau_t)}{p(z|\tau_t)} dz d\tau_{t+k} d\tau_t \tag{15}$$

We introducce $q_\phi(z|\tau_{t+k}, \tau_t)$ to approximate to variational distribution to approximate the conditional distribution $p(z, \tau_t, \tau_{t+k})$. Then we obtain the lower variationally bound object as follows:

$$\begin{aligned} I(\tau_{t+k}; z|\tau_t) &= \int p(z, \tau_t, \tau_{t+k}) log \frac{p(z|\tau_{t+k}, \tau_t)}{p(z|\tau_t)} dz d\tau_{t+k} d\tau_t \\ &= E_{z,\tau_{t+k},\tau_t \sim p}[log \frac{q_\phi(z|\tau_{t+k}, \tau_t)}{p(z|\tau_t)}] \\ &+ E_{\tau_{t+k},\tau_t \sim p}[D_{KL}(q_\phi(z|\tau_{t+k}, \tau_t)||p(z, \tau_t, \tau_{t+k}))] \\ &\geq E_{z,\tau_{t+k},\tau_t \sim p}[log \frac{q_\phi(z|\tau_{t+k}, \tau_t)}{\pi^{high}(z|h_{\lfloor t/k \rfloor}).}]. \end{aligned} \tag{16}$$

Since $E_{\tau_{t+k},\tau_t \sim p}[D_{KL}(q_\phi(z|\tau_{t+k}, \tau_t)||p(z|\tau_t, \tau_{t+k}))] \geq 0$. Then we approximately assmue that $p(z|\tau_t) \approx \pi^{high}(z|h_{\lfloor t/k \rfloor})$. Thus we get the lower bound in Theorm 1.

When this divergence is small enough. We can use this bound replace MI. When the expectation of Kullback-Leibler (KL) divergence between $q_\phi(z|\tau_{t+k}, \tau_t)$ and $p(z|\tau_{t+k}, \tau_t)$ is zero, the MI is equal to its lower bound. It is easy to use maximizing the likelihood to optimize $q_\phi(z|\tau_{t+k}, \tau_t)$ to approximate $p(z|\tau_{t+k})$.

B Implementation Details

Table 1 and 2 summarize the hyperparameters used by HLSI and the baselines in the three environments.

Table 1: hyperparameters for cooperative navigation and predator prey

Hyperparameter	HLSI	MADDPG	I2C	IQL	HSD	IS	COMA
discount(γ)	0.99	0.99	0.99	0.99	0.99	0.99	0.99
repay buffer size	$5{\times}10^5$	$5{\times}10^5$	$5{\times}10^5$	$5{\times}10^5$	$5{\times}10^5$	$5{\times}10^5$	$5{\times}10^5$
batch size	128	128	128	128	128	128	128
optimizer	ADAM	ADAM	ADAM	ADAM	ADAM	ADAM	ADAM
learning rate of low-level policy	$5{\times}10^{-4}$	$5{\times}10^{-4}$	$1{\times}10^{-2}$	$5{\times}10^{-4}$	$1{\times}10^{-4}$	$5{\times}10^{-4}$	$1{\times}10^{-4}$
learning rate of high-level policy	$5{\times}10^{-4}$	-	-	-	$1{\times}10^{-4}$	-	-

Table 2: hyperparameters for Level-based foraging

Hyperparameter	HLSI	MADDPG	I2C	IQL	HSD	IS	COMA
discount(γ)	0.99	0.99	0.99	0.99	0.99	0.99	0.99
repay buffer size	$5{\times}10^5$	$5{\times}10^5$	$5{\times}10^5$	$5{\times}10^5$	$5{\times}10^5$	$5{\times}10^5$	$5{\times}10^5$
batch size	128	128	128	128	128	128	128
optimizer	ADAM	ADAM	ADAM	ADAM	ADAM	ADAM	ADAM
learning rate of low-level policy	$3{\times}10^{-4}$	$3{\times}10^{-4}$	$1{\times}10^{-2}$	$3{\times}10^{-4}$	$1{\times}10^{-4}$	$3{\times}10^{-4}$	$1{\times}10^{-4}$
learning rate of high-level policy	$3{\times}10^{-4}$	-	-	-	$1{\times}10^{-4}$	-	-

C Pseudo Code

Algorithm 1 learning hierarchical multi-agent cooperation with long short-term intention

```
Initialize parameters $\theta^i_{Q_h}$, $\theta^i_{Q_l}$,$\theta^i_{\pi_h}$, $\theta_{b_i}$ for each agent i.
for episode = 1, . . . , $N$ do
    initialize observation $o_1$
    for t = 0,. . . ,T do
        for each agent i do
            if t mod k then
                choose a macro action $z_i^{\lfloor t/k \rfloor} \sim \pi_i^{high}(z_i^{\lfloor t/k \rfloor} | h^{\lfloor t/k \rfloor})$
            share long-term intention $h^{\lfloor t/k \rfloor}$ if adjacent agent j hasn't save it
            integrate received long term intention $l_{ji}$ with attention module to obtain $m_i^t$
            infer other's short intention $I_{-i}^t$ with behavior inference network
            long short-term fusion module fuse other's short-term intentions $I_{-i}^t$ and long-term intentions $m_i^t$ to obtain $f_i^t$
            choose action $a_i^t \sim \pi_i^{low}(a_i^t | \tau_i^t, f_i^t, z_i^{\lfloor t/k \rfloor})$
        EndFor
        excute action $a_t$ to obtain feedback $o_{t+1}$,$r_t$
        store tansition in Buffer
        if t mod $update\_frequency$ then
            for each agent i do
                Update $\theta^i_{Q_h}$, $\theta^i_{Q_l}$ by minimizing the loss (8) and loss(11)
                Update $\theta^i_{\pi_h}$ through gradient (10)
                Update $\theta_{b_i}$ by minimizing the loss (13)
            EndFor
        EndIf
    EndFor
EndFor
```

D Additional Detailed Indexes

We test HLSI and other baselines in three different environment. The results of more detailed indexes are from 100 test episodes among 5 random seeds. The ”collision” means that the number of collisions in the whole episode. The ”occupid” and ”captued” mean that the mean occupid landmarks and captured preys from 20th time step to the end of this episode.

Table 3: detailed indexes for CN

	HLSI	HLSI-NC	MADDPG	I2C	IQL	HSD	IS	COMA
reward	-184.52±58.44	-234.62±49.21	-435.99±60.03	-354.08±79.86	-281.99±58.09	-422.80±75.87	-304.39±79.60	-613.13±142.94
collisions	0.10±0.51	0.02±0.19	3.56±3.55	0.44±1.61	0.52±1.40	5.5±5.52	0.28±1.49	0.08±0.55
occupied	6.48±0.89	2.71±1.17	1.39±0.76	1.38±0.69	3.07±0.98	1.41±0.88	2.71±1.17	0.44±0.65

Table 4: detailed indexes for PP

	HLSI	HLSI-NC	MADDPG	I2C	IQL	HSD	IS	COMA
reward	-108.68±23.73	-143.19±53.88	-324.83±57.39	-354.92±52.08	-169.94±51.49	-439.42±56.56	-232.96±50.03	-433.93±121.83
collisions	0.06±0.31	0.13±0.36	1.95±2.02	0.56±0.97	0.20±0.52	1.40±1.42	0.16±0.62	0.00±0.00
captured	4.56±0.50	3.94±0.80	0.53±0.47	0.36±0.33	2.77±0.70	0.16±0.32	0.71±0.47	0.19±0.44

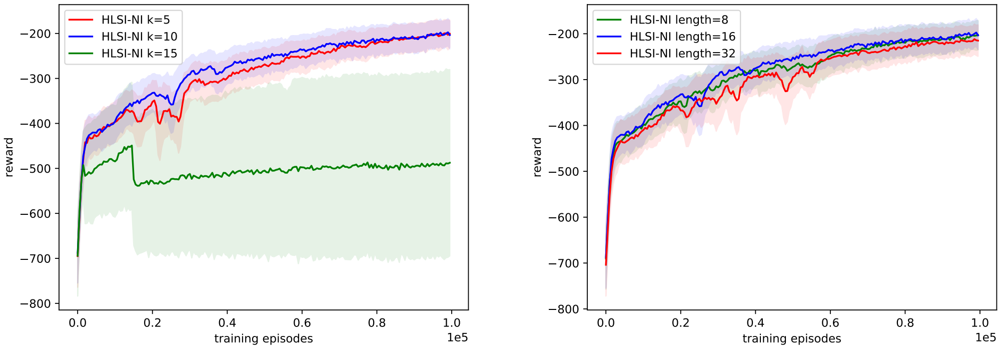
Table 5: detailed indexes for LBF

	HLSI	HLSI-NC	MADDPG	I2C	IQL	HSD	IS	COMA
reward	0.85±0.25	0.83±0.25	0.06±0.11	0.00±0.00	0.51±0.39	0.57±0.34	0.02±0.06	0.02±0.08

E Additional Ablation Study

We conduct two additional ablation studies to compare the effect of different communication interval k and the length of one-hot variable in cooperative navigation environment. Firstly, we compare three different k to evaluate how it affect the performance of agent's cooperation. Then, we compare three different length of one-hot variables.

Figure 6(a) shows that choosing a relatively long communication interval don't impair model's performance, however, HLSI-NI becomes very instable when communication interval becomes larger. Thus we can try to choose relatively long communication interval in different task. Figure 6(b) illustrates the different length of one-hot variables don't affect the performance of communicating with high-level policy's output of the penultimate layer network. Thus the k is ten and the length of one-hot variables is sixteen in our experiments.



(a) (b)

Figure 7: (a) performance for HLSI-NI comared with different communication interval k during training. (b) performance for HLSI-NI comared with different length of one-hot variables during training.