M2I2: Learning Efficient Multi-Agent Communication via Masked State Modeling and Intention Inference

Chuxiong Sun[1*], Peng He[3*], Qirui Ji[1,2*], Zehua Zang[1,2], Jiangmeng Li[1†], Rui Wang[1,4], Wei Wang[3†]

[1]National Key Laboratory of Space Integrated Information System, Institute of Software, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
[3]Beijing University of Posts and Telecommunications
[4]National Key Laboratory of Complex System Modeling and Simulation Technology
{chuxiong2016, jiqirui2022, zehua2020, jiangmeng2019, wangrui}@iscas.ac.cn
{hepeng123, wangwei}@bupt.edu.cn

Abstract

Communication is essential in coordinating the behaviors of multiple agents. However, existing methods primarily emphasize content, timing, and partners for information sharing, often neglecting the critical aspect of integrating shared information. This gap can significantly impact agents' ability to understand and respond to complex, uncertain interactions, thus affecting overall communication efficiency. To address this issue, we introduce M2I2, a novel framework designed to enhance the agents' capabilities to assimilate and utilize received information effectively. M2I2 equips agents with advanced capabilities for masked state modeling and joint-action prediction, enriching their perception of environmental uncertainties and facilitating the anticipation of teammates' intentions. This approach ensures that agents are furnished with both comprehensive and relevant information, bolstering more informed and synergistic behaviors. Moreover, we propose a Dimensional Rational Network, innovatively trained via a meta-learning paradigm, to identify the importance of dimensional pieces of information, evaluating their contributions to decision-making and auxiliary tasks. Then, we implement an importance-based heuristic for selective information masking and sharing. This strategy optimizes the efficiency of masked state modeling and the rationale behind information sharing. We evaluate M2I2 across diverse multi-agent tasks, the results demonstrate its superior performance, efficiency, and generalization capabilities, over existing state-of-the-art methods in various complex scenarios.

Extended version — https://arxiv.org/abs/2501.00312

1 Introduction

Reinforcement Learning (RL) has achieved significant milestones in various complex real-world applications, from Game AI (Osband et al. 2016; Silver et al. 2017, 2018; Vinyals et al. 2019), Robotics (Andrychowicz et al. 2020; Zang et al. 2026) and Autonomous Driving (Leurent 2018) to applications in Large Language Models (Shao et al. 2024; Wang et al. 2025; Gu et al. 2025; Chen et al. 2025). However, the landscape shifts markedly when applied to Multi-Agent Reinforcement Learning (MARL) (Lowe et al. 2017; Rashid et al. 2018; Yu et al. 2022; Chen et al. 2022; Zang et al. 2025), where unique challenges emerge. A principal challenge is the issue of partial observability, where agents must make decisions based on limited local observations, lacking a comprehensive view of the entire environment. In addressing this challenge, multi-agent communication emerges as a potent solution. By enabling agents to *share* and *integrate* information, this strategy facilitates a deeper collective understanding of their environment, stabilizing the learning process and promoting synchronized actions among agents.

Despite advancements, existing methods in multi-agent communication primarily focus on sending policies, such as creating meaningful messages (Zhang, Zhang, and Lin 2019, 2020; Yuan et al. 2022), optimizing the timing (Singh, Jain, and Sukhbaatar 2018; Kim et al. 2019; Hu et al. 2021) and selecting appropriate partners (Ding, Huang, and Lu 2020; Niu, Paleja, and Gombolay 2021; Xue et al. 2022) for information exchange. However, these methods exhibit a significant gap in effectively integrating received information to enhance decision-making at receiving end. Typically, a large volume of received messages, processed by basic mechanisms like concatenation (Sukhbaatar, Szlam, and Fergus 2016) is fed directly into policy networks. This approach treats the information integration task as a black box, presupposing that neural networks can autonomously discern the most decision-important information, overlooking the intricacies of cognition and collaborative decision-making. In contrast, human cognitive processes (Etel and Slaughter 2019) demonstrate a superior ability for utilizing received information to perceive the environment and reduce the uncertainty of decision-making. This level of decision-making complexity, inherent in human cognition, is something that current multi-agent communication methods fail to capture.

Inspired by human cognitive processes and recent advancements in representation learning, as illustrated by models like BERT (Devlin et al. 2018) and Masked Auto-Encoder (MAE)

*These authors contributed equally.
†Jiangmeng Li and Wei Wang are corresponding authors.

(He et al. 2022), we redefine the challenge of information integration in multi-agent communication as one of representation learning task. In this context, agents are tasked with developing representations for cooperative decision-making from a limited set of messages. These messages, constrained by partial observability and limited communication resources, only reflect a subset of the environmental states, often proving inadequate for a comprehensive understanding of environmental dynamics. Furthermore, not all received messages are beneficial for decision-making; some may introduce noise that disrupts the process. Consequently, we argue that an ideal representation in this context must be both **sufficient**—offering a comprehensive breadth of information for a deep understanding of the environment, and **informative**—sharply focused on data crucial for facilitating cooperative decision-making.

Following this principle, we introduce M2I2, a novel approach incorporating two self-supervised auxiliary tasks to enhance efficiency of information integration. To meet the standard of "sufficient", M2I2 utilizes masked modeling techniques to reconstruct global states from received messages, furnishing agents with comprehensive information for informed decision-making. Essentially, M2I2 introduces a state-level MAE designed for multi-agent communication. A distinctive aspect of this model is its unique masking mechanism, where the masks are dynamically determined by the communication strategies of the sending agents. Our empirical studies highlight that traditional random mask generating techniques (Devlin et al. 2018; He et al. 2022; Liu et al. 2022) fall short in addressing the complexities encountered in MARL. To this end, we develop the Dimensional Rational Network (DRN) to dynamically adjust the importance of each dimension of observed information. DRN is trained via a meta-learning paradigm, which takes into account the impacts on both decision-making and auxiliary tasks. After exploring the rationale of dimensional observations, we further propose an importance-based heuristic to discern which dimensions of observations should be masked at both training and execution stages, thereby enhancing the efficiency of masked state modeling and communication rationality.

Regarding the “informative" aspect, M2I2 integrates an inverse model to predict joint actions from sequential state representations, enabling agents to focus on information pivotal to their decisions. Furthermore, the inverse model enables agents to infer their teammates' intentions during decentralized decision-making processes. This capability is essential for facilitating team communication that goes beyond mere information exchange, enabling a deeper understanding of teammates' intentions and insights that can impact collective strategies. By introducing the self-supervised objective, M2I2 facilitates a deeper integration of received information, allowing agents to align closely with each other's intentions and leading to more efficient and informed decision-making across various scenarios.

To validate the effectiveness of M2I2, we conduct comprehensive evaluations across a range of multi-agent tasks with differing complexities, from Hallway and MPE to SMAC. Compared to state-of-the-art communication methods(Das et al. 2019; Yuan et al. 2022; Xue et al. 2022; Guan et al. 2022; Duan, Lu, and Xuan 2024), M2I2 demonstrates superior performance, enhanced efficiency, and remarkable generalization capabilities. Our main **contributions** are summarized in three-fold:

- To the best of our knowledge, M2I2 represents the first instance of incorporating self-supervised objectives, i.e., reconstructing global states and predicting joint actions, into the process of information integration under the condition of partial observability and restricted communication resources.
- We integrate a meta-learning paradigm to model the contribution of each dimensional piece of information towards both decision-making and self-supervised objectives, therefore directing agents to transmit and focus on only the most relevant and important information.
- Empirically, our proposed method not only facilitates efficient message integration, but also significantly improves communication efficiency, effectively bridging a vital research gap in MARL.

2 Related Works

Multi-agent communication has emerged as an indispensable component in MARL. Research in this domain has primarily concentrated on three fundamental questions:

Determining the optimal content of communication (what to communicate). CommNet (Sukhbaatar, Szlam, and Fergus 2016), as a pioneering work in this area, facilitated agents in learning continuous messages. Following CommNet, several methods have been developed to further refine the message learning process. VBC (Zhang, Zhang, and Lin 2019) aims to filter out noisy parts while retaining valuable content by limiting the variance of messages. TMC (Zhang, Zhang, and Lin 2020) introduces regularizers to reduce temporally redundant messages. NDQ (Wang et al. 2020) employs information-theoretic regularizers to develop expressive and succinct messages. MAIC (Yuan et al. 2022) enabled agents to customize communications for specific recipients, advancing tailored message learning.

Deciding appropriate timing and partners for information exchange (when and whom to communicate). To enhance communication efficiency, approaches such as IC3Net (Singh, Jain, and Sukhbaatar 2018) and ATOC (Kim et al. 2019) have introduced gating networks to eliminate superfluous communication links. Similarly, SchedNet (Kim et al. 2019), IMMAC (Sun et al. 2021) and T2MAC(Sun et al. 2024) have modeled the significance of observations, using heuristic mechanisms to gate non-essential communication. Further, methods such as MAGIC (Niu, Paleja, and Gombolay 2021), ToM2C(Wang et al. 2022), I2C (Ding, Huang, and Lu 2020) and SMS (Xue et al. 2022) have been developed to identify the most suitable recipients. These approaches focus on modeling the contribution of shared information to the decision-making processes of the recipients, aiming to direct communication where it most influences decision-making.

Integrating incoming messages and making decisions (how to utilize received information). TarMAC (Das et al. 2019) has explored how agents can effectively assimilate crucial information from an abundance of raw messages. MA-

SIA (Guan et al. 2022) take a different approach, employing an Auto-Encoder and a forward model for information integration and becoming the first to introduce self-supervised learning into multi-agent communication. DRMAC(Sun et al. 2025) aims to mitigate both dimensional redundancy and confounders. However, existing methods are under the strong assumption that agents have access to all observations from their peers. In this work, we challenge and relax this assumption by introducing the masked state modeling technique, extending the approach to more realistic environments where communication resources are constrained.

3 Preliminary

In this work, we focus on fully cooperative multi-agent tasks, characterized by partial observability and necessity for inter-agent communication. These tasks are modeled as Decentralized Partially Observable Markov Decision Processes (Dec-POMDPs) (Oliehoek, Amato et al. 2016), represented by the tuple $G = (N, S, O, A, \mathbb{O}, P, R, \gamma, M)$. In this formulation, $N \equiv \{1, \ldots, n\}$ denotes the set of agents, S represents the global states, O describes the observations available to each agent, A signifies the set of available actions, $\mathbb{O}$ is the observation function mapping states to observations, P is the transition function illustrating the dynamics of the environment, R is the reward function dependent on the global states and joint actions of the agents, γ is the discount factor, and M specifies the set of messages that can be communicated among the agents. At each time-step t, each agent $i \in N$ has access to its own observation $o_i^t \in O$ determined by the observation function $\mathbb{O}(o_i^t|s_t)$. Additionally, each agent can receive messages $c_i^t = \sum_{j \neq i} m_j^t$ from teammates $j \in N$. Utilizing both the observed and received information, agents then make local decisions. As each agent selects an action, the joint action a_t results in a shared reward $r_t = R(s_t, a_t)$ and transitions the system to the next state s_{t+1} according to the transition function $P(s_{t+1}|s_t, a_t)$. The objective for all agents is to collaboratively develop a joint policy π to maximize the discounted cumulative return $\sum_{t=0}^{\infty} \gamma^t r_t$.

4 Methodology

4.1 Overall Framework

The framework of M2I2 is shown in Figure 1, with its core components highlighted for effective multi-agent communication. A key component of M2I2 includes a message encoder and a state decoder, functioning collectively as an extendable module to reconstruct the environmental states in an auto-encoding manner. This design allows for the reconstruction of global states from received messages (i.e. limited observations), thereby providing agents with sufficient information to make well-informed cooperative decisions. Furthermore, M2I2 integrates an inverse model capable of predicting joint actions based on consecutive state representations. This model is pivotal in equipping agents with the ability to infer the intentions of their teammates while making decisions. Another standout feature of M2I2 is DRN, which is adept at evaluating the importance of various observed information based on their gradient contributions to both auxiliary and RL tasks. The DRN is continually refined through a meta-learning paradigm, which effectively avoid the trivial solution and local optimum issues during training. By identifying and emphasizing important information, DRN enables agents to share and focus on important data, thereby optimizing the communication process for efficiency and effectiveness.

4.2 Communication Process of M2I2

The communication process of M2I2 can be summarized as the following four steps.

Selectively masking unnecessary observations for information sharing: At each time-step, agents utilize the DRN to evaluate the importance of observed information in supporting decision-making and auxiliary tasks. This importance is quantified as $\omega_i = \omega_{id} | d \in [1, D]$, where i represents the ID of the agent and D is the dimensionality of observed information. To optimize communication efficiency while ensuring effective decision-making, a topK mechanism is applied for generating observation masks, which is formulated as:

$$\text{topK}(\omega_i) = \begin{cases} \omega_{id}, & \text{if } d \text{ in top-k largest dimensions} \\ 0, & \text{others.} \end{cases} \tag{1}$$

This process allows each agent to selectively share the most important dimensions of the observations while non-essential dimensions are masked to zero. The resulting shared information is represented as:

$$m_i^t = o_i^t \otimes \text{topK}(\omega_i), \tag{2}$$

where m_i^t denotes the messages, i.e. masked and weighted observations, and $\otimes$ is an element-wise Hadamard product function. The DRN, central to this selectively observation mask process, is trained using a meta-learning paradigm (detailed in **Section 4.4**), which enables it to dynamically adjust its assessments based on both decision-making and auxiliary task performances.

Integrating received information: Upon receiving messages, M2I2 integrates a scaled dot-product self-attention module (Vaswani et al. 2017) to adeptly process incoming messages. Specifically, the received messages are transformed into corresponding queries Q, keys K, and values V. The process of integrating this information is mathematically represented as follows:

$$z_i^t = f_{\theta_E}(\text{softmax}(\frac{QK^T}{\sqrt{D_k}})V) \tag{3}$$

where θ_E represents the parameters of Message Encoder and D_k represents the dimension of a single key. This message encoder exhibits two notable benefits. Firstly, the encoder's design makes it adaptable to diverse communication contexts, accommodating varying numbers and arrangements of agents. Secondly, by utilizing a weighted sum mechanism, the self-attention module integrates information without excessively expanding the agents' local policy spaces.

Implicitly Inferring the global states and teammates' intention: Following this, M2I2 encodes the received messages into a compact representation. Unlike traditional methods that rely solely on RL objectives, which often struggle to learn effective representation from the limited and noisy messages,

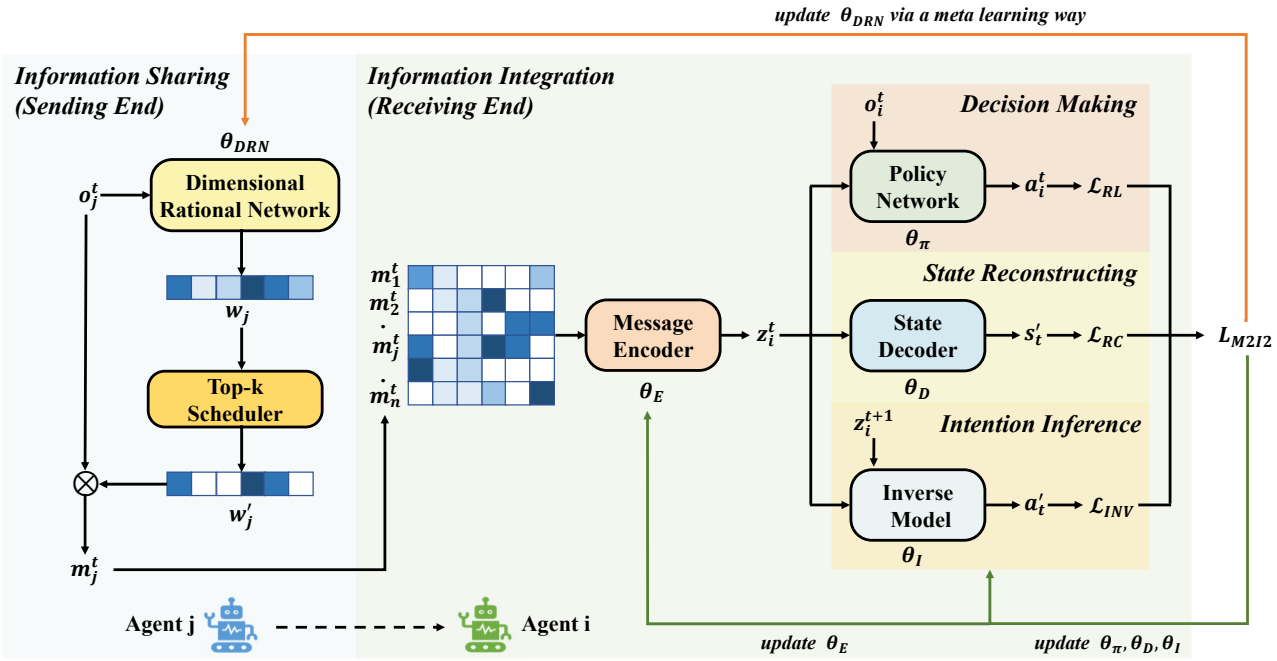

Figure 1: Framework of M2I2. Similar to other CTDE approaches in MARL, M2I2 only leverages global states and joint actions during centralized training phase. However, M2I2 distinguishes itself through its self-supervised auxiliary tasks. These tasks enable agents to develop representations from received messages, enhancing their ability to comprehend global states and infer teammates' joint actions. This capability becomes particularly valuable during the decentralized execution phase, where agents must operate based on limited observations.

M2I2 incorporates two self-supervised objectives. These objectives are specifically designed to develop a representation that is both "sufficient" for a thorough understanding of the environment and "promising" for aiding cooperative decision-making. *Although the involved self-supervised auxiliary tasks are conducted only during training, they can implicitly enhance the message encoder's ability to interpret the environment and predict teammates' intentions during decentralized decision-making process.*

Making cooperative decisions: The culmination of the M2I2 process is reflected in the agents' ability to make cooperative decisions. Here, the enriched integrated information, blended with each agent's personal observations, is channeled into the policy network. The process of decision-making is mathematically represented as follows:

$$a_i^t = \pi_i(o_i^t, z_i^t; \theta_\pi) \tag{4}$$

where θ_π represents the parameters of policy network. This convergence of individual perception and collective insights is crucial, as it empowers agents to make decisions that is not only informed, but also aligned with the overarching goals and strategies of the team.

4.3 Self-Supervised Auxiliary Tasks for Efficient Multi-Agent Communication

Given the inherent constraints in agents' perceptual capabilities and the communication constraints, the information encoded by the message encoder often captures only a fraction of the environment's state. To ensure that agents have access to sufficient information for effective decision-making, we employ a state decoder. This decoder is tasked with reconstructing the global state of the environment, represented by $s_t' = g_{\theta_D}(z_i^t)$. The associated loss function is computed using the mean squared error between the reconstructed and global states:

$$\mathcal{L}_{RC}(\theta_E, \theta_D) = \mathbb{E}_{z_i^t, s_t} \|s_t' - s_t\|_2^2. \tag{5}$$

By combining the message encoder with the state decoder, we effectively create an extendable masked state modeling. This masked modeling is characterized by a unique masking process, generated both by the environment and the agents themselves. This approach enables the integrated representation z_t to effectively represent the global states of the environment, thus overcoming the challenges posed by their limited observational scope and communication capacity.

To augment the capability of agents in focusing on information promising to their decisions and aligning with the intentions of their counterparts, we introduce an inverse model, denoted as $I_{\theta_I} : \mathcal{Z} \times \mathcal{Z} \to A^n$, where $\mathcal{Z}$ is the space of state representations. This model is crafted to predict the joint actions that agents take to transition from one state representation to the next. Formally, given a triplet (z_t, a_t, z_{t+1}) composed of two consecutive state representations and joint actions taken by agents, we parameterise the conditional likelihood as $p(a_t') = I_{\theta_I}(z_t, z_{t+1})$, where I_{θ_I} embodies a two hidden layers MLP followed by a softmax operation. The parameters of both inverse model θ_I and message encoder

θ_E are optimized via a maximum likelihood approach. The corresponding loss function is formulated as:

$$\mathcal{L}_{INV}(\theta_E, \theta_I) = \mathbb{E}_{z_i^t, a_t} \|a'_t - a_t\|_2^2, \tag{6}$$

where this loss function measures the discrepancy between the predicted joint actions and the actual joint actions. At first, the objective encourages agents to focus on information that are controllable and expressive pertinent to cooperative decision-making. This focus is crucial for agents to effectively handle elements that they can influence, enhancing their relevance in a coordinated environment. Moreover, the deeper integration of received information facilitated by the model allows agents to implicitly infer the intentions of others during execution. This capability significantly bolsters agents' potential to align their actions with the intentions of their teammates. Such alignment is not only technically beneficial, but also aligns with cognitive research findings (Etel and Slaughter 2019), which underscore the importance of intention understanding in effective social interactions.

4.4 DRN for Importance Modeling

DRN is designed to discern the importance of different dimensions of observed information, specifically tailoring to the needs of decision-making and auxiliary tasks. The primary challenge here lies in the dynamic nature of the MARL and the variability in communication needs across different stages of a mission. Unlike static scenarios, the importance of information can change dramatically, requiring the DRN to adapt continuously and efficiently. This challenge transcends the realm of simple optimization problems, typically addressed with first-order gradients. To effectively navigate this complexity, we employ a meta-learning approach (Liu, Davison, and Johns 2019; Li et al. 2022a; Qiang et al. 2023; Ji et al. 2024), as it is well-suited to circumvent trivial solutions and local optima that often hinder training efficiency. It allows the DRN to dynamically adjust its understanding of information importance in a sophisticated manner, aligning closely with the overarching goals of decision-making and auxiliary tasks.

It is important to note that within our training framework, only the parameters of θ_{DRN} are refined using this meta-learning approach. The other parameters of the system are updated using conventional first-order gradient methods. Specifically, in the first regular training step, we focus on training the combined set of parameters $\theta = (\theta_E, \theta_D, \theta_I, \theta_\pi)$ by jointly minimizing the auxiliary tasks and RL losses, which is formalized by

$$\arg\min_{\theta} \mathcal{L}_{M2I2}(\theta, \theta_{DRN}), \tag{7}$$

where $\mathcal{L}_{M2I2}(\theta, \theta_{DRN}) = \mathcal{L}_{RL} + \beta(\mathcal{L}_{RC} + \mathcal{L}_{INV})$, $\mathcal{L}_{RL}$ is the RL objective and β is a coefficient that controls the balance between RL objective and auxiliary objectives.

In the second meta-learning-based step, θ_{DRN} is updated by using the second-derivative technique (Liu, Davison, and Johns 2019; Li et al. 2022b). This technique is crucial for adjusting θ_{DRN} to better discern the importance of various information dimensions that significantly impact both RL and auxiliary tasks. The update process involves calculating the gradients of θ_{DRN} in relation to the combined performance metrics from these tasks, encapsulated by $\mathcal{L}_{M2I2}$. Formally, we update θ_{DRN} by

$$\arg\min_{\theta_{DRN}} \mathcal{L}_{M2I2}(\theta_{trial}, \theta_{DRN}), \tag{8}$$

where $\theta_{trial} = (\theta_E^{trial}, \theta_D^{trial}, \theta_I^{trial}, \theta_\pi^{trial})$ is the trial weights of the θ after one gradient update using the M2I2 loss defined in Equation 7. We formulate the updating of such trial weights as follows:

$$\theta_{trial} = \theta - \ell_\theta \nabla_\theta \mathcal{L}_{M2I2}, \tag{9}$$

where ℓ_θ is the learning rate. Note that the calculation of trial weights excludes the step of gradient back-propagation. Thus, θ_{DRN} is updated through the second-derivative gradient of θ. During the trial step, DRN evaluates how changes in parameters affect task performance, capturing this impact through gradients. This allows DRN to learn meta-knowledge about how different dimensions contribute to outcomes, enabling it to adjust importance weights effectively over time.

Overall, DRN models the gradient contribution of each observation dimension to $\mathcal{L}_{M2I2}$. Dimensions contributing more receive higher weights. Unlike traditional attention mechanisms(Waswani et al. 2017) in Transformers, which capture data similarity, DRN's importance modeling is directly tied to the objective function and guided by second-order gradients. This design ensures that θ_{DRN} is continuously fine-tuned by the gradient contributions of $\mathcal{L}_{M2I2}$, allowing DRN to dynamically evaluate the importance of each observed dimension. Hence, M2I2 ensures priority communication of information most beneficial for the receiving agents' tasks, including state reconstruction, intention inference, and decision-making. Our visualization study in **Appendix E** further validates this approach: it reveals that the DRN not only distinguishes varying levels of importance across different agent types and observation categories but also dynamically adjusts these importance weights over time, adapting to the evolving demands of the task.

5 Experiment

In this section, our experimental design is meticulously structured to address three fundamental questions:

- **RQ1.** How does M2I2's performance and efficiency compare to leading communication methods?
- **RQ2.** What specific components within M2I2 are instrumental to its performance?
- **RQ3.** Is M2I2 versatile enough to be applied across a range of tasks, and can it be seamlessly integrated with multiple existing baselines?

5.1 Setup

Benchmarks. In order to demonstrate the effectiveness and generality of M2I2, we conducted extensive experiments across four popular multi-agent communication benchmarks: Hallway (Wang et al. 2020), Predator-Prey (PP) (Lowe et al. 2017), SMAC (Samvelyan et al. 2019) and SMAC-Communication (Wang et al. 2020). Each of these benchmarks provides a substantial testbed for evaluating multi-agent communication strategies. Detailed descriptions of each environment can be found in **Appendix C**.

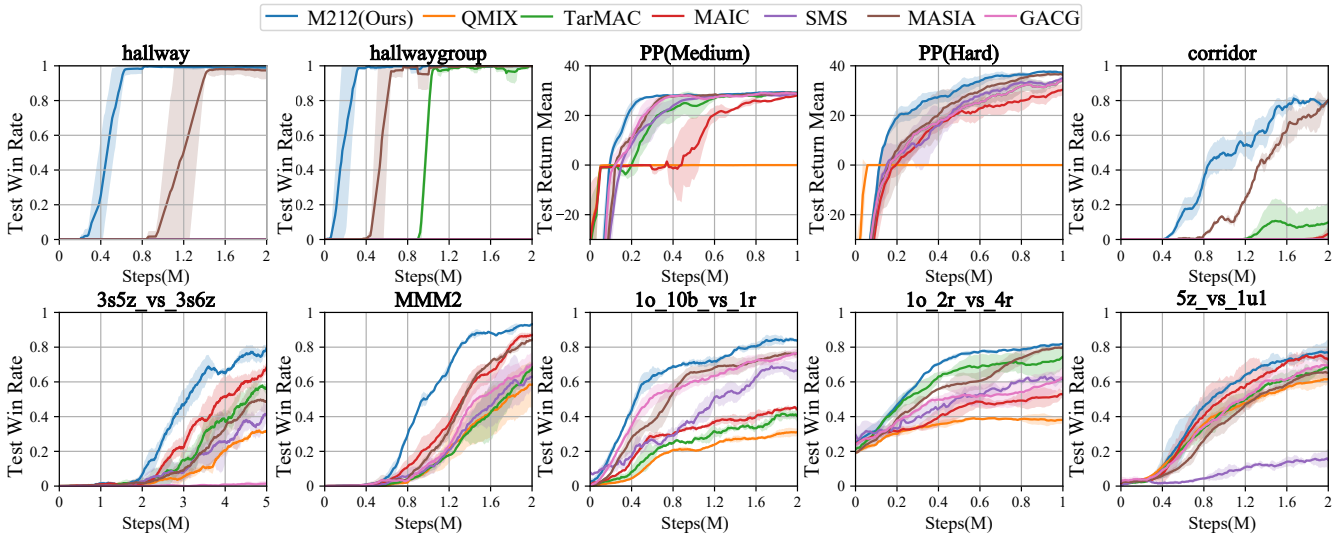

Figure 2: Performance on multiple benchmarks. All results are reported as the median performance with a 95% confidence interval over five random seeds.

Baselines. For comparative analysis, we select a diverse set of baselines. This includes QMIX (Rashid et al. 2018), a well-established MARL algorithm that operates without a communication mechanism. To assess our method's performance in the context of communication-enhanced MARL, we also include contemporary state-of-the-art communication methods: TarMAC (Das et al. 2019), MAIC (Yuan et al. 2022), SMS (Xue et al. 2022), MASIA (Guan et al. 2022) and GACG (Duan, Lu, and Xuan 2024). Each of these baselines represents a significant stride in the development of communication strategies within the MARL framework, providing a robust backdrop for evaluating the efficacy and innovation of our proposed approach.

Hyperparameters. To ensure reproducibility, the intricate details of our method's architecture, and our hyperparameter choices are extensively detailed in the **Appendix C**.

5.2 Performance (RQ1)

Our evaluation begins with a comparative analysis of the learning curves of M2I2 against a range of baseline methods across diverse environments. This comparison is aimed at assessing the comprehensive performance of M2I2. For a fair comparison, we evaluate M2I2 under the QMIX-based implementation, aligning with the setup used for most of the selected baselines (i.e., TarMAC, MAIC, and MASIA). As depicted in Figure 2, M2I2 demonstrates a notable performance advantage, consistently outperforming all baselines by a significant margin in each tested environment, indicating M2I2's strong applicability across scenarios of varying complexity. Specifically, in Hallway, where the reward signals of environment is sparse, many methods exhibit poor performance or fail to learn effectively. In contrast, M2I2 rapidly achieves a 100% win rate. This success can be attributed to our proposed auxiliary tasks, which appear to significantly aid agents in understanding the locations and intentions of their teammates. In PP, where the communication-free method QMIX struggles, most communication methods demonstrate effectiveness. Notably, M2I2 achieves the best sample efficiency, swiftly identifying the optimal policy. In SMAC and SMAC-Communication, several full communication methods face challenges in scenarios with large joint observation spaces. For instance, TarMAC shows competence in $1o\_2r\_vs\_4r$ and $5z\_vs\_1ul$ but underperforms in $corridor$ and $1o\_10b\_vs\_1r$. This could be attributed to the naive approach of these methods in feeding observations from all agents directly into the policy network, thereby increasing the complexity of policy learning. In contract, M2I2 consistently excels in all six SMAC scenarios, irrespective of the varying difficulties, number of agents, and terrain types. This further underscores the broad applicability and potency of M2I2 in diverse multi-agent communication contexts.

Communication Efficiency	Hallway	PP	SMAC	SMAC-Communication
TarMAC	49.6%	32.32	19.1%	17.4%
MAIC	0.0%	29.75	27.1%	10.1%
SMS	0.0%	51.63	13.7%	41.8%
MASIA	98.6%	32.52	46.5%	28.5%
M2I2(Ours)	**165.6%**	**55.76**	**98.7%**	**59.3%**

Table 1: Communication efficiency, defined as performance improvement with communication normalized by communication frequency.

5.3 Efficiency (RQ1)

Efficiency is a long-standing issue in multi-agent communication, as many real-world applications operate under limited communication resources. Therefore, it is crucial to achieve promising performance while maintaining a low communication resource cost. Notably, the performance of M2I2, as

reported in Figure 2, was achieved with a 60% communication frequency, where 60% is a hyper parameter defined by our proposed top-k mechanism in **Section 4.1**. To further understand the communication efficiency of M2I2, we adopted a mechanism inspired by MAGIC (Niu, Paleja, and Gombolay 2021) to measure communication efficiency. Specifically, we calculated the performance improvement for each communication algorithm by subtracting the baseline performance of their communication-free versions. For the SMS algorithm, the communication-free baseline used was DOP (Wang et al. 2020), while for other algorithms, QMIX served as the baseline. Subsequently, we examined the communication frequency for each method. M2I2 and SMS both operated at approximately 60% communication frequency, in contrast to other methods which utilized 100% communication frequency. Finally, we calculated the communication efficiency for each method by dividing the performance improvement by the communication frequency. As indicated in Table 1, M2I2 demonstrated a substantial lead in communication efficiency across all tested scenarios, further validating its effectiveness. This analysis not only underscores M2I2's ability to maintain high performance with reduced communication demands but also highlights its significant advantages in terms of resource efficiency.

5.4 Ablation (RQ2)

To assess the contribution of each module in M2I2, we perform an ablation study across three SMAC-Communication scenarios. Specifically, to isolate the role of the two auxiliary tasks introduced in M2I2, we evaluate two ablated variants: one without the MAE module and another without the inverse model. As illustrated in Figure 3, removing the MAE results in a noticeable performance drop, indicating that the MAE-based masked state modeling is a fundamental component of our approach. This outcome is expected, as the DRN module was specifically introduced to enhance the masked state modeling process. Moreover, the inverse model typically requires two consecutive states as input to predict the intermediate action. However, under the CTDE paradigm, global state information is not available during execution. To address this limitation, our state modeling mechanism constructs latent state representations from the agents' received messages, thereby effectively capturing global information during execution. Then, our inverse model leverages two representations obtained through state modeling training as input. Removing masked state modeling would render these latent states meaningless for training, ultimately degrading overall performance. Finally, to evaluate the contribution of the DRN itself, we replace it with a naive random masking strategy. We observe that this substitution causes performance to drop in all three scenarios, confirming the importance of the DRN component. Additionally, we provide fine-grained ablation studies in the **Appendix D**, such as an analysis of the impact of meta-learning on DRN.

Furthermore, to gain insight into how varying communication frequencies impact M2I2's performance, we executed an ablation study with communication rates set at 0.8, 0.6, 0.4. The findings, depicted in Figure 4, consistently show the best performance at a communication rate of 0.6. This result

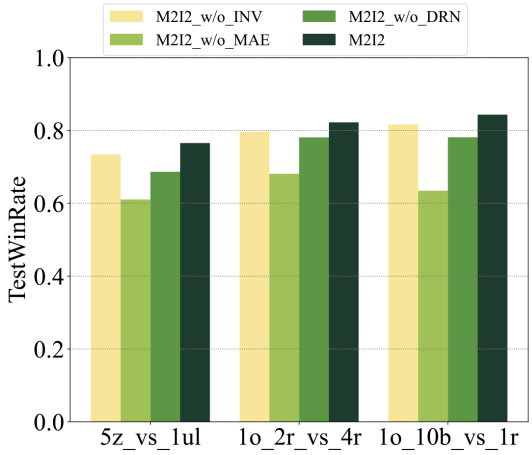

Figure 3: Ablation for modules with in M2I2.

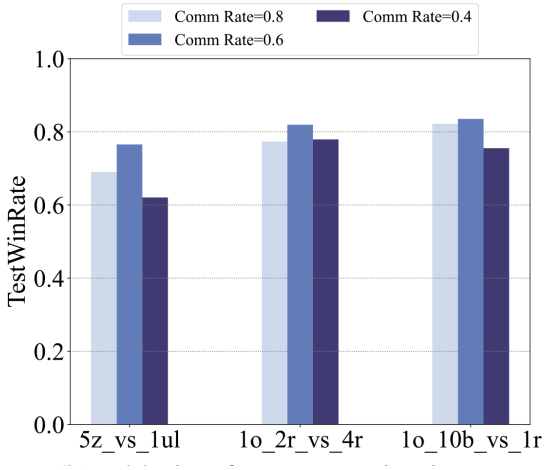

Figure 4: Ablation for communication rate.

suggests that an excessively high communication rate can introduce redundant and misleading information, whereas too low a rate may lead to critical information being overlooked. Intriguingly, reducing information by 0.6 using DRN still outperforms a random mask reduction of 0.4, underscoring the DRN's proficiency in discerning and prioritizing key information. This delicate balance achieved by M2I2, efficiently filtering out non-essential data while preserving crucial information, significantly enhances the overall communication efficiency in complex multi-agent environments.

5.5 Generalization (RQ3)

Our previous experiments have conclusively shown M2I2's robust performance in a variety of environments, encompassing scenarios with diverse complexities and scales. Building on this, we extend our evaluation of M2I2 to assess its generality across various MARL baselines, including QMIX, VDN, QPLEX, MAPPO and MADDPG. We present the test win rates for the scenario $1o\_2r\_vs\_4r$ in Figure 5. Remarkably, M2I2 demonstrates consistently superior performance across all these baselines, often achieving a significant margin of

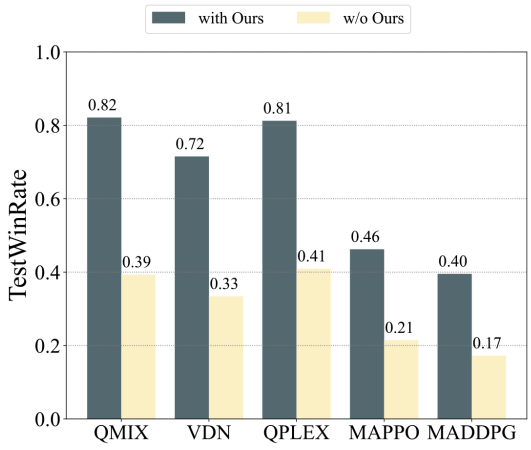

Figure 5: Generalization for multiple MARL baseline.

improvement. This observation shows that M2I2 is effective not only with off-policy algorithms but also with on-policy approaches, and not only with Q-learning methods but also with policy gradient methods, demonstrating M2I2's broad applicability and effectiveness within the MARL domain.

6 Conclusion

In this work, we delve into the complexities of multi-agent information integration. We introduce M2I2, an approach that incorporates two auxiliary tasks to enhance communication efficiency. We specifically design a MAE and an inverse model. These elements play a crucial role in guiding the processes of information integration, thereby significantly enhancing the agents' ability to navigate uncertain environments and dynamically adapt to their teammates. To substantiate our claims, we conduct exhaustive experiments across a multitude of benchmarks. The results from these tests not only validate the effectiveness of M2I2 but also its efficiency and adaptability in various multi-agent scenarios. These findings highlight M2I2's potential to significantly advance the field of multi-agent communication. In future work, we plan to delve deeper into which observations are being masked by M2I2 and what kind of representations the model is learning. This enhanced understanding will help further refine the model's application and effectiveness.

Acknowledgements

This work is supported by the National Natural Science Foundation of China No. 62406313, 2023 Special Research Assistant Grant Project of the Chinese Academy of Sciences.

References

Andrychowicz, O. M.; Baker, B.; Chociej, M.; Jozefowicz, R.; McGrew, B.; Pachocki, J.; Petron, A.; Plappert, M.; Powell, G.; Ray, A.; et al. 2020. Learning dexterous in-hand manipulation. *The International Journal of Robotics Research*, 39(1): 3–20.

Chen, Y.; Mao, H.; Mao, J.; Wu, S.; Zhang, T.; Zhang, B.; Yang, W.; and Chang, H. 2022. PTDE: Personalized training with distilled execution for multi-agent reinforcement learning. *arXiv preprint arXiv:2210.08872*.

Chen, Y.; Yan, L.; Sun, W.; Ma, X.; Zhang, Y.; Wang, S.; Yin, D.; Yang, Y.; and Mao, J. 2025. Improving retrieval-augmented generation through multi-agent reinforcement learning. *arXiv preprint arXiv:2501.15228*.

Das, A.; Gervet, T.; Romoff, J.; Batra, D.; Parikh, D.; Rabbat, M.; and Pineau, J. 2019. Tarmac: Targeted multi-agent communication. In *International Conference on Machine Learning*, 1538–1546.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. N. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

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.

Duan, W.; Lu, J.; and Xuan, J. 2024. Group-Aware Coordination Graph for Multi-Agent Reinforcement Learning. In Larson, K., ed., *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, 3926–3934. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Etel, E.; and Slaughter, V. 2019. Theory of mind and peer cooperation in two play contexts. *Journal of Applied Developmental Psychology*, 60: 87–95.

Gu, Z.; Wang, J.; Zuo, R.; Sun, C.; Song, Z.; Zheng, C.; and Qiang, W. 2025. Group Causal Policy Optimization for Post-Training Large Language Models. *arXiv preprint arXiv:2508.05428*.

Guan, C.; Chen, F.; Yuan, L.; Wang, C.; Yin, H.; Zhang, Z.; and Yu, Y. 2022. Efficient Multi-agent Communication via Self-supervised Information Aggregation. *Advances in Neural Information Processing Systems*, 35: 1020–1033.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009.

Hu, G.; Zhu, Y.; Zhao, D.; Zhao, M.; and Hao, J. 2021. Event-triggered communication network with limited-bandwidth constraint for multi-agent reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems*, 34(8): 3966–3978.

Ji, Q.; Li, J.; Hu, J.; Wang, R.; Zheng, C.; and Xu, F. 2024. Rethinking Dimensional Rationale in Graph Contrastive Learning from Causal Perspective. In Wooldridge, M. J.; Dy, J. G.; and Natarajan, S., eds., *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, 12810–12820. AAAI Press.

Kim, D.; Moon, S.; Hostallero, D.; Kang, W. J.; Lee, T.; Son, K.; and Yi, Y. 2019. Learning to schedule communication in multi-agent reinforcement learning. *arXiv preprint arXiv:1902.01554*.

Leurent, E. 2018. A survey of state-action representations for autonomous driving.

Li, J.; Qiang, W.; Zheng, C.; Su, B.; and Xiong, H. 2022a. MetAug: Contrastive Learning via Meta Feature Augmentation. In Chaudhuri, K.; Jegelka, S.; Song, L.; Szepesvári, C.; Niu, G.; and Sabato, S., eds., *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, 12964–12978. PMLR.

Li, J.; Qiang, W.; Zheng, C.; Su, B.; and Xiong, H. 2022b. MetAug: Contrastive Learning via Meta Feature Augmentation. In Chaudhuri, K.; Jegelka, S.; Song, L.; Szepesvári, C.; Niu, G.; and Sabato, S., eds., *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, 12964–12978. PMLR.

Liu, F.; Liu, H.; Grover, A.; and Abbeel, P. 2022. Masked autoencoding for scalable and generalizable decision making. *Advances in Neural Information Processing Systems*, 35: 12608–12618.

Liu, S.; Davison, A.; and Johns, E. 2019. Self-supervised generalisation with meta auxiliary learning. *Advances in Neural Information Processing Systems*, 32.

Lowe, R.; Wu, Y. I.; Tamar, A.; Harb, J.; Abbeel, O. P.; and Mordatch, I. 2017. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in neural information processing systems*, 6379–6390.

Niu, Y.; Paleja, R. R.; and Gombolay, M. C. 2021. Multi-Agent Graph-Attention Communication and Teaming. In *AAMAS*, 964–973.

Oliehoek, F. A.; Amato, C.; et al. 2016. *A concise introduction to decentralized POMDPs*, volume 1. Springer.

Osband, I.; Blundell, C.; Pritzel, A.; and Van Roy, B. 2016. Deep exploration via bootstrapped DQN. In *Advances in neural information processing systems*, 4026–4034.

Qiang, W.; Li, J.; Su, B.; Fu, J.; Xiong, H.; and Wen, J. 2023. Meta Attention-Generation Network for Cross-Granularity Few-Shot Learning. *Int. J. Comput. Vis.*, 131(5): 1211–1233.

Rashid, T.; Samvelyan, M.; De Witt, C. S.; Farquhar, G.; Foerster, J.; and Whiteson, S. 2018. QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning. *arXiv preprint arXiv:1803.11485*.

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*.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Zhang, M.; Li, Y. K.; Wu, Y.; and Guo, D. 2024. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *CoRR*, abs/2402.03300.

Silver, D.; Hubert, T.; Schrittwieser, J.; Antonoglou, I.; Lai, M.; Guez, A.; Lanctot, M.; Sifre, L.; Kumaran, D.; Graepel, T.; et al. 2018. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, 362(6419): 1140–1144.

Silver, D.; Schrittwieser, J.; Simonyan, K.; Antonoglou, I.; Huang, A.; Guez, A.; Hubert, T.; Baker, L.; Lai, M.; Bolton, A.; et al. 2017. Mastering the game of go without human knowledge. *nature*, 550(7676): 354–359.

Singh, A.; Jain, T.; and Sukhbaatar, S. 2018. Learning when to communicate at scale in multiagent cooperative and competitive tasks. *arXiv preprint arXiv:1812.09755*.

Sukhbaatar, S.; Szlam, A.; and Fergus, R. 2016. Learning Multiagent Communication with Backpropagation. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, 2252–2260. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781510838819.

Sun, C.; He, P.; Wang, R.; and Zheng, C. 2025. Revisiting Communication Efficiency in Multi-Agent Reinforcement Learning from the Dimensional Analysis Perspective. In Das, S.; Nowé, A.; and Vorobeychik, Y., eds., *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2025, Detroit, MI, USA, May 19-23, 2025*, 1977–1986. International Foundation for Autonomous Agents and Multiagent Systems / ACM.

Sun, C.; Wu, B.; Wang, R.; Hu, X.; Yang, X.; and Cong, C. 2021. Intrinsic Motivated Multi-Agent Communication. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '21, 1668–1670. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450383073.

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.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All You Need.

Vinyals, O.; Babuschkin, I.; Czarnecki, W. M.; Mathieu, M.; Dudzik, A.; Chung, J.; Choi, D. H.; Powell, R.; Ewalds, T.; Georgiev, P.; et al. 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575(7782): 350–354.

Wang, J.; Qiang, W.; Song, Z.; Zheng, C.; and Xiong, H. 2025. Learning to Think: Information-Theoretic Reinforcement Fine-Tuning for LLMs. *CoRR*, abs/2505.10425.

Wang, T.; Wang, J.; Zheng, C.; and Zhang, C. 2020. Learning Nearly Decomposable Value Functions Via Communication Minimization. In *ICLR 2020 : Eighth International Conference on Learning Representations*.

Wang, Y.; Han, B.; Wang, T.; Dong, H.; and Zhang, C. 2020. Dop: Off-policy multi-agent decomposed policy gradients. In *International Conference on Learning Representations*.

Wang, Y.; Zhong, F.; Xu, J.; and Wang, Y. 2022. ToM2C: Target-oriented Multi-agent Communication and Cooperation with Theory of Mind. In *International Conference on Learning Representations*.

Waswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. In *NIPS*.

Xue, D.; Yuan, L.; Zhang, Z.; and Yu, Y. 2022. Efficient Multi-Agent Communication via Shapley Message Value. In Raedt, L. D., ed., *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, 578–584. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Yu, C.; Velu, A.; Vinitsky, E.; Gao, J.; Wang, Y.; Bayen, A.; and Wu, Y. 2022. The surprising effectiveness of ppo in cooperative multi-agent games. *Advances in Neural Information Processing Systems*, 35: 24611–24624.

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.

Zang, Z.; Li, J.; Sun, C.; Wang, R.; Liu, L.; and Sun, F. 2026. Visual reinforcement learning via sequential consistency preserved policy contrast from optimal transport view. *Neural Networks*, 193: 108019.

Zang, Z.; Sun, C.; Liu, L.; Sun, F.; and Zheng, C. 2025. Loss of Plasticity: A New Perspective on Solving Multi-Agent Exploration for Sparse Reward Tasks. In Das, S.; Nowé, A.; and Vorobeychik, Y., eds., *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2025, Detroit, MI, USA, May 19-23, 2025*, 2299–2308. International Foundation for Autonomous Agents and Multiagent Systems / ACM.

Zhang, S. Q.; Zhang, Q.; and Lin, J. 2019. Efficient communication in multi-agent reinforcement learning via variance based control. In *Advances in Neural Information Processing Systems*, 3235–3244.

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.