T2MAC: Targeted and Trusted Multi-Agent Communication through Selective Engagement and Evidence-Driven Integration

Chuxiong Sun[1 2*], Zehua Zang[1 3*], Jiabao Li[4*], Jiangmeng Li[1 2†], Xiao Xu[2], Rui Wang[1 2 3], Changwen Zheng[1 3]

[1]Science & Technology on Integrated Information System Laboratory, Institute of Software Chinese Academy of Sciences
[2]State Key Laboratory of Intelligent Game
[3]University of Chinese Academy of Sciences
[4]School of Automation and Electrical Engineering, University of Science and Technology Beijing
{chuxiong2016,zehua2020}@iscas.ac.cn, M202110548@xs.ustb.edu.cn, jiangmeng2019@iscas.ac.cn,
xuxiao0825@gmail.com, {wangrui,changwen}@iscas.ac.cn,

Abstract

Communication stands as a potent mechanism to harmonize the behaviors of multiple agents. However, existing works primarily concentrate on broadcast communication, which not only lacks practicality, but also leads to information redundancy. This surplus, one-fits-all information could adversely impact the communication efficiency. Furthermore, existing works often resort to basic mechanisms to integrate observed and received information, impairing the learning process. To tackle these difficulties, we propose Targeted and Trusted Multi-Agent Communication (T2MAC), a straightforward yet effective method that enables agents to learn selective engagement and evidence-driven integration. With T2MAC, agents have the capability to craft individualized messages, pinpoint ideal communication windows, and engage with reliable partners, thereby refining communication efficiency. Following the reception of messages, the agents integrate information observed and received from different sources at an evidence level. This process enables agents to collectively use evidence garnered from multiple perspectives, fostering trusted and cooperative behaviors. We evaluate our method on a diverse set of cooperative multi-agent tasks, with varying difficulties, involving different scales and ranging from Hallway, MPE to SMAC. The experiments indicate that the proposed model not only surpasses the state-of-the-art methods in terms of cooperative performance and communication efficiency, but also exhibits impressive generalization.

1 Introduction

Reinforcement Learning (RL) has achieved remarkable milestones in a myriad of intricate real-world domains, ranging from Game AI (Osband et al. 2016; Silver et al. 2017, 2018; Vinyals et al. 2019) and Robotics (Andrychowicz et al. 2020) to Autonomous Driving (Leurent 2018). However, when delving into cooperative multi-agent settings, distinct challenges surface. The issue of partial observability stands out, where agents are confined to their local observations, missing out on the broader perspective of the entire environment. Complicating matters further, Multi-Agent Reinforcement Learning (MARL) grapples with the non-stationarity of the environment. From an individual agent's perspective, the environmental dynamics shift incessantly, adding another layer of complexity to the learning process.

Multi-agent communication offers a compelling solution to the issues outlined by granting agents the capability to derive a deeper understanding of their surroundings through collective insights. This approach ensures stable learning and encourages harmonized actions among agents. However, historical methods have focused on the content and timing of communication (Sukhbaatar, Szlam, and Fergus 2016; Singh, Jain, and Sukhbaatar 2018; Kim et al. 2019; Wang et al. 2020b; Zhang, Zhang, and Lin 2020; Yuan et al. 2022). Once an agent elects to share its message, it is broadcast to the entire agent group. This indiscriminate broadcasting is not only resource-intensive but also potentially inefficient. A pivotal realization is that only some agents carry valuable insights, and flooding the network with redundant information can be counterproductive to learning. Interestingly, humans know when to communicate intrinsically, with whom, and how to customize their messages to the recipient. Mirroring these human instincts could significantly refine the information exchange process, allowing agents to curate their messages and recipients selectively.

Moreover, the essence of messages relayed by agents is a distillation of their individual observational experiences. Assimilating these messages aptly can enrich agents' perception of an uncertain environment, leading to more refined policies. Regrettably, the existing techniques—whether they're steeped in basic aggregation (Jiang and Lu 2018) or are more avant-garde with representation learning (Das et al. 2019; Guan et al. 2022)—tend to treat the fusion of information as a black box, presupposing that the policy networks can innately sift out vital data and diminish decision-making uncertainty. In this context, the information integration process might prove to be both uncertain and inefficient, especially in intricate scenarios. As such, there's a pressing need for a novel and theory-grounded approach that can adeptly merge messages while tackling the inherent

*These authors contributed equally.
†Corresponding author.

underlying uncertainties.

With this vision, we introduce the Targeted and Trusted Multi-Agent Communication (T2MAC) framework, which embodies the principles of discerning and streamlined communication, drawing inspiration from human inclinations to engage selectively with trusted and relevant counterparts, ensuring more efficient information integration, and fostering a more adaptive multi-agent collaboration in dynamic environments. Specifically, each agent is skilled at analyzing observations to extract evidence. In this context, evidence denotes metrics instrumental in guiding the decision-making processes. This evidence plays a dual role: guiding local decision-making and serving as the basis for crafting messages that are meticulously tailored to specific agent contexts. Moreover, we evaluate the variations in uncertainty prior to and post-communication to measure the impact and significance of specific communication behavior. Armed with these insights, we craft binary pseudo-labels based on the significance of communication and devise an auxiliary task. This task is specifically designed to train a communication selector network, empowering it to identify the ideal communication counterparts. By adopting this strategy, we guarantee that only the most relevant and credible data is exchanged among the agents. Upon receipt, messages are integrated at the evidence level rather than the conventional observation or feature level. To capture the intricacies of decision-making, we leverage the Dirichlet distribution. This allows us to model decision policies, anchoring them on evidence that's been sourced from a myriad of perspectives. Concretely, we integrate Subjective Logic (SL) (Jsang 2018) to link the Dirichlet parameters with belief and uncertainty, therefore quantifying the uncertainty for decision-making and jointly modeling the probability of each action. Then, we utilize Dempster-Shafer theory of evidence (DST) (Dempster 1967) to integrate evidence observed from multiple agents, producing a comprehensive belief and uncertainty that considers all available evidence, ensuring trusted message integration and decision-making. We subjected T2MAC to rigorous testing across various MARL environments, such as Hallway, MPE, and SMAC. Compared to prominent multi-agent communication strategies like TarMac (Das et al. 2019), MAIC (Yuan et al. 2022), SMS (Xue et al. 2022), and MASIA (Guan et al. 2022), T2MAC consistently excelled in both performance and efficiency. Additionally, its versatility shone through across diverse scenarios.

2 Related Works

MARL has undergone remarkable progression in recent epochs (Lowe et al. 2017; Sunehag et al. 2017; Rashid et al. 2018; Yu et al. 2022). Within the MARL ambit, multi-agent communication has emerged as an indispensable aspect, particularly salient for cooperative endeavors constrained by partial observability. Research in this domain can be broadly segmented into three main categories.

Deciding What to Communicate. Historically, communication vocabularies are set in stone during training, as illustrated by (Foerster et al. 2016). This seemingly efficient strategy unintentionally limits the depth and flexibility of agent communication. In response, CommNet (Sukhbaatar, Szlam, and Fergus 2016) introduces a paradigm shift by allowing agents to create dynamic, continuous messages. With its design for continuous interactions, CommNet ensures that messages are timely and sensitive to environmental changes. Building on this foundation, both VBC (Zhang, Zhang, and Lin 2019) and TMC (Zhang, Zhang, and Lin 2020) further optimize message learning processes. Furthermore, NDQ (Wang et al. 2020b) and MAIC (Yuan et al. 2022) are designed to craft messages tailored for individual agents.

Deciding When and With Whom to Communicate. Effective communication timing and partner selection are pivotal in Multi-Agent Communication. A gating network showcased in (Singh, Jain, and Sukhbaatar 2018; Jiang and Lu 2018) generates binary decisions, allowing agents the freedom to communicate or abstain. Advancing this idea, (Kim et al. 2019; Mao et al. 2019; Wang et al. 2020a; Sun et al. 2021) implement a weight-based scheduler, prioritizing agents holding vital observations. Enriching this approach, I2C (Ding, Huang, and Lu 2020), MAGIC (Niu, Paleja, and Gombolay 2021), and SMS (Xue et al. 2022) harness methods like causal inference, graph-attention, and Shapley message value to pinpoint ideal communication recipients.

Incorporating Messages for Cooperative Decision-Making. A prominent subset of the MARL methodologies posits an egalitarian weightage to all incoming messages. Such an approach fails to recognize the significance of filtering vital information from a sea of communications. Therefore, we introduce representation learning paradigms to address this lacuna for discerning message assimilation. For instance, TarMac (Das et al. 2019) adopts soft attention mechanisms to weight messages, while MASIA (Guan et al. 2022) consolidates received messages into concise representations using an autoencoder.

To our knowledge, no existing MARL method simultaneously addresses targeted and trusted communication. T2MAC stands as the pioneering approach, enabling agents to efficiently select communication partners and distill tailored evidence and integrate messages, resulting in trustworthy cooperative decisions.

3 Background

In this study, we concentrate on fully cooperative multi-agent reinforcement learning tasks characterized by partial observability while also allowing inter-agent communication. These tasks are an evolved form of Decentralized Partially Observable Markov Decision Processes (Dec-POMDPs). Their framework uses the tuple $G = (N, S, O, A, \mathbb{O}, P, R, \gamma, M)$. In this formulation: $N = (agent_1, ..., agent_n)$ depicts the collective of agents. S encompasses global states, offering a comprehensive environmental overview. O refers to the accessible local observations. A signifies a set of available actions. $\mathbb{O}$ refers to the observation function, which describes how agents perceive the environment based on the global state. P acts as the transition function, illustrating environmental dynamics. R is a reward function contingent on global states and joint actions.

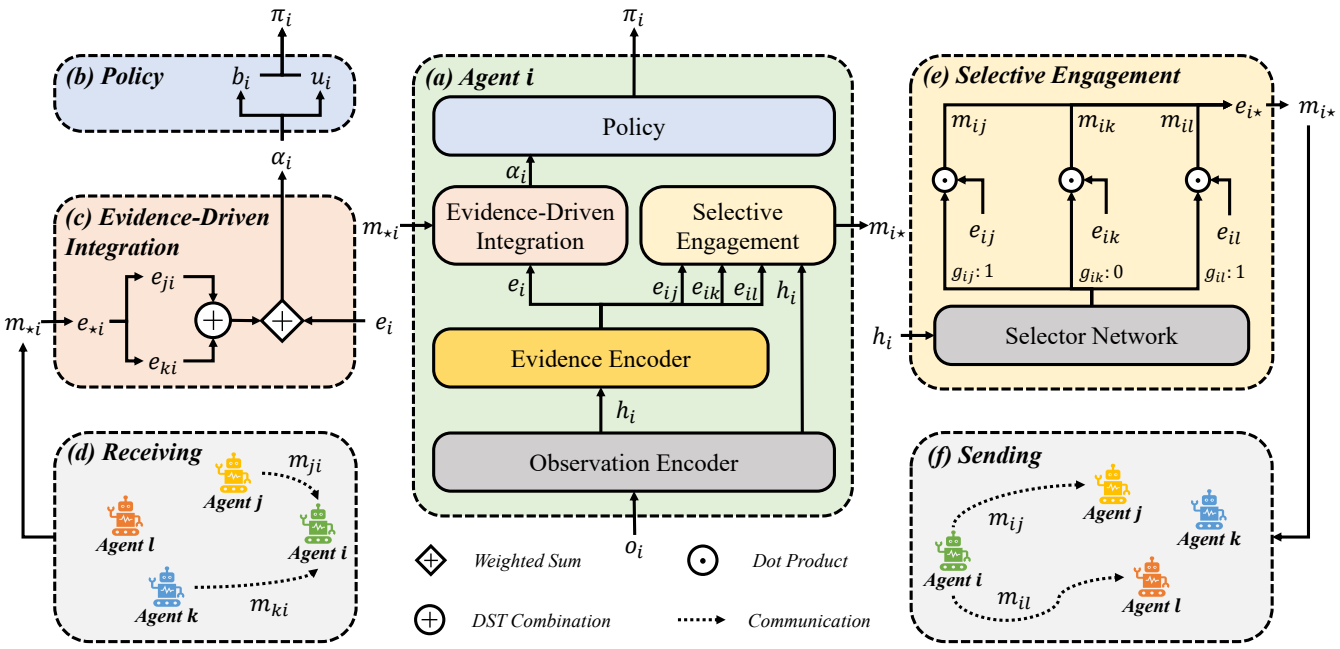


Figure 1: Framework of T2MAC.

γ represents the discount factor, M delineates the set of communicable messages.

At each time step, agents access only local observations, which are derived from the global state through the observation function, $\mathbb{O}(o_i^t|s)$. Simultaneously, agents are equipped with the capability to share messages, denoted as m_i^t. These messages might encompass observations, intentions, or past experiences. Crucially, each agent can judiciously decide when to communicate, streamlining the efficiency of the communication process. As messages are received, agents integrate their incoming information, leading to the aggregated message $c_i^t = \sum_{j \neq i} m_j^t$. This composite data then guides their localized decision-making, encapsulated by $a_i^t = \pi(o_i^t, c_i^t)$. Following this, the environment reacts to the joint action, $a = (a_1^t, ..., a_n^t)$, transitioning to the subsequent state s'. Simultaneously, this joint action then yields a shared team reward, $r = R(s, a)$. The overarching goal is to pinpoint an optimal joint policy geared towards maximizing the expected cumulative team reward, expressed as $\mathbb{E}_{s,a}[\sum_{t=0}^{\infty} \gamma^t r]$.

4 Methodology

As depicted in Fig. 1, the distinctive characteristics of T2MAC can be highlighted in these four aspects:

- T2MAC's policy is characterized as a Dirichlet distribution, facilitating the assimilation of evidence from various sources for informed and trusted decisions.
- The evidence encoder serves a dual purpose: extracting evidence for its own decisions and crafting tailored messages for specific teammates.
- Through the selective engagement, T2MAC can pinpoint optimal moments and counterparts for communication, ensuring the dissemination of only the most pertinent and reliable data.
- The evidence-driven integration combines incoming messages at an evidence level, refraining from treating the fusion process as a black box.

In the following sections, we will illustrate the key components of T2MAC in detail.

4.1 Theory of Evidence

For communication to be precise and reliable, it's essential to factor in the uncertainties intrinsic to individual decisions. To this effect, we have incorporated the theory of evidence into multi-agent communication. Within this context, evidence pertains to metrics sourced from observations supporting decision-making processes. To get a grasp on this evidence and uncertainty, we employ the Dirichlet distribution, which has proven efficacious in mitigating the overconfidence issue (Sensoy, Kaplan, and Kandemir 2018; Malinin, Mlodozeniec, and Gales 2020; Malinin and Gales 2018). This distribution is characterized by its concentration parameters, represented as $\alpha = [\alpha^1, ..., \alpha^K]$ where K is the number of actions. These parameters share an intimate relationship with uncertainty. Building on this, we harness SL to discern the concentration parameters. SL offers a theoretical framework for extracting the probabilities (belief masses) of disparate actions and the overarching uncertainty (uncertainty mass) tied to policy-prediction challenges. Delving deeper into decision-making quandaries, SL seeks to allocate a belief mass to each action while assigning an overarching uncertainty mass to the entire scenario based on observed evidence. Consequently, all mass values remain non-negative

and their cumulative value equals one:

$$u_i + \sum_{k=1}^{K} b_i^k = 1 \tag{1}$$

where $u_i \geq 0$ signifies the overall uncertainty for $agent_i$, $b_i^k \geq 0$ denotes the belief of $agent_i$ associated with the k^{th} action.

Moreover, SL elegantly bridges the evidence observed by $agent_i$, denoted as $e_i = [e_i^1, , ..., e_i^K]$, with the parameters constituting the Dirichlet distribution for $agent_i$, $\alpha_i = [\alpha_i^1, , ..., \alpha_i^K]$. Here, by employing ReLU in the final layer, all evidence values are ensured to be non-negative. The parameter α_i^k is directly influenced by e_i^k, specifically, $\alpha_i^k = e_i^k + 1$. Subsequently, the belief mass b_i^k and the overarching uncertainty u_i can be deduced as:

$$b_i^k = \frac{e_i^k}{S_i} = \frac{\alpha_i^k - 1}{S_i} \; and \; u_i = \frac{K}{S_i} \tag{2}$$

where $S_i = \sum_{k=1}^{K}(e_i^k + 1) = \sum_{k=1}^{K} \alpha_i^k$ represents the strength of the Dirichlet distribution (Jsang 2018). Eq. 2 captures an intuitive phenomenon: the more evidence accumulated for the k^{th} action, the higher the probability attributed to that action. Inversely, when there's scant evidence, the encompassing uncertainty escalates. This belief assignment can be interpreted as a form of subjective reasoning.

To enhance decision-making precision and trustworthiness, we propose leveraging evidence collected by different agents as a foundation for decision-making. Consequently, we develop an evidence encoder to deduce bespoke evidence tailored for each agent. At each time-step, $agent_i$ not only produces evidence e_i for its own local decision but also extracts a collection of evidence - $(e_{i1}, ..., e_{ij}, ..., e_{in})$, aimed at aiding its teammates in making more reliable choices. Such evidence then acts as the communication medium, enabling us to generate messages tailored for specific agents. The tailored message from $agent_i$ to $agent_j$ can be denoted as $m_{ij} = e_{ij}$.

4.2 Selective Engagement

As we've discussed earlier, broadcast communication falls short in practical applications and results in redundant information. The timing of information exchange and the choice of communication partners are paramount. For precise and trustworthy message exchanges, it's vital to identify truly instrumental connections from the vast web of interactions. At a holistic level, we aim to share evidence-backed data, thus enabling recipients to make informed and reliable decisions. To bring this vision to fruition, we meticulously quantify the strength and relevance of each communication link between agents by performing an ablative decision-making analysis. This approach primarily seeks to quantify the variability in decision uncertainty attributable to communication. To delve deeper into the mechanics, consider the communication from $agent_i$ to $agent_j$, denoted as m_{ij}. This communication's value is mathematically expressed as:

$$v_{ij} = u_j - \hat{u_j} \tag{3}$$

where u_j represents the decision uncertainty for recipient $agent_j$ before communication, whereas $\hat{u_j}$ is the uncertainty post communication.

To foster targeted and trusted communication, we develop a communication selector network. This network aids agents in determining the right moments and partners for communication, ensuring that only the most valuable and credible information is shared. We also set a constant threshold to generate binary pseudo-labels. If the deduced communication value is below the set threshold, it implies that the message received doesn't substantially benefit the recipient agent, leading the connection to be tagged as 'cut', denoted mathematically as $y_{ij} = 0$. However, if the communication value exceeds the threshold, it signifies the message's importance, prompting its tag to be 'retain' with $y_{ij} = 1$. This systematic labeling forms the foundation for optimizing the communication selector network, with the binary cross-entropy loss steering the fine-tuning process.

$$\mathcal{L}_{BCE} = \mathbb{E}_{i,j\sim\mathbb{Z}^n}[y_{ij}\times\log(p_{ij})+(1-y_{ij})\times\log(1-p_{ij})] \tag{4}$$

where $\mathbb{Z}^n$ is the set of integers from 1 to n, p_{ij} is the output of the communication selector network, representing the likelihood of $agent_i$ choosing to communicate with $agent_j$.

4.3 Evidence-Driven Integration

In T2MAC, messages exchanged among agents encapsulate evidence observed from diverse perspectives. Agents can better understand the uncertain environment by adeptly integrating these messages, resulting in more sophisticated policies. To this end, we incorporate the DST to integrate incoming messages. This approach facilitates the combination of evidence from different sources, culminating in a degree of belief that comprehensively reflects all gathered evidence. The rule of message integration for evidence is presented as:

$$\mathcal{M} = \mathcal{M}_i \oplus \mathcal{M}_j \tag{5}$$

where $\mathcal{M}_i = \{\{b_i^k\}_{k=1}^K, u_i\}$ and $\mathcal{M}_j = \{\{b_j^k\}_{k=1}^K, u_j\}$ symbolize the joint masses derived from two distinct perspectives of evidence and $\oplus$ represents DST combination. Meanwhile, $\mathcal{M} = \{\{b^k\}_{k=1}^K, u\}$ encapsulates the consolidated joint mass, integrating evidence from both standpoints. The more specific integration rule can be formulated as follows:

$$b^k = \frac{1}{1-C}(b_i^k b_j^k + b_i^k u_j + b_j^k u_j), \; u = \frac{1}{1-C} u_i u_j \tag{6}$$

where $C = \sum_{k\neq k'} b_i^k b_j^{k'}$ represents the degree of disagreement between the two sets of mass values. To account for this discord, DST employs the normalization factor $\frac{1}{1-C}$ to ensure a coherent integration of the evidence from both sets. Intuitively, when encountering evidence and beliefs from multiple sources, DST aims to merge the common elements and sidesteps conflicting beliefs through normalization factors. The integration rule ensures:

1. If both perspectives exhibit high uncertainty (with significant values of u_i and u_j), the resultant prediction should be treated cautiously, yielding a lower confidence level (represented by a smaller value of b^k).

2. Conversely, if both viewpoints possess low uncertainty (denoted by minimal values of u_i and u_j), the resulting prediction is likely to be made with a high degree of confidence (manifesting as a larger value of b^k);
3. In situations where only one viewpoint exhibits low uncertainty (meaning either u_i or u_j is significantly large), the final prediction predominantly relies on the more confident viewpoint.

Upon receiving distinct messages from other agents, we derive the aforementioned mass for each perspective. Subsequently, leveraging Dempster's rule of combination, we can integrate the beliefs stemming from these varied viewpoints. More specifically, the fusion of belief and uncertainty masses across different messages is governed by the subsequent rule:

$$\mathcal{M} = \mathcal{M}_1 \oplus \mathcal{M}_2 \oplus ...\mathcal{M}_n \tag{7}$$

Once we have determined the joint mass $\mathcal{M} = \{\{b^k\}_{k=1}^K, u\}$, the associated joint evidence gleaned from the messages, along with the parameters of the Dirichlet distribution, can be derived as follows:

$$S = \frac{K}{u}, e^k = b^k \times S \; and \; \alpha^k = e^k + 1 \tag{8}$$

Leveraging DST, we attain an efficient and theoretically-founded method for message integration. This method skillfully amalgamates messages and simultaneously addresses enduring intrinsic policy uncertainties. Importantly, the fusion of information isn't treated as a black box, given that the combination rules of DST lack learnable parameters. Furthermore, DST offers a more comprehensible and theoretical perspective on the message integration process.

Following the assimilation of incoming messages and the acquisition of integrated evidence, each agent makes a local decision influenced by both its observed and received evidence. For $agent_i$, this procedure is represented as:

$$a_i^t = \pi_i(\hat{e}_i) \tag{9}$$

where $\hat{e}_i$ symbolizes the evidence post-integration for $agent_i$ at time-step t. For details of the communication process and the training paradigm of T2MAC, please refer to pseudo-code provided in this section.

Algorithm 1: T2MAC

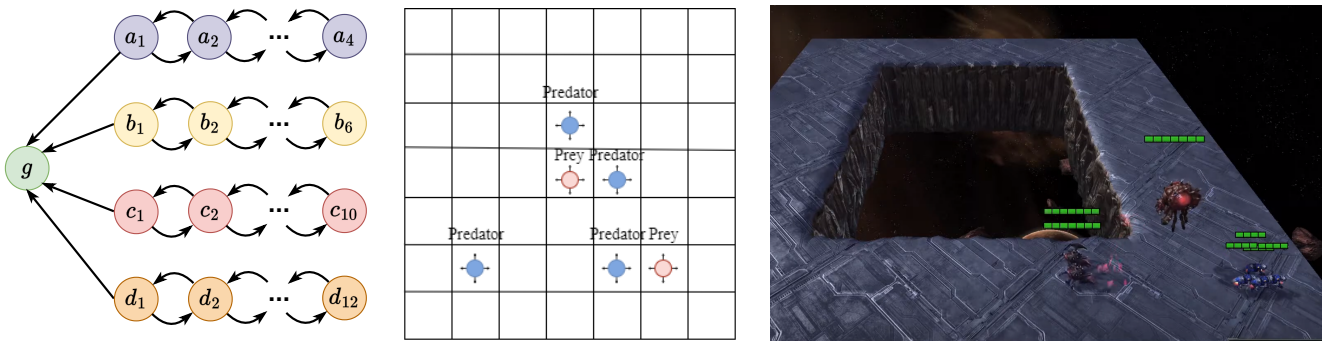
```
Initialize replay buffer D
Initialize the Observation encoder, Evidence encoder, Selective Engagement and Q network with random parameters
Set learning rate α and max training episode E
for episode in 1, ..., E do
  for each agent i do
    Sending Phase: Encode the hidden feature h_i^t from observation o_i^t
    Encode evidence e_i^t for local decision
    Encode evidence and generate tailored messages for specific teammates (e_i1, ..., e_ij, ..., e_in)
    Select ideal communication partners using communication selector network
    Receiving Phase: Combing received messages m_⋆i^t from other agents by DST combine
    Select action a_i^t by combined evidence
    Compute the importance for each communication link and generating labels y_ij for communication selector network
  end for
  Store the trajectory in replay buffer D
  Sample a minibatch of trajectories from D
  Update observation encoder, evidence encoder and policy network using MARL loss function
  Update Selective Engagement by Equation 4
end for
```

5 Experiments

In this section, we carefully design experiments to address three pivotal questions: (1) How does T2MAC's performance measure against top-tier communication methods? (2) What characterizes T2MAC's communication efficiency? (3) Can T2MAC scale across various tasks and seamlessly integrate with multiple baselines?

5.1 Setup

As illustrated in Fig. 2, we extensively evaluate T2MAC across three notable cooperative multi-agent tasks. Beginning with Hallway (Wang et al. 2020b), this environment is relatively direct, built around multiple Markov chains. Here, agents start at random positions within different chains and aim to reach the goal state simultaneously under partial observability. To escalate the complexity, we augment the number of agents and the length of the Markov chains, leading to a substantial increase in the exploration space. On the other hand, MPE (Lowe et al. 2017) is a vital MARL benchmark set in a 2D grid. We focus on the Cooperative Navigation (CN) and Predator Prey (PP) scenarios. In CN, the task for agents is to navigate to different landmarks, whereas, in PP, their objective is to capture unpredictably moving prey. To introduce varying difficulty levels, we employ different grid sizes for both scenarios. The Cooperative Navigation: Medium scenario is set on a 7×7 grid, while the Cooperative Navigation: Hard occupies a 9×9 grid. The Predator Prey: Medium scenario is set on a 5×5 grid, while the Predator Prey: Hard occupies a 7×7 grid. SMAC (Samvelyan et al. 2019) is derived from the well-known real-time strategy game StarCraft II. It delves into micromanagement challenges where each unit is steered by an independent agent making decisions under partial observability. To emphasize the importance of communication, we adopt the setup from (Wang et al. 2020b), which not only restricts the agents' sight range but also throws them into intricate maps, characterized either by their labyrinthine terrains or the unpredictable spawning dynamics of units. For comparative analysis, we draw from a diverse set of baselines. This includes non-communication paradigms like the leading MARL methods QMIX (Rashid et al. 2018) and DOP (Wang et al. 2020). Meanwhile, our baselines include con-



(a) Hallway (b) MPE (c) SMAC

Figure 2: Multiple environments considered in our experiments.

temporary state-of-the-art communication methods, such as TarMAC (Das et al. 2019), MAIC (Yuan et al. 2022), SMS (Xue et al. 2022), and MASIA (Guan et al. 2022).

In conclusion, our experimental design integrates a medley of challenging tasks and robust baselines, establishing a solid foundation for evaluation. Our overarching goal with this varied selection is to place T2MAC in diverse scenarios and test its adaptability, scalability, and overall performance. To ensure transparency and reproducibility, the intricate details of our method's architecture and our hyperparameter choices are extensively detailed in Table 1.

Module	Architecture
Obs Encoder	Linear(obs_dim, 64) Linear(64, 64) Linear(64, 64) RNN(64, 64)
Evidence Encoder	n*Linear(64, K)
Selector Network	Linear(64, n)

Table 1: Hyperparameters of T2MAC

5.2 Results

Performance We begin our evaluation by comparing the learning curves of T2MAC with various baselines across various environments to test its overarching performance. As illustrated in Fig. 3, T2MAC emerges superior in almost all environments, highlighting its robust performance. In Hallway, as the difficulty intensifies, many baselines falter, unable to adapt effectively. Among them, only MASIA stands out, delivering commendable results, primarily due to its ability to assist agents in reconstructing global information. Intriguingly, our T2MAC works even under such demanding conditions, achieving performance on par with MASIA. This might be largely attributed to its adeptness at sharing and integrating relevant evidence. In SMAC, T2MAC delivers consistent and impressive performance across all three maps. However, when looking at all scenarios in their entirety, other methods exhibit signs of instability. For instance, SMS struggles to adapt in the $5z\_vs\_1ul$, while TarMAC fails in the $1o\_10b\_vs\_1r$. Such observations accentuate, to some extent, the broad applicability and robustness inherent to T2MAC. In CN and PP, T2MAC maintains its sustained sample efficiency. Upon reaching a convergence, its performance remains fiercely competitive. Furthermore, an interesting observation is that all methods incorporating communication significantly outperform those that don't. This emphasizes that our chosen environments and scenarios intrinsically demand proficient communication. Such an outcome not only underscores the importance of communication in these contexts but also validates the aptness of our experimental setup in benchmarking communication methods.

Methods	Performance Improvement	Comm Rate	Comm Efficiency
TarMAC	17.0%	100.0%	17.0%
MAIC	12.3%	100.0%	12.3%
SMS	27.9%	66.7%	41.8%
MASIA	30.2%	100.0%	30.2%
T2MAC(Ours)	**37.2%**	**56.0%**	**66.4%**

Table 2: Communication Efficiency

Efficiency In addition to analyzing the overarching performance, we also focus on understanding communication efficiency. In many real-world situations, communication resources—like bandwidth and transmission channels—are inherently scarce. Overloading these resources doesn't always yield proportional benefits in performance. To quantify this efficiency, we calculate the performance improvement attributable to communication and then normalize this by the communication rate. Here, communication rate denotes how frequently communication occurs throughout the learning process. To gauge performance improvement, we introduce a communication-free variant for each communication method. This allows us to make a side-by-side comparison to effectively highlight the tangible advantages offered by each method. Specifically, for SMS, this communication-

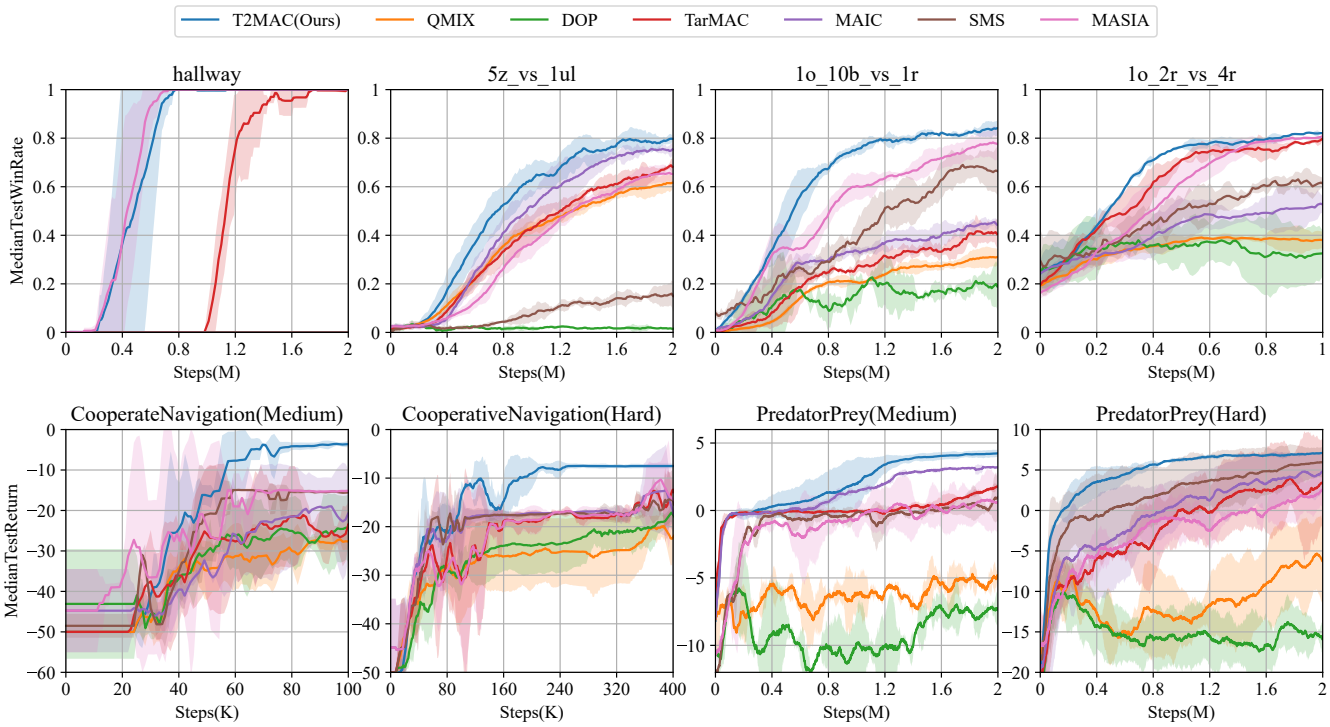


Figure 3: Performance on multiple benchmarks.

free baseline is DOP, while for the others, it's QMIX. We've carried out this analytical assessment predominantly in the most challenging environment, SMAC. As shown in Table. 2, T2MAC consistently outperforms baselines in terms of both improvement, communication rate, and communication efficiency. Such results underscore the capability of T2MAC to process communication dynamics, including when to communicate, with whom, and how to trade-off between performance and efficiency.

Generality Our prior experiments have demonstrated the robustness of T2MAC across diverse environments, scenarios of varying complexities, and different scales. To further evaluate the generality of T2MAC, we apply it across a wide range of established MARL baselines, including QMIX, DOP, and MAPPO. The test win rate for the scenario $1o\_10b\_vs\_1r$ is illustrated in Fig. 4. Notably, across all these baselines, T2MAC consistently achieves superior performance, often by a notable margin. This positive performance improvement demonstrates the broad applicability and potency of T2MAC in the realm of MARL.

Ablation To better understand the impact of each component within T2MAC, we perform an ablation study on the scenario $1o\_10b\_vs\_1r$. Here's a breakdown of the configurations evaluated: **T2MAC**: This refers to the complete method proposed in our work. **QMIX**: This serves as our baseline for comparison, representing the core functionality without the enhancements introduced in T2MAC. **T2MAC(Fullcomm)**: This is a variant of T2MAC that does not incorporate selective engagement. Here, communication occurs continuously amongst agents without deciding when or with whom to communicate. **T2MAC(Nocomm)**: This is a more stripped-down version of T2MAC, excluding both selective engagement and evidence-driven integration. Essentially, it's a version of T2MAC where communication is completely omitted, but the Dirichlet Distribution remains in the Q-value network. As illustrated in Fig. 5, the results demonstrate the contributions of each component: **From QMIX to T2MAC(Nocomm):** The shift from Categorical distribution to Dirichlet distribution makes sense. The Dirichlet distribution's advantage might stem from its ability to model second-order probabilities, introducing an additional layer of decision-making uncertainty which potentially enhances learning and adaptation. **From T2MAC(Nocomm) and T2MAC(Fullcomm)**: The sizable performance gap between these two underscores the significance of evidence-driven information exchange and integration. This sheds light on the efficacy of trust-based communication, where agents not only share but also assess the reliability of information before acting upon it. **From T2MAC(Fullcomm) to T2MAC**: The contrast in performance between these two configurations underlines the importance of targeted communication. Instead of a blanket communication strategy, selective engagement, whereby agents communicate at strategic junctures with specific partners, can enhance the overall efficiency and performance of the system.

Furthermore, to provide a clear ablation analysis for evidence-driven integration, we have conducted additional comparisons in the $1o\_10b\_vs\_1r$ scenario

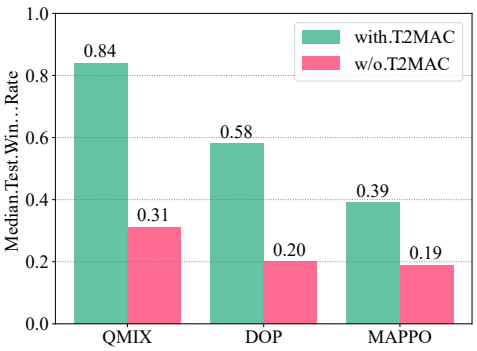


Figure 4: Generality.

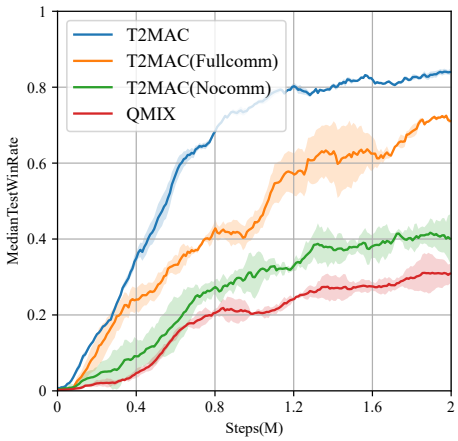


Figure 5: Ablation for trusted communication and selective engagement.

with a summation-based integration method (COMMNET(Sukhbaatar, Szlam, and Fergus 2016)) and a black-box method (TarMAC(Das et al. 2019)). As shown in Fig. 6, the results demonstrate that the evidence-driven integration proposed by T2MAC has a clear advantage, confirming its effectiveness.

6 Conclusions

In this work, we tackle the intricacies inherent in multi-agent communication. Previous works focus on broadcast communication and treat the fusion of information as a block box, which inevitably diminishes communication efficiency. To this end, we present the T2MAC framework. This novel approach empowers agents with the capacity to craft messages specifically tailored for distinct agents. Beyond mere message customization, T2MAC strategically chooses the best timings and relies on trusted partners for communication, ensuring an efficient integration of incoming messages and facilitating trusted decision-making. Rooted in solid theoretical principles, this approach stands out for its efficiency. Furthermore, to substantiate our claims, we conduct comprehensive experiments across multiple benchmarks, the results of which underscore the effectiveness, efficiency, and adaptability of the T2MAC.

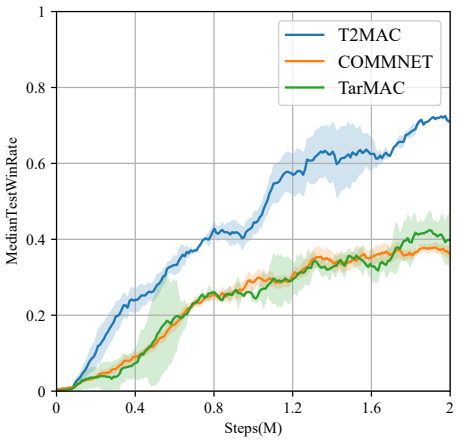


Figure 6: Ablation for evidence-driven integration.

Acknowledgements

The authors would like to thank the editors and reviewers for their valuable comments. This work is supported by the Youth Innovation Promotion Association CAS, No. 2021106, the China Postdoctoral Science Foundation, No. 2023M743639, the 2022 Special Research Assistant Grant project, No. E3YD5901, and the CAS Project for Young Scientists in Basic Research, Grant No. YSBR-040.

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.

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.

Dempster, A. P. 1967. Upper and Lower Probabilities Induced by a Multivalued Mapping. *The Annals of Mathematical Statistics*, 38(2): 325 – 339.

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.

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.

Jiang, J.; and Lu, Z. 2018. Learning attentional communication for multi-agent cooperation. In *Advances in neural information processing systems*, 7254–7264.

Jsang, A. 2018. *Subjective Logic: A formalism for reasoning under uncertainty*. Springer Publishing Company, Incorporated.

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.

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.

Malinin, A.; and Gales, M. J. F. 2018. Predictive Uncertainty Estimation via Prior Networks. In Bengio, S.; Wallach, H. M.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, 7047–7058.

Malinin, A.; Mlodozeniec, B.; and Gales, M. J. F. 2020. Ensemble Distribution Distillation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Mao, H.; Zhang, Z.; Xiao, Z.; Gong, Z.; and Ni, Y. 2019. Learning agent communication under limited bandwidth by message pruning. *arXiv preprint arXiv:1912.05304*.

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

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.

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*.

Sensoy, M.; Kaplan, L. M.; and Kandemir, M. 2018. Evidential Deep Learning to Quantify Classification Uncertainty. In Bengio, S.; Wallach, H. M.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, 3183–3193.

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.; 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.

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*.

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, R.; He, X.; Yu, R.; Qiu, W.; An, B.; and Rabinovich, Z. 2020a. Learning Efficient Multi-agent Communication: An Information Bottleneck Approach. In *ICML 2020: 37th International Conference on Machine Learning*.

Wang, T.; Wang, J.; Zheng, C.; and Zhang, C. 2020b. 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*.

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.

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.