

CoDe: Communication Delay-Tolerant Multi-Agent Collaboration via Dual Alignment of Intent and Timeliness

Shoucheng Song^{1,2}, Youfang Lin^{1,2}, Sheng Han^{1,2}, Chang Yao^{1,2}, Hao Wu^{1,2},
Shuo Wang^{1,2}, Kai Lv^{1,2*}

¹School of Computer Science & Technology, Beijing Jiaotong University, Beijing, China

²Beijing Key Laboratory of Traffic Data Mining and Embodied Intelligence, Beijing, China
{insis_songsc, yflin, shhan, 22120453, 21241019, 19112029, lvkai}@bjtu.edu.cn

Abstract

Communication has been widely employed to enhance multi-agent collaboration. Previous research has typically assumed delay-free communication, a strong assumption that is challenging to meet in practice. However, real-world agents suffer from channel delays, receiving messages sent at different time points, termed *Asynchronous Communication*, leading to cognitive biases and breakdowns in collaboration. This paper first defines two communication delay settings in MARL and emphasizes their harm to collaboration. To handle the above delays, this paper proposes a novel framework **Communication Delay-tolerant Multi-Agent Collaboration (CoDe)**. At first, CoDe learns an intent representation as messages through future action inference, reflecting the stable future behavioral trends of the agents. Then, CoDe devises a dual alignment mechanism of intent and timeliness to strengthen the fusion process of asynchronous messages. In this way, agents can extract the long-term intent of others, even from delayed messages, and selectively utilize the most recent messages that are relevant to their intent. Experimental results demonstrate that CoDe outperforms baseline algorithms in three MARL benchmarks without delay and exhibits robustness under fixed and time-varying delays.

Introduction

Communication is crucial for enhancing the collaborative capability of multi-agents, especially within distributed systems. Through internal communication, agents restricted by local observations can strengthen their understanding of teammates and the environment, leading to more cooperative behaviors (Foerster et al. 2016; Das et al. 2019; Ding, Huang, and Lu 2020; Yuan et al. 2022; Sun et al. 2024). However, previous works frequently rely on ideal communication conditions, such as unlimited bandwidth, noise-free channels, and delay-free transmissions. These strong assumptions are typically hard to achieve in the real world.

Prior efforts have primarily focused on communication architectures to achieve sparse communication under bandwidth constraints, which involves gating mechanisms (Jiang and Lu 2018; Han, Dastani, and Wang 2023), event-triggered mechanisms (Hu et al. 2020), and predefined rules

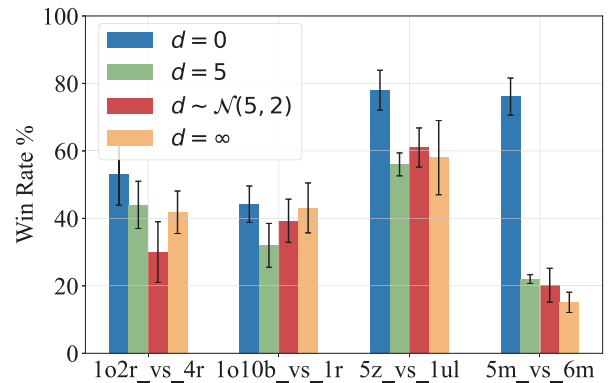


Figure 1: Performance of Communication-Enabled MARL Algorithm in Delayed Environments. Channel delays are configured as four types: 0, 5, infinity, and values sampled from a Gaussian distribution $\mathcal{N} \sim (5, 2)$. Notably, the unit of delay is the decision time interval.

(Zhang, Zhang, and Lin 2019, 2020; Sheng et al. 2022; Yuan et al. 2022). Moreover, several studies (Mitchell, Blumenkamp, and Prorok 2020; Xue et al. 2021; Yuan et al. 2024; Ding et al. 2024; Yu et al. 2024) also consider the impact of channel noise and malicious attacks and make further enhancements.

Nevertheless, only a few researchers have concentrated on the delay issue in MARL. DAMARL (Chen et al. 2020) addresses interaction delays between agents and the environment, assuming agents receive observations after a fixed time interval, without involving inter-agent communication. DACOM (Yuan et al. 2023) centers on communication delays among agents, integrating waiting time into the decision process. It assumes the delay is shorter than a single decision interval, thus ignoring message asynchronicity issues. Actually, channel delays often extend across decision intervals in more general scenarios, meaning that agents receive and utilize delayed messages from varying time points.

We define the above issue of delays across decision intervals as *Asynchronous Communication*. With delays, message transmission among agents lags, forcing them to rely on outdated messages to understand the environment and team-

*Corresponding Author: Kai Lv (lvkai@bjtu.edu.cn)

mates, which may have already changed. To verify this concern, we test conventional communication-enable MARL algorithms in SMAC (Samvelyan et al. 2019), setting 4 end-to-end channel delays. As depicted in Fig. 1, delays across decision intervals can impair the collaborative capability of original algorithms. Asynchronous messages can cause significant misguidance, potentially resulting in performance worse than without communication. This insight suggests that it is critical to research the robustness against communication delays across decision intervals. Here, we first summarize two types of communication delays in MARL:

- **Fixed-Delay Setting:** Point-to-point channels encounter a fixed delay that exceeds the decision interval. Messages received are outdated but sent simultaneously.
- **Time-Varying-Delay Setting:** Point-to-point delay fluctuates over time and surpasses the decision interval. Messages received are sent at different times.

To achieve robustness against the above delays, we propose a novel communication MARL framework named CoDe. Firstly, in terms of message expressiveness, we conceptualize intents as agents’ future behavioral trends. Within a seq-to-seq structure, we design two regularization losses, the inference loss and the continuity loss, to extract and refine intents. The former enhances the representative capacity of behavioral trends by inferring future action sequences, while the latter reinforces stability over time by measuring intent similarity across adjacent time intervals. Unlike past or present information, which can become diluted over time, delayed intents can still assist the recipient in understanding the situation. Subsequently, we design a dual alignment mechanism to integrate multi-source asynchronous messages. The first alignment prioritizes messages related to their intents via an attention module, and the second one focuses on new messages based on the timeliness of transmission. We implement the two types of delay settings in commonly used MARL benchmarks such as SMAC (Samvelyan et al. 2019), GRF (Kurach et al. 2020), and Hallway (Wang et al. 2020b). Extensive experiment results show that CoDe exhibits significant delay robustness.

Our contributions can be outlined as follows:

- According to our knowledge, we are the first to investigate asynchronous communication in MARL and propose two delay settings.
- We introduce an intent extraction framework based on future action inference, which can represent agents’ stable future behavioral trends.
- We design a dual alignment mechanism of intent and timeliness to integrate multi-source asynchronous messages, mitigating the impact of delays.
- We realize two delay settings in SMAC, GRF, and Hallway. CoDe consistently outperforms baseline under zero-delay, fixed-delay, and time-varying-delays.

Related Works

Communication-Constrained MARL. Considering the constraints of limited channel bandwidth, several works concentrated on the sparsity of multi-agent communication (Fo-

erster et al. 2016; Zhang, Zhang, and Lin 2019; Wang et al. 2020b; Zhang, Zhang, and Lin 2020; Yuan et al. 2022; Sun et al. 2024; Cui 2024). NDQ (Wang et al. 2020b) designed two information-theoretic regularizers that maximize the mutual information between action selection and messages while minimizing the entropy of inter-agent messages. TMC (Zhang, Zhang, and Lin 2020) employed a time-window technique to mitigate the temporal variability of messages, thereby reducing the communication frequency. MAIC (Yuan et al. 2022) introduced an entropy regularization term to increase the concentration of attention and pruned communication links with low scores. T2MAC (Sun et al. 2024) quantified the necessity of communication by leveraging uncertainty reduction through different evidence ablation. Furthermore, some works (Mitchell, Blumenkamp, and Prorok 2020; Xue et al. 2021; Yuan et al. 2024; Ding et al. 2024; Yu et al. 2024) focused on noisy or adversarial communication settings. To our best knowledge, there is a lack of research on the channel delay issue in multi-agent scenarios, which has motivated our study.

Delay in RL. In the Markov Decision Process (MDP), delays are typically unavoidable, such as delays in state perception, action execution, and reward feedback. DATS (Chen et al. 2021) incorporated multi-step action delays into the environment model, effectively mitigating the impact of delays. DA-MAVL (Zhang et al. 2023) introduced a methodology for learning coarse-correlated equilibria under time-varying delays in reward feedback. DAMARL (Chen et al. 2020) researched observation delays in multi-agent settings by incorporating future actions into local observations to facilitate decision-making. DACOM (Yuan et al. 2023) first integrated communication delays into multi-agent reinforcement learning, primarily focusing on decision-making of waiting time. Additionally, DACOM impractically assumed the delays within the interval between two consecutive decisions. In contrast, our research investigates explicitly the more general settings where channel delays surpass the decision interval, leading to asynchronous communication.

Background

Considering the problem of asynchronous communication, we first formulate multi-agent decision-making as Delay-Tolerant Decentralized Partially Observable Markov Decision Process (DT-Dec-POMDP), which can be formed as $\langle \mathcal{N}, \mathcal{S}, \Omega, \mathcal{O}, \mathcal{A}, \mathcal{R}, \mathcal{P}, \gamma, \mathcal{M}, \mathcal{D} \rangle$. Specifically, \mathcal{N} represents multiple agents in the system. \mathcal{S} denotes the global state space that reflects an exhaustive environment overview. \mathcal{O} signifies the partial observation space for agents and Ω represents the observation function that maps states to observations. \mathcal{A} refers to the joint action space. \mathcal{R} stands for the reward function used to assess the quality of decisions. \mathcal{P} acts as the environmental dynamics model, illustrating the influence of joint actions on states. γ is the discount factor that balances long-term and short-term returns. Expanding the original Dec-POMDP, \mathcal{M} delineates the message-passing within the system, which is characterized by point-to-point communication, facilitating the direct exchange of information between individual agents. $\mathcal{D} \in \{\mathcal{D}_f, \mathcal{D}_v\}$ rep-

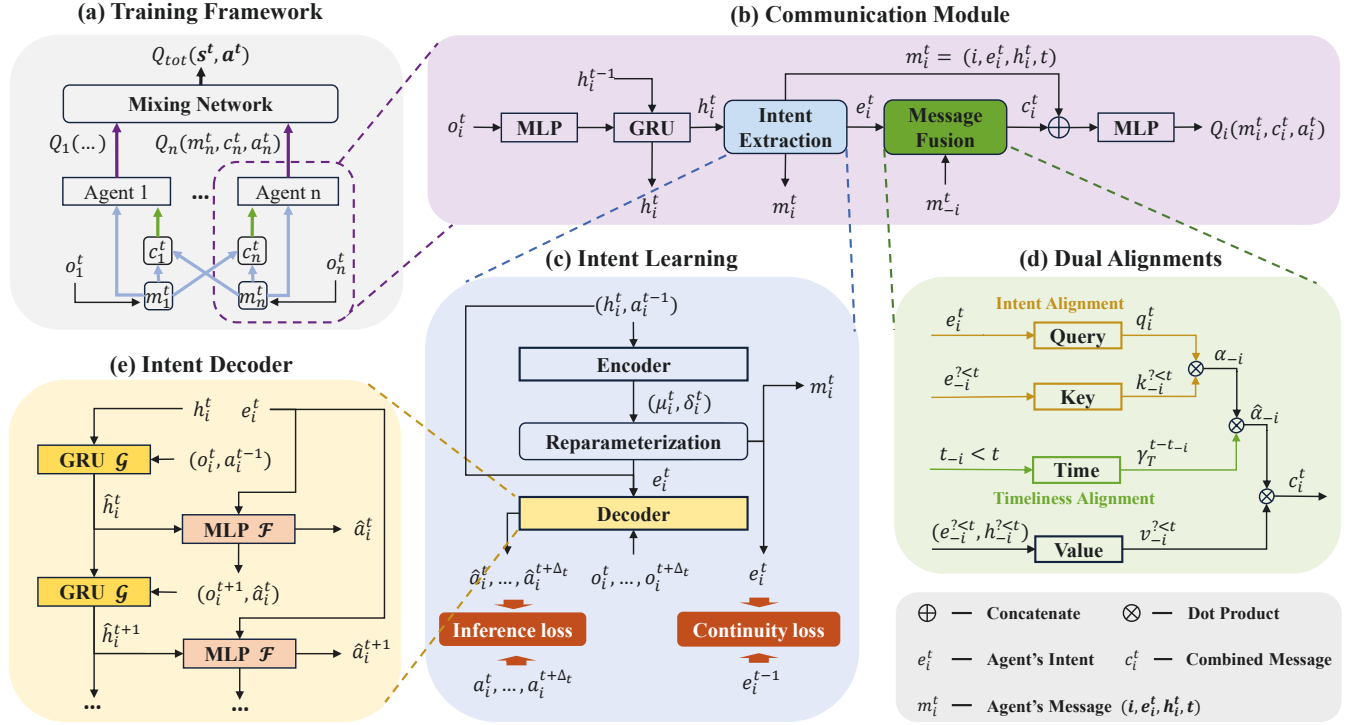


Figure 2: Overall Framework of CoDe. (a) The training framework; (b) The communication module consisting of intent extraction, message propagation, and message fusion; (c) The intent learning module via two designed losses; (d) The dual alignment module via a modified attention structure, in which “ $-i$ ” identifies the index of agents other than i and “ $? < t$ ” represents certain timestamp before t ; (e) The sequence prediction model used to decode the intent to future actions.

resents the instantaneous delay on the point-to-point channel, indicating that the messages received by agents at each moment were sent in the previous few moments.

- **Fixed Delay** D_f : $\forall_{i,j,t} d_{i,j}^t = d_f$, where $d_{i,j}^t$ denotes the delay between agent i and j at time t . d_f is a fixed value;
- **Time-Varying Delay** D_v : $d_{i,j}^t \sim \mathcal{N}(\mu, \delta)$, where \mathcal{N} represents the predefined Gaussian distribution of delay.

Method

Our intuitive perspective suggests that in a delayed environment, messages should reflect the sender’s stable behavioral trends while the receiver selectively utilizes outdated messages. Even if messages are delayed, receivers should still be able to extract some insights about the sender.

CoDe contains the mixing network in Fig. 2 (a) and the shared individual communication module in Fig. 2 (b). The core of CoDe is twofold: one aspect focuses on optimizing message design at the sender’s end to extract and convey long-term intents (Fig. 2 (c, e)); the other aspect centers on optimizing message fusion at the receiver’s end, employing a dual alignment of intent and timeliness to integrate multi-source asynchronous messages effectively (Fig. 2 (d)).

Intent Learning Via Future Action Inference

CoDe aims to learn an intent representation that can reflect agents’ future behavioral trends to enhance inter-agent com-

munication. Learned intents improve the collaborative abilities of agents in two aspects: firstly, transmitting intents aids in the comprehension of behaviors between agents, and secondly, intents can then act as verification for message fusion, enabling the selective use of multi-source messages.

Under channel delays, intents should possess two fundamental properties: (1) expressiveness of future behavior and (2) stability within a time window. Here, two regularization losses are devised to reinforce these two properties.

Inference Loss. To begin with, the inference loss is designed to augment the capacity to represent future action sequences. We believe that the historical trajectory of agents contains their behavioral motivations, which support agents in making coherent decisions. Hence, CoDe relies on the past to envision the future via an approximate seq-to-seq structure, depicted in Fig. 2 (c).

To handle the historical trajectory $\tau = \{o_i^1, o_i^2, \dots, o_i^t\}$, CoDe utilizes a GRU structure to obtain a compressed representation h_i^t . Subsequently, the preceding action a_i^{t-1} is concatenated and fed into the encoder to obtain the distribution of intents $\mathcal{N}(\mu_i^t, \delta_i^t)$. Following reparameterization, we obtain the intent vector e_i^t that preserves gradients.

Afterward, a sequence prediction model is introduced to forecast actions for the next K steps, as shown in Fig. 2 (e). Specifically, CoDe employs a GRU structure \mathcal{G} and an action prediction network \mathcal{F} for intent decoding rather than

the commonly used transformer structure due to the limited sequence length. Firstly, we initialize the hidden state of \mathcal{G} with the historical trajectory encoding h^t at the current time step to enrich agents' comprehension of the past. In the following K steps, the agent's current observations o^{t+k} and the predicted action \hat{a}^{t+k-1} from the previous time step are employed as inputs to \mathcal{G} . Then, the hidden state \hat{h}^{t+k} processed through \mathcal{G} and the intent e^t from the initial time are passed to the prediction network \mathcal{F} to forecast action. The difference between predicted actions \hat{a}^{t+k} and ground-truth actions a^{t+k} serves as the inference loss to train the network and refine the intent representation, defined as Eq. 1. In the case of discrete action space, $\|\cdot, \cdot\|$ denotes the cross-entropy loss. In contrast, for a continuous space, it signifies the mean squared error loss.

$$\mathcal{L}_{inf} = \mathbb{E}_{\tau \sim \mathcal{B}} \left[\sum_{k=1}^{K-1} \|\mathcal{F}(e_i^t, \mathcal{G}(o_i^{t+k}, \hat{a}_i^{t+k-1})), a_i^{t+k}\| + \|\mathcal{F}(e_i^t, \mathcal{G}(o_i^t, a_i^{t-1})), a_i^t\| \right]. \quad (1)$$

Continuity Loss. Given that intent reflects future behavioral trends, it is expected to exhibit continuity over the short term. We compute the cosine similarity between intents generated at adjacent time steps and maximize it to enhance the consistency of intents, formulated as:

$$\mathcal{L}_c = \mathbb{E}_{\tau \sim \mathcal{B}} \left[-\frac{e_i^{t-1} \cdot e_i^t}{\|e_i^{t-1}\| \|e_i^t\|} \right]. \quad (2)$$

Demanding stable intents via Eq. 2 can narrow the intent distribution, limiting exploration. Hence, we introduce an auxiliary loss \mathcal{L}_k to increase the diversity of intents. Formally, we compute the Kullback-Leibler divergence between the intent distribution and a standard normal distribution $D_{KL}(\mathcal{N}(\mu_i^t, \delta_i^t) \parallel \mathcal{I})$. \mathcal{L}_k can be formulated as Eq. 3, in which n denotes the dimension of intents.

$$\mathcal{L}_k = \mathbb{E}_{\tau \sim \mathcal{B}} \left[-\frac{1}{2} \sum_{i=1}^n (\delta_i^2 + \mu_i^2 - 1 - \log(\delta_i^2)) \right]. \quad (3)$$

In addition to the training loss for subsequent reinforcement learning, we design two regularization losses and an auxiliary loss to optimize intent representation, balancing their effects through hyperparameters formed as:

$$\mathcal{L}_{int} = \lambda_{inf} \mathcal{L}_{inf} + \lambda_c \mathcal{L}_c + \lambda_k \mathcal{L}_k. \quad (4)$$

Dual Alignment Of Intent and Timeliness

We posit that, even though the sender can promptly convey its intents, the receiver encounters challenges in managing asynchronous messages from multiple sources. In the CoDe framework, the structure of a message is represented as (i, e_i^t, h_i^t, t) . e_i^t represents the intent involving the agent's future behavioral trend, h_i^t signifies the message content encompassing the agent's insights, and t denotes the delivery time. To prioritize the integration of messages, we propose a dual alignment mechanism in Fig. 2 (d). On the one hand, the learned intent can serve as credentials, prompting agents with similar intents to pay more attention to each other. On the other hand, delay also acts as a criterion, with newer messages carrying greater significance for the recipient.

Intent Alignment. Leveraging intent as the credential, we utilize an attention mechanism (Vaswani et al. 2017) for the initial alignment. The attention mechanism is extensively employed in multi-source information fusion for its capacity to extract features regardless of sequence. In our design, the receiver's intent e_i acts as the query, the sender's intent e_{-i} serves as the key, and the sender's content combined with intents (e_{-i}, h_{-i}) serves as the value. Formally, the attention score is calculated as follows:

$$\alpha_{i,j} = \frac{\exp\left(\frac{1}{\sqrt{d_k}} \cdot e_i W^Q \cdot (e_j W^K)^\top\right)}{\sum_{u=1}^{-i} \exp\left(\frac{1}{\sqrt{d_k}} \cdot e_i W^Q \cdot (e_u W^K)^\top\right)}, \quad (5)$$

where W^Q and W^K represent parameter matrices for linearly transforming queries and keys. $\frac{1}{\sqrt{d_k}}$ serves as a scaling factor for dot-product attention. $\alpha_{i,j}$ defines the correlation between agent j 's message and agent i . To enhance attention concentration, we drew inspiration from (Yuan et al. 2022), introducing an entropy regularization term to minimize uncertainty by reducing entropy.

$$\mathcal{L}_e = -\lambda_e \sum_{i=1}^n \mathcal{H}(\alpha_{i,\cdot}) = -\lambda_e \sum_{i \neq j}^n \alpha_{i,j} \log \alpha_{i,j}. \quad (6)$$

Timeliness Alignment. Due to delays, the time lag between sending and receiving an intent can be substantial. In such instances, the sender's intent may have evolved, and aligning with outdated intents could easily lead to the breakdown of collaborative behaviors. Here, we employ a straightforward yet effective approach to address asynchronous messages by a secondary decay. In detail, CoDe incorporates a temporal discount factor γ_T for exponential decay to the weight of intent alignment $\alpha_{i,j}$. As shown in Eq. 6, Δt represents the time interval since the message m_j was sent.

$$\hat{\alpha}_{i,j} = \alpha_{i,j} * \gamma_T^{\Delta t}. \quad (7)$$

Following the dual alignment process, we derive attention weights that balance intent and timeliness. Subsequently, weighted message fusion is carried out, expressed as

$$c_i^t = \sum_{j=1}^{-i} \hat{\alpha}_{i,j} (e_j, h_j) W^V, \quad (8)$$

where W^V denotes the linear transformation for part of messages (e_j, h_j) . Notably, we assume zero channel delays during training. Hence, Eq. 5, 7, and 8 vary between the training and testing phases.

Overall Training Objective

After message fusion, the integrated information is incorporated into the MARL training process. This paper adopts the foundational framework QMIX (Rashid et al. 2020) while also adaptable to any value decomposition algorithm like VDN (Sunehag et al. 2017) or QPLEX (Wang et al. 2020a). \mathcal{L}_{RL} is defined as the standard temporal difference (TD) loss using Q_{tot} from the mixing network. By integrating intent learning and the dual alignment mechanism, the complete training loss is as follows:

$$\mathcal{L}_{tot} = \mathcal{L}_{RL} + \mathcal{L}_{int} + \mathcal{L}_e. \quad (9)$$

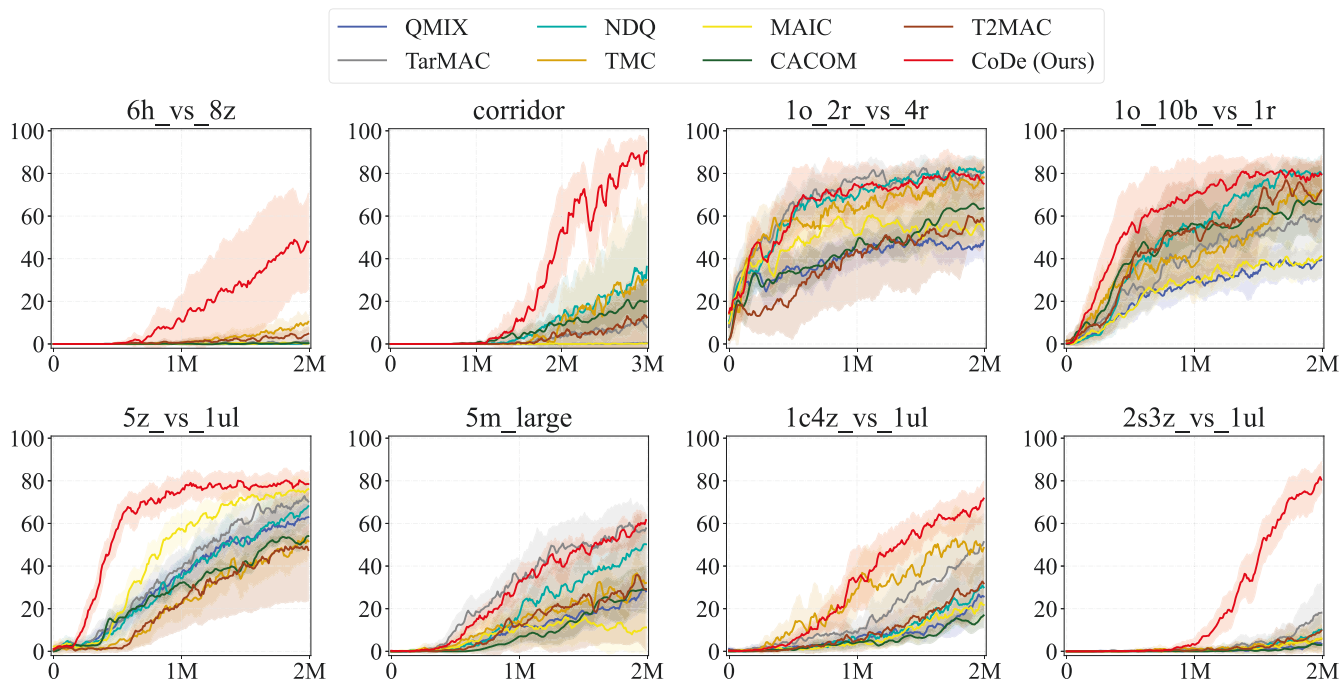


Figure 3: Algorithms Performance in SMAC under Zero Communication Delay. Each curve represents the average result of 5 random seeds. The last three are our proposed new maps.

Delayed Communication Protocol

After training, we comprehensively describe the decision-making process in delayed scenarios, as illustrated in Fig. 2 (b). At each time step, the agent perceives the local environment to gather partial observations o_i^t , which are then encoded into a historical trajectory representation h_i^t . The current intent distribution is derived by feeding h_i^t and the previous action a_i^{t-1} into the intent encoder, and the intent e_i^t is obtained after reparameterization sampling. Then, the agent constructs and broadcasts messages m_i^t . Concurrently, as a recipient, the agent maintains a message buffer that stores the most recent messages from other agents and performs dual alignments to integrate asynchronous messages $m_{-i}^{?<t}$ from diverse sources. The integrated information c_i^t , alongside its own messages m_i^t , is jointly fed into the local Q-values $Q_i(\dots)$ for decision-making. The pseudocode for training and testing can be found in the Appendix.

Experiments

To evaluate CoDe, we conduct experiments on three benchmarks, including Hallway (Wang et al. 2020b), SMAC (Samvelyan et al. 2019), and GRF (Kurach et al. 2020). In the section, we select seven algorithms as baselines, including QMIX (Rashid et al. 2020), TarMAC (Das et al. 2019), NDQ (Wang et al. 2020b), TMC (Zhang, Zhang, and Lin 2020), MAIC (Yuan et al. 2022), CACOM (Li and Zhang 2023) and T2MAC (Sun et al. 2024). We aim to validate the following three questions experimentally: (1) Does the intent-based communication mechanism exhibit superior performance in a zero-delay environment? (2) Does our pro-

posed CoDe demonstrate robustness under various communication delay conditions? (3) Does the learned intent effectively represent the future action trend of agents?

Performance in Delay-free Environments

We evaluate the algorithm under zero-delay conditions to assess the effectiveness of inter-agent intent transmission. Since CoDe does not consider communication concurrency, for fairness, the experiments in this section employ a fully connected communication structure for all other baselines.

SMAC. As the most popular MARL benchmark, SMAC exhibits limited suitability under communication conditions due to constrained environmental randomness and local observability. Hence, apart from two original super-hard maps and three communication-enhanced maps introduced by NDQ, we additionally devise three supplementary maps, as shown in the Appendix. In our designed maps, the initial positions of the units are randomized, necessitating agent exploration and identification of both allies and enemies. Furthermore, to defeat enemies, agents must develop tactical strategies, underscoring the critical role of communication. As depicted in Fig. 3, CoDe consistently demonstrates impressive performance and sampling efficiency across nearly all maps. Comparably, most baselines struggle to produce effective results on the newly introduced maps. In scenarios with high levels of randomness and the requirement for collaborative exploration, we believe that agents should transmit messages about themselves rather than suggestions to others. As for how the messages are utilized, it should be left to the receiving agents to determine.

Delay	Map	1o_2r_vs_4r	1o_10b_vs_1r	5z_vs_1ul	5m_large	1c4z_vs_1ul	2s3z_vs_1ul
3	NDQ	63.1 ± 7.4	52 ± 3.1	39.2 ± 4.5	39.6 ± 3.1	6.3 ± 1.4	8.2 ± 1.7
	TarMAC	33.3 ± 4.5	47.2 ± 5.2	48.4 ± 7.9	28.2 ± 6	<u>61.1 ± 10.5</u>	<u>43.1 ± 6.9</u>
	MAIC	48 ± 7.5	26.2 ± 8.4	<u>66.8 ± 5.8</u>	8.4 ± 3.4	14.4 ± 6.4	6 ± 3.6
	T2MAC	56.3 ± 3.3	<u>55.3 ± 9.1</u>	43.8 ± 7.6	13.1 ± 5.2	15.3 ± 3.7	15.1 ± 4.4
	CoDe	<u>60.8 ± 7.5</u>	59.3 ± 8.1	67.7 ± 4.3	<u>33.7 ± 1.4</u>	72 ± 5.8	89.3 ± 3.3
5	NDQ	48.2 ± 3.9	<u>46.7 ± 4.5</u>	37.8 ± 3.2	<u>25.4 ± 4.1</u>	5.5 ± 1.7	7.7 ± 0.8
	TarMAC	24.7 ± 4.8	39.5 ± 6.5	42.7 ± 9.6	12 ± 3.7	<u>56.4 ± 5.9</u>	<u>32.8 ± 5.4</u>
	MAIC	44.6 ± 5.7	25.1 ± 6.1	<u>61.3 ± 7.2</u>	6.2 ± 4.1	11.1 ± 5.4	4.6 ± 2.8
	T2MAC	<u>55.1 ± 6.4</u>	43.5 ± 8.8	39.3 ± 6.7	3.5 ± 2.4	13.7 ± 4.5	17.1 ± 5.8
	CoDe	52.3 ± 6.1	49.8 ± 9.3	68.6 ± 7.5	29.7 ± 2.3	64 ± 6.1	87.3 ± 4.9

Table 1: Results in SMAC under Fixed Communication Delay. Bold and underlined represent the first and second performances.

Delay	Map	1o_2r_vs_4r	1o_10b_vs_1r	5z_vs_1ul	5m_large	1c4z_vs_1ul	2s3z_vs_1ul
$\mathcal{N}(3, 2)$	NDQ	66 ± 7.8	59.2 ± 4	41.3 ± 6	<u>34.8 ± 3.8</u>	6.4 ± 1.3	8.1 ± 2.3
	TarMAC	47.1 ± 4.6	45.2 ± 5.4	44.7 ± 5.3	10.5 ± 1.6	<u>48.2 ± 3.1</u>	<u>20.9 ± 2.6</u>
	MAIC	49.1 ± 4.5	27.6 ± 4.4	<u>67.7 ± 5.8</u>	16.9 ± 3.8	14.6 ± 4.1	6.8 ± 1.6
	T2MAC	59.5 ± 4.1	55.1 ± 6.1	47.5 ± 5.3	15.4 ± 3.4	17.7 ± 3.3	18 ± 2.5
	CoDe	69.2 ± 6.3	62.1 ± 7.5	72.2 ± 4.8	36.7 ± 2.1	71 ± 5.3	88 ± 2.9
$\mathcal{N}(5, 2)$	NDQ	50.3 ± 3.1	<u>50.8 ± 3.4</u>	38.8 ± 6.1	36.7 ± 4.6	5.7 ± 1.2	7.6 ± 1.1
	TarMAC	38.1 ± 6.9	38.3 ± 4.1	38.5 ± 5.5	4.5 ± 1.4	<u>49 ± 7.8</u>	<u>21.8 ± 2.3</u>
	MAIC	44.9 ± 5	27.3 ± 4.2	<u>59.6 ± 4.1</u>	7.9 ± 2.1	14.9 ± 4.1	6.2 ± 4.3
	T2MAC	<u>57.8 ± 6</u>	48.1 ± 7.2	42.1 ± 5	6.4 ± 7.6	16.8 ± 2.8	14.9 ± 2.6
	CoDe	58.9 ± 8.5	54.1 ± 10.4	66.7 ± 5	<u>28.3 ± 3.8</u>	67.4 ± 4.6	88 ± 2.7

Table 2: Results in SMAC under Time-Varying Communication Delay.

GRF. We further evaluate CoDe on the more challenging GRF; the results are depicted in Fig. 4. Compared with baselines, CoDe achieves high team scoring rates across two tasks. Due to the more vital interactivity involving actions like passing among agents in GRF, conveying one’s behavioral intents enhances the coordination of agents’ actions.

Hallway. In the Hallway benchmark, multiple agents navigate parallel linear tracks by moving left and right. The maximum reward is achieved when all agents simultaneously reach the goal for the first time. We devise two tasks to control the coordinated movement of 4 and 5 agents. As illustrated in Fig. 4, CoDe achieves remarkable performance across two tasks compared to the baseline. Building upon pervasive communication, we attribute this performance to future behavioral intents within CoDe. In high-risk environments such as Hallway, incorporating long-term intents proves advantageous for coordinating the agents’ actions and mitigating risky decisions. In contrast, prior works like NDQ and TarMAC tend to convey behavioral suggestions to other agents, which leads to a relatively unclear comprehension of the sender’s state and intents for the recipients.

Performance in Delayed Environments

Next, we implement two communication delay settings to assess the algorithm’s robustness, involving sampling point-to-point channel delays from either a Gaussian distribution

or a fixed value at each time step. Notably, to handle out-of-time-order messages, we discard any older timestamped messages received after the arrival of a message with a new timestamp. Additionally, considering the need for extra storage space in CoDe, we augment the storage capacity of other baseline algorithms for fairness. In cases where new messages are unavailable, agents can rely on the most recent messages to make decisions.

Results. We perform two sets of experiments in SMAC, one with fixed delays and the other with varying ones. In Tab. 1 and 2, we observe that CoDe maintains stable performance across various delay settings, validating the effectiveness of the dual alignment mechanism. Under fixed delays, the performance of CoDe slightly declines compared to varying delays. This can be attributed to the synchronized message sources, which weaken timeliness alignment. The consistent timestamps of messages lead to uniform down-weighting during information fusion, causing agents to focus more on themselves and reducing collaboration. Notably, prior works, such as NDQ and TarMAC, which center on communication sparsity, also exhibit a degree of delay robustness. The sparse communication structure implies that only a small subset of messages influences decision-making, alleviating the negative impact of delays.

Ablation. To validate the dual alignment mechanism, we execute an ablation experiment in Tab. 3. Results show that

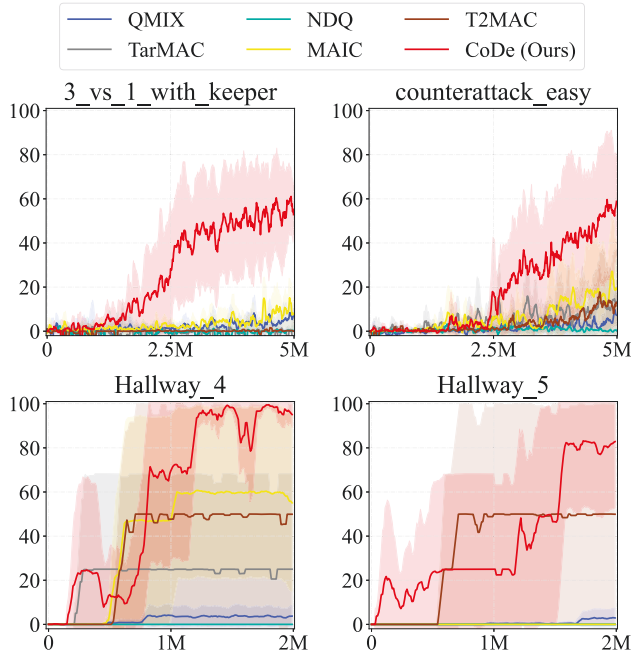


Figure 4: Performance in GRF and Hallway under Zero Communication Delay. The first line represents the result in GRF, while the second line represents the result in Hallway.

Maps	CoDe	CoDe w/o IA	CoDe w/o TA	CoDe w/o DA
1o_10b_vs_1r	62.1	35.6	58.5	33.8
1c4z_vs_1ul	71	60.4	65.8	59.5
2s3z_vs_1ul	88	79.6	83.7	78.3

Table 3: Ablation over alignment modules. “IA” means the intent alignment, “TA” means the timeliness alignment, and “DA” means dual alignments.

removing any of the alignment modules leads to a decline in the delay robustness of CoDe, demonstrating the effectiveness of dual alignment. Comparably, intent alignment exhibits a more significant impact.

Analysis Over The Learned Intent

We analyze the learned intent by visualizing its capability to represent future behaviors while conducting ablation experiments on various losses during the intent training process.

Visualization. We decode the learned intents, converting them into future sequences of actions, and visualize them within the Hallway environment. From Fig. 5, it can be observed that decoded action sequences reflect the accurate behavioral trends of agents at the current time step.

Ablation. A set of ablation experiments is implemented to verify the effectiveness of the proposed inference loss and continuity loss. Fig. 6 captures the performance changes of the ablated CoDe in both delay-free and delayed scenarios.

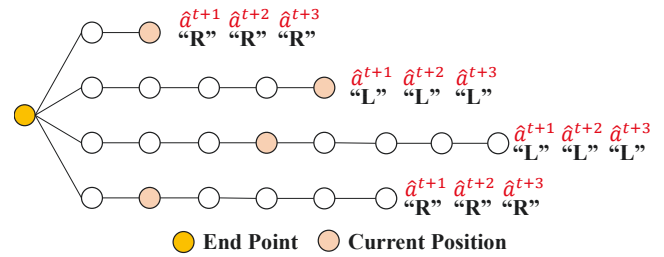


Figure 5: Visualization of the learned intent. “L” and “R” represent the left and right actions, respectively. $\hat{a}^{t+?}$ refers to the expected action at a future time, as decoded from the intent at time t .

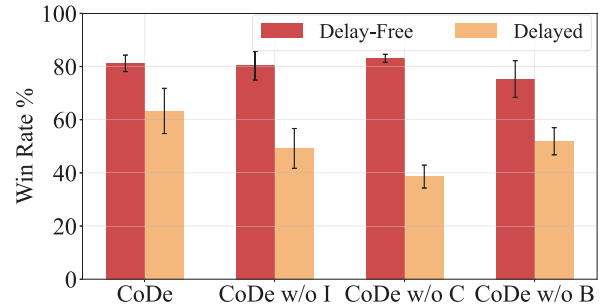


Figure 6: Ablation over two regularization losses. “I” means the inference loss, “C” means the continuity loss, and “B” means both losses.

Evidently, in the absence of delays, removing losses does not yield significant performance variations. This can be attributed to the immediate message transmission, which does not demand long-term expressive and stable intent representation. However, in the presence of delays, any form of ablation leads to a noticeable performance deterioration, confirming the effectiveness of the two losses. Regarding CoDe w/o B, the learned intent vectors tend to converge towards random variables upon eliminating two losses, diminishing their communicative efficacy. Consequently, the impact of delayed random intent is also relatively minor.

Conclusions

This study explores a new field concerning the asynchronous communication issue in MARL under channel delays. To address this issue, we propose a new communication framework called CoDe. For senders, CoDe extracts and conveys intents that reflect future behavioral trends. For receivers, CoDe introduces a dual alignment mechanism of intents and timeliness to merge asynchronous messages. Empirical findings indicate that CoDe outperforms baselines across various delay scenarios. However, CoDe is limited by the length of future action prediction and struggles with more severe delays, a challenge we aim to address in future work. We aspire that this work will inspire increased interest in the research of MARL within communication delay scenarios.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 62206013).

References

- Chen, B.; Xu, M.; Li, L.; and Zhao, D. 2021. Delay-aware model-based reinforcement learning for continuous control. *Neurocomputing*, 450: 119–128.
- Chen, B.; Xu, M.; Liu, Z.; Li, L.; and Zhao, D. 2020. Delay-aware multi-agent reinforcement learning for cooperative and competitive environments. *arXiv preprint arXiv:2005.05441*.
- Cui, J. 2024. Communication and Generalization in Multi-Agent Learning. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '24, 2728–2730. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9798400704864.
- Das, A.; Gervet, T.; Romoff, J.; Batra, D.; Parikh, D.; Rabat, M.; and Pineau, J. 2019. Tarmac: Targeted multi-agent communication. In *International Conference on machine learning*, 1538–1546. PMLR.
- Ding, S.; Du, W.; Ding, L.; Guo, L.; and Zhang, J. 2024. Learning Efficient and Robust Multi-Agent Communication via Graph Information Bottleneck. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 17346–17353.
- Ding, Z.; Huang, T.; and Lu, Z. 2020. Learning individually inferred communication for multi-agent cooperation. *Advances in neural information processing systems*, 33: 22069–22079.
- Foerster, J.; Assael, I. A.; De Freitas, N.; and Whiteson, S. 2016. Learning to communicate with deep multi-agent reinforcement learning. *Advances in neural information processing systems*, 29.
- Han, S.; Dastani, M.; and Wang, S. 2023. Model-based Sparse Communication in Multi-agent Reinforcement Learning. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, 439–447.
- Hu, G.; Zhu, Y.; Zhao, D.; Zhao, M.; and Hao, J. 2020. Event-triggered multi-agent reinforcement learning with communication under limited-bandwidth constraint. *arXiv preprint arXiv:2010.04978*.
- Jiang, J.; and Lu, Z. 2018. Learning attentional communication for multi-agent cooperation. *Advances in neural information processing systems*, 31.
- Kurach, K.; Raichuk, A.; Stańczyk, P.; Zajac, M.; Bachem, O.; Espeholt, L.; Riquelme, C.; Vincent, D.; Michalski, M.; Bousquet, O.; et al. 2020. Google research football: A novel reinforcement learning environment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 4501–4510.
- Li, X.; and Zhang, J. 2023. Context-aware Communication for Multi-agent Reinforcement Learning. *arXiv preprint arXiv:2312.15600*.
- Mitchell, R.; Blumenkamp, J.; and Prorok, A. 2020. Gaussian process based message filtering for robust multi-agent cooperation in the presence of adversarial communication. *arXiv preprint arXiv:2012.00508*.
- Rashid, T.; Samvelyan, M.; De Witt, C. S.; Farquhar, G.; Foerster, J.; and Whiteson, S. 2020. Monotonic value function factorisation for deep multi-agent reinforcement learning. *Journal of Machine Learning Research*, 21: 1–51.
- Samvelyan, M.; Rashid, T.; De Witt, C. S.; Farquhar, G.; Nardelli, N.; Rudner, T. G.; Hung, C.-M.; Torr, P. H.; Foerster, J.; and Whiteson, S. 2019. The starcraft multi-agent challenge. *arXiv preprint arXiv:1902.04043*.
- Sheng, J.; Wang, X.; Jin, B.; Yan, J.; Li, W.; Chang, T.-H.; Wang, J.; and Zha, H. 2022. Learning structured communication for multi-agent reinforcement learning. *Autonomous Agents and Multi-Agent Systems*, 36(2): 50.
- Sun, C.; Zang, Z.; Li, J.; Li, J.; Xu, X.; Wang, R.; and Zheng, C. 2024. T2MAC: Targeted and Trusted Multi-Agent Communication through Selective Engagement and Evidence-Driven Integration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 15154–15163.
- Sunehag, P.; Lever, G.; Gruslys, A.; Czarnecki, W. M.; Zambaldi, V.; Jaderberg, M.; Lanctot, M.; Sonnerat, N.; Leibo, J. Z.; Tuyls, K.; et al. 2017. Value-decomposition networks for cooperative multi-agent learning. *arXiv preprint arXiv:1706.05296*.
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- Wang, J.; Ren, Z.; Liu, T.; Yu, Y.; and Zhang, C. 2020a. Qplex: Duplex dueling multi-agent q-learning. *arXiv preprint arXiv:2008.01062*.
- Wang, T.; Wang, J.; Zheng, C.; and Zhang, C. 2020b. Learning Nearly Decomposable Value Functions Via Communication Minimization. In *International Conference on Learning Representations*.
- Xue, W.; Qiu, W.; An, B.; Rabinovich, Z.; Obraztsova, S.; and Yeo, C. K. 2021. Mis-spoke or mis-lead: Achieving robustness in multi-agent communicative reinforcement learning. *arXiv preprint arXiv:2108.03803*.
- Yu, L.; Qiu, Y.; Yao, Q.; Shen, Y.; Zhang, X.; and Wang, J. 2024. Robust Communicative Multi-Agent Reinforcement Learning with Active Defense. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 17575–17582.
- Yuan, L.; Chen, F.; Zhang, Z.; and Yu, Y. 2024. Communication-robust multi-agent learning by adaptable auxiliary multi-agent adversary generation. *Frontiers of Computer Science*, 18: 186331.
- Yuan, L.; Wang, J.; Zhang, F.; Wang, C.; Zhang, Z.; Yu, Y.; and Zhang, C. 2022. Multi-agent incentive communication via decentralized teammate modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 9466–9474.

Yuan, T.; Chung, H.-M.; Yuan, J.; and Fu, X. 2023. DA-COM: Learning delay-aware communication for multi-agent reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 11763–11771.

Zhang, S. Q.; Zhang, Q.; and Lin, J. 2019. Efficient communication in multi-agent reinforcement learning via variance based control. *Advances in neural information processing systems*, 32.

Zhang, S. Q.; Zhang, Q.; and Lin, J. 2020. Succinct and robust multi-agent communication with temporal message control. *Advances in neural information processing systems*, 33: 17271–17282.

Zhang, Y.; Zhang, R.; Gu, Y.; and Li, N. 2023. Multi-agent reinforcement learning with reward delays. In *Learning for Dynamics and Control Conference*, 692–704. PMLR.