Task-Agnostic Contrastive pre-Training for Inter-Agent Communication

Peihong Yu
University of Maryland
College Park, USA
peihong@umd.edu

Manav Mishra
IISER Bhopal
Bhopal, India
mishra20@iiserb.ac.in

Syed Zaidi
University of Maryland
College Park, USA
szaidi@terpmail.umd.edu

Pratap Tokekar
University of Maryland
College Park, USA
tokekar@umd.edu

ABSTRACT

The "sight range dilemma" in cooperative Multi-Agent Reinforcement Learning (MARL) presents a significant challenge: limited observability hinders team coordination, while extensive sight ranges lead to distracted attention and reduced performance. While communication can potentially address this issue, existing methods often struggle to generalize across different sight ranges, limiting their effectiveness. We propose TACTIC, Task-Agnostic Contrastive pre-Training strategy Inter-Agent Communication. TACTIC is an adaptive communication mechanism that enhances agent coordination even when the sight range during execution is vastly different from that during training. The communication mechanism encodes messages and integrates them with local observations, generating representations grounded in the global state using contrastive learning. By learning to generate and interpret messages that capture important information about the whole environment, TACTIC enables agents to effectively "see" more through communication, regardless of their sight ranges. We comprehensively evaluate TACTIC on the SMACv2 benchmark across various scenarios with broad sight ranges. The results demonstrate that TACTIC consistently outperforms traditional state-of-the-art MARL techniques with and without communication, in terms of generalizing to sight ranges different from those seen in training, particularly in cases of extremely limited or extensive observability.

KEYWORDS

MARL, Communication, Contrastive Learning

ACM Reference Format:
Peihong Yu, Manav Mishra, Syed Zaidi, and Pratap Tokekar. 2025. Task-Agnostic Contrastive pre-Training for Inter-Agent Communication. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 9 pages.

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

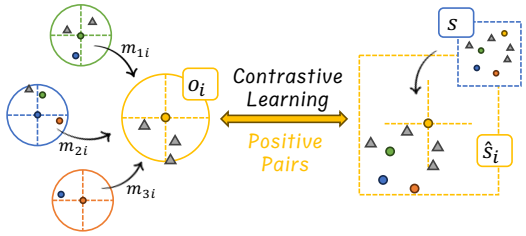

Figure 1: TACTIC utilizes contrastive learning to align the integration of local observations o_i and messages $\{m_{ji}\}$ with the full egocentric state $\hat{s}_i$ for each agent i, enabling agents to "see" beyond their limited sight ranges through communication.

1 INTRODUCTION

Multi-agent Reinforcement Learning (MARL) provides a framework for addressing complex coordination tasks across various domains such as robotics [1, 30, 32], autonomous vehicles [15, 19], and network optimization [9, 34]. In MARL, agents often operate under partial observability, where each agent's perception is limited to a certain "sight range" around itself, which results in a fundamental challenge known as the *sight range dilemma* [20]. The dilemma lies in balancing tension between an agent's need for local information to make decisions and the broader context required for effective team coordination. Agents with narrow sight ranges often struggle to coordinate effectively due to limited environmental information, while those with extensive sight ranges can become overwhelmed by excessive data, leading to inefficient learning and reduced performance.

Researchers have proposed various approaches to addressing this challenge, primarily focusing on communication mechanisms that allow agents to share information. Those methods include targeted communication strategies [2, 11, 24], attention mechanisms [13, 18, 21], or graph-based methods [7, 22, 26]. For example, QMIX-Att [6] integrates attention into the QMIX framework for selective message aggregation, and TarMAC [2] uses signature-message pairs for context-aware communication. While effective, these methods assume the same sight ranges during training and during execution, limiting their capacity to adapt to varying visibility conditions, as illustrated in Figure 2.

Generalizing across varying sight ranges provides considerable benefits for MARL systems. It facilitates more efficient and cost-effective deployments by enabling a single, adaptable model to handle diverse observability conditions, eliminating the need for separate models for each scenario. This flexibility is crucial in real-world applications, where systems must adjust to different visibility conditions. For example, autonomous vehicles need to adapt to visibility fluctuations due to weather or time of day, while search and rescue robots may encounter visual obstructions from debris or smoke. In this work, we are particularly interested in the cases where the sight range during execution is fixed but different from that seen during training.

We hypothesize that *agents can better generalize across different sight ranges if they can communicate in a way that leads to a more comprehensive understanding of the global environment.* Based on this hypothesis, we propose a novel approach that aligns the integrated local observations and messages with each agent's egocentric (global) state, as illustrated in Figure 1. The egocentric state serves as an ideal alignment target, providing a comprehensive yet agent-specific view of the environment during training. We use contrastive learning to achieve this alignment, encouraging agents to develop a communication protocol that bridges the gap between limited local observations and the broader environmental context. This process enables agents to effectively "see" more through communication, regardless of their actual sight ranges.

In addition to the contrastive learning objective, we introduce two auxiliary losses: a reconstruction loss and a dynamics loss. The reconstruction loss helps ensure that the learned representations retain essential information from the original observations, while the dynamics loss encourages the model to capture the temporal relationships in the environment. Crucially, our method is task-agnostic in nature, as it does not rely on task-specific reward information when learning to communicate, solely focusing on capturing the underlying environment information, which further enhances the flexibility and adaptability of our method across diverse scenarios.

To this end, we introduce **TACTIC** (**T**ask-**A**gnostic **C**ontrastive pre-**T**raining for **I**nter-Agent **C**ommunication), a novel strategy designed to enhance generalization across varying sight ranges in cooperative MARL. TACTIC operates through two key stages: (1) **Offline contrastive pretraining**, where we use contrastive learning on an offline dataset to pretrain two key communication modules: a message generator and a message-observation integrator. (2) **Online policy integration**, where the pre-trained communication modules are frozen and incorporated into agents' online policy learning, enabling dynamic communication adaptation during task execution while preserving the learned task-agnostic properties.

We summarize the main contributions of this work as follows:

- A task-agnostic communication mechanism that enables adaptive message generation and interpretation;
- A cooperative MARL framework with communication called TACTIC that alleviates the sight range dilemma;
- A comprehensive evaluation of TACTIC in the SMACv2 environment showing TACTIC's superior performance regarding generalizability across sight ranges and training efficiency.

Our experimental results on the SMACv2 (StarCraft Multi-Agent Challenge) benchmark show that TACTIC outperforms existing state-of-the-art MARL with communication techniques. Our method demonstrates robust generalization capabilities, enabling effective coordination across various sight ranges (Figure 2).

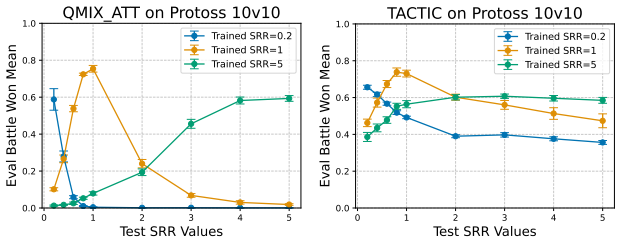

Figure 2: QMIX-ATT and TACTIC's performances on Protoss 10v10 from SMACv2 [4] with varying sight ranges (SRs). Different SRs are achieved by applying different sight-range ratios (SRRs) to the agents' original SRs in the implementation. Policies trained at SRR=0.2, 1, and 5 are tested across a broader set of SRRs. QMIX-ATT Struggles to generalize to unseen SRs, while TACTIC generalizes much better.

2 RELATED WORK

Communication in MARL. In cooperative MARL, communication is a critical component for addressing partial observability and improving agent coordination [12, 25]. Learning effective communication in this context involves several challenges, including determining *who* communicates, *how* messages are conveyed, and *what* information is transmitted under bandwidth or sight-range constraints.

Several approaches have been proposed to address these challenges. Targeted communication methods focus on identifying specific agents for message exchange, while graph-based and attention-based models structure communication based on relational dynamics between agents. Graph-based methods, such as the graph-attention network proposed by Niu et al. [13], enable agents to dynamically adjust communication based on relevance, improving scalability in complex environments. Techniques that manage bandwidth limitations address the need to optimize when and what information should be shared. For example, attention-based methods such as TarMAC [2] leverage signature-message pairs and attention mechanisms to enable dynamic, context-aware communication between agents. Information-theoretic approaches, such as NDQ [28], introduce regularization to minimize communication overhead while maximizing the informativeness of messages.

Recent advancements also aim to make multi-agent communication more interpretable and flexible. Lin et al. [10] proposed grounding communication by autoencoding raw observations into messages, allowing agents to develop a shared understanding of communication symbols. Similarly, MASIA [5] aggregates raw observations into latent representations that can be used to reconstruct the global state, providing agents with a more holistic view of the environment. Du et al. [3] introduced methods to learn correlated communication topologies, which reduce redundancy and optimize coordination among agents by refining communication pathways. In parallel, Zhang et al.[31] introduced Temporal Message Control (TMC), a technique that applies temporal smoothing

to reduce the number of inter-agent messages, achieving robust and efficient communication in resource-constrained environments without sacrificing performance.

Our work builds on these advancements by addressing adaptive communication under sight-range limitations, proposing a novel approach that optimizes communication frequency and content based on changing agent observations. This differs from previous works by integrating bandwidth and perceptual constraints into a unified framework, enabling efficient communication without reliance on pre-defined structures.

Contrastive Learning in MARL. Contrastive learning is a representation learning technique that aims to bring similar (positive) samples closer together in the learned feature space while pushing dissimilar (negative) samples farther apart. This is typically achieved using a contrastive loss function.. Methods such as Contrastive Predictive Coding (CPC) by Oord et al. [27] laid the groundwork for learning predictive representations by contrasting positive and negative samples. Building on this, TACO [35] adapts contrastive learning to RL by learning useful representations through temporal abstraction. More recently, the use of supervised contrastive loss [8], which incorporates multiple positive and negative samples per anchor point, has been explored. This extension enables richer representation learning by capturing a broader set of relevant relationships, which is particularly valuable in RL tasks with multiple favorable outcomes.

In the multi-agent domain, contrastive learning has seen adaptation for both non-communicative and communicative settings. For MARL without communication, methods like COLA [29] have demonstrated the ability to improve coordination among agents by using contrastive objectives to refine agent policies based on shared goals. This approach emphasizes the utility of contrastive learning in situations where direct agent-to-agent communication is absent or limited. In contrast, for settings where agent communication is possible, contrastive learning has been utilized to optimize communication strategies between agents. Lo et al. [12] apply contrastive learning to improve multi-agent communication, enabling agents to develop more efficient communication protocols that reduce unnecessary message exchanges while maintaining performance. Additionally, methods such as the one proposed by Singh et al. [23] focus on learning when to communicate, which is crucial in reducing communication overhead in resource-constrained environments. Zhang et al. [33] further refine this by incorporating variance-based control mechanisms, allowing agents to communicate only when necessary, improving overall system efficiency.

Our work extends these ideas by integrating contrastive learning into multi-agent communication under sight range limitations. This approach differs from previous work in that we train the communication module offline specifically with the goal of generalizing across sight ranges.

3 PRELIMINARIES

Dec-POMDP with Communication. We consider the fully cooperative MARL problem with communication, which can be modeled as Decentralized Partially Observable Markov Decision Process (Dec-POMDP) [14] and formulated as a tuple $\langle \mathcal{N}, \mathcal{S}, \mathcal{A}, P, \Omega, O, R, \gamma, C \rangle$. The sets $\mathcal{N} = \{1, ..., n\}$ denotes the indexing of the agents, $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, Ω is observation space, and C denotes all possible communication messages. Each agent $i \in \mathcal{N}$ acquires an observation $o_i = O(s, i) \in \Omega$, where O is the observation function and $s \in \mathcal{S}$. A joint action $\boldsymbol{a} = \langle a_1, ..., a_n \rangle$ leads to the next state $s' \sim P(s'|s, \boldsymbol{a})$ and a shared global reward $r = R(s, \boldsymbol{a})$ where R is the reward function.

Each agent selects actions based on the observation-action history $\tau_i \in \mathcal{T} \equiv (\Omega \times \mathcal{A})^*$ [1] using a policy $\pi(a_i|\tau_i, m_i)$ where $m_i = [m_{ji} \in C, j \in \mathcal{N}]$ denotes the incoming messages for agent i and m_{ji} is the message sent from agent j to agent i. The policy is shared across agents during training.

The overall objective is to find a joint policy $\boldsymbol{\pi}(\boldsymbol{\tau}, \boldsymbol{a})$ to maximize the global value function

$$Q^{\boldsymbol{\pi}}(\boldsymbol{\tau}, \boldsymbol{a}) = \mathbb{E}_{s,\boldsymbol{a}} \left[\sum_{t=0}^{\infty} \gamma^t R(s, \boldsymbol{a}) \mid s_0 = s, \boldsymbol{a_0} = \boldsymbol{a}, \boldsymbol{\pi} \right], \tag{1}$$

where $\boldsymbol{\tau}$ is the joint observation-action history of all agents and $\gamma \in [0, 1)$ is the discount factor. We follow the *Centralized Training and Decentralized Execution* (CTDE) paradigm and adopt the architecture of QMIX [16] to form our algorithm [2].

Contrastive Learning. Contrastive learning is a powerful paradigm in representation learning, particularly in the context of deep learning, where it aims to learn embeddings by contrasting positive and negative samples. The fundamental idea is to pull together representations of similar instances (positives) while pushing apart those of dissimilar instances (negatives). In the supervised setting, the SupCon (Supervised Contrastive) loss [8] extends this framework by allowing for multiple positive samples per anchor, thereby leveraging label information more effectively. The loss objective can be mathematically expressed as:

$$L_{supcon} = \sum_i \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \left(\frac{\exp(z_i \cdot z_p / \upsilon)}{\sum_{a \in A(i)} \exp(z_i \cdot z_a / \upsilon)} \right) \tag{2}$$

where z_i denotes the normalized embedding of the anchor sample, $P(i)$ is the set of positive samples corresponding to the anchor, $A(i)$ is the set of all samples in the batch excluding the anchor, and υ is a temperature parameter that controls the sharpness of the distribution. This formulation enhances the clustering of similar instances in the embedding space and improves robustness against natural corruptions.

4 TACTIC: TOWARDS TASK-AGNOSTIC ADAPTIVE COMMUNICATION IN MARL

In this section, we introduce our algorithm TACTIC. TACTIC consists of two stages: (1) offline training of a communication mechanism that works well with varied sight ranges, and (2) online training of the agent coordination policy. The task-agnostic offline training stage solely utilizes the environment states and local observations of agents and doesn't rely on environmental reward signals or any policy. The online policy training stage learns a task-specific policy with the communication module from the first stage frozen. We explain the offline stage in Section 4.1 and the online stage in Section 4.2 in detail.

[1] * denotes the product over time

[2] For clarity, we drop the time superscripts for states and actions

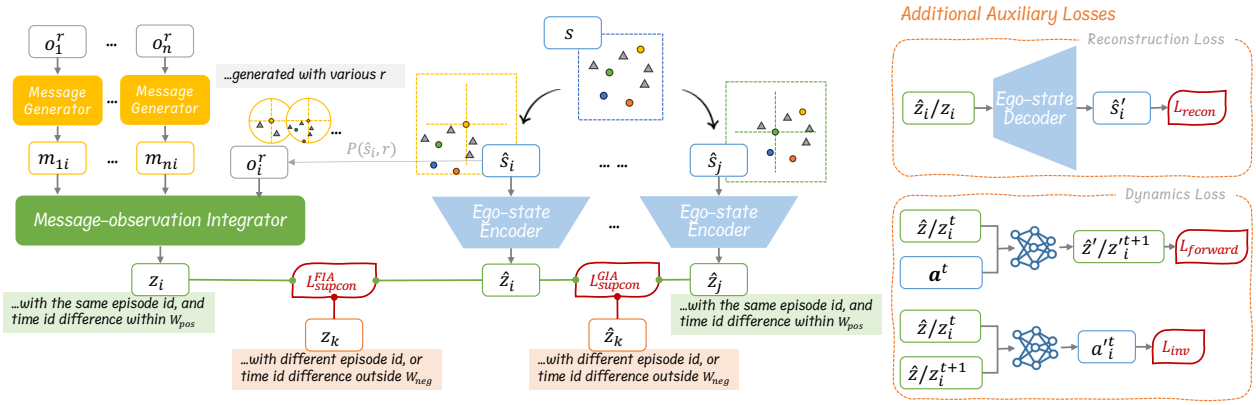

Figure 3: The offline training pipeline of the adaptive communication mechanism(Section 4.1). It includes three key components: an egocentric state encoder, an adaptive message generator, and a message-observation integrator. The training pipeline consists of two contrastive learning processes: Global Information Alignment (GIA) for aligning the features generated from the egocentric state encoder across all agents and timesteps, and Feature Integration Alignment (FIA) for aligning features from the message-observation integrator and the egocentric state encoder on an individual agent. Two auxiliary loss functions are introduced in the total loss function to enhance training: a deconstruction loss for learning to recover the egocentric states and a dynamic loss for learning temporally coherent representations.

4.1 Offline Training of Communication Mechanism

One main challenge in Multi-Agent Reinforcement Learning (MARL) is to develop effective joint policies when each agent has only a partial observation of the environment. By learning to communicate more effectively, agents are expected to overcome their limited observability and achieve better coordination. Furthermore, the mechanism used to generate and use messages between agents should be flexible enough to handle different sight ranges.

To this end, we present an approach where, in an offline training stage, we use contrastive learning on a pre-collected dataset to develop a communication mechanism that can adapt to different observation ranges. The rationale is that, by learning to create and integrate messages that capture important information about the whole environment, agents can effectively **"see" more through communication**, regardless of their current sight range.

The offline dataset $\mathcal{D}$ consists of a set of trajectories, where each trajectory $\tau = \{(s^t, o^t_{1:n}, a^t_{1:n})\}_{t=1}^{T}$ represents a sequence of T timesteps for n agents. The dataset can be collected through random exploration of the environment (details in Section 5.3). The overall offline training pipeline for the adaptive communication mechanism is illustrated in Figure 3, where three key components are present: an egocentric state encoder, an adaptive message generator, and a message-observation integrator:

(1) The *egocentric state encoder* takes an egocentric state $\hat{s}_i$ and generates its corresponding feature embedding $\hat{z}_i$, where the $\hat{s}_i$ are obtained from the global state s, which preserves all information from s but represents from the perspective of agent i.
(2) The *adaptive message generator* takes the partial observation o^r_i of an agent with *varying* sight ranges and outputs the message $\{m_{ij}\}_{j=1}^{n}$ it communicates to other agents. We obtain o^r_i by randomly sampling a sight range r and applying a masking operation over the egocentric state $\hat{s}_i$, denoted as $P(\hat{s}_i, r)$.
(3) The *message-observation integrator* takes an agent's partial observation o^r_i and all the messages $\{m_{ji}\}_{i=1}^{n}$ it receives from other agents, integrating them into a feature embedding z_i.

With the three components at hand, the offline training consists of two contrastive learning processes:

(1) *Global Information Alignment (GIA)*: a supervised contrastive (SupCon) loss for aligning the feature embeddings $\hat{z}_i$ generated from the egocentric state encoder across all agents, ensuring that the egocentric state encoder captures consistent and relevant information from agents about the whole environment;
(2) *Feature Integration Alignment (FIA)*: a supervised contrastive (SupCon) loss for aligning the integrated feature z_i calculated by the message-observation integrator with the generated feature $\hat{z}_i$ from the egocentric state encoder for each specific agent, pushing the adaptive message generator to learn to synthesize the most informative messages to communicate and allowing the message-observation integrator to reflect a more complete picture of the environment (i.e., the agent's egocentric state) given the agent's limited observation and the messages they receive.

In FIA, the function $P(\hat{s}_i, r)$ applies augmentation by randomly varying the sight range r (resulting in o^r_i). This process exposes the adaptive message generator to diverse scenarios with changing sight ranges, rather than fixed ones, thereby enhancing its generalizability. For contrastive learning in both GIA and FIA, we define positive and negative pairs based on the offline dataset $\mathcal{D}$. Features from agents within the same episode (i.e., from the same trajectory)

and within a timestep window of length W_{pos} form positive pairs. Conversely, features from agents in different episodes or separated by more than W_{neg} timesteps constitute negative pairs. Note that both GIA and FIA are task-agnostic as they don't interact with any environmental reward signals.

To further improve the learned representations, we incorporate two additional auxiliary learning objectives: (1) a reconstruction loss (L_{recon}): a decoder network learns to recover the egocentric state $\hat{s}_i$ from the feature embeddings produced by either the egocentric state encoder or the message-observation integrator; and (2) a dynamic loss (L_{dyn}): this includes both forward and inverse dynamics predictions using MLP networks. The forward model predicts $\hat{z}_i^{t+1}$ (or z_i^{t+1}) given $\hat{z}_i^t$ (or z_i^t) and $\boldsymbol{a}^t$, while the inverse model predicts a_i^t given consecutive feature embeddings ($\hat{z}_i^t$, $\hat{z}_i^{t+1}$) or (z_i^t, z_i^{t+1}). These auxiliary objectives promote comprehensive, temporally coherent representations that better support adaptive communication in multi-agent scenarios.

The final loss for the overall offline training is a weighted sum of the contrastive losses and the auxiliary losses:

$$L_{total} = L_{supcon}^{GIA} + L_{supcon}^{FIA} + \alpha L_{recon} + \beta L_{dyn} \tag{3}$$

where α and β are weighting factors.

4.2 Online Training of Agent Policy

After offline pretraining, we integrate the *adaptive message generator* and *message-observation integrator* into the QMIX [16] framework for online policy training (illustrated in Figure 4). In this stage, each agent i processes its observation and receives messages (synthesized with the *adaptive message generator* from other agents) to generate an integrated representation z_i using the *message-observation integrator*. This representation z_i, together with the agent's last action, informs the agent's action selection via a GRU-based Q-network. The QMIX architecture then combines individual agents' Q-values through a mixing network conditioned on the global state to produce a centralized Q-value, which is used to compute the loss for training agent policies and the mixing network. Importantly, the parameters of the pre-trained *adaptive message generator* and *message-observation integrator* remain fixed during this online training phase and are not updated.

In the remaining sections, we present the experimental setup and results that evaluate the effectiveness of our methodology.

5 EXPERIMENTS

In this section, we report our evaluation of TACTIC. Our experiments aim to address the following key questions:

- Q1. Can the policy trained with the adaptive communication mechanism generalize across different sight ranges?
- Q2. Does the offline-trained communication mechanism enhance the efficiency of online policy training?
- Q3. How do data quality and varying loss terms impact the performance of training effectiveness and generalization?

5.1 Experimental Setup

Our experiments are conducted in the SMACv2 environment [4]. Compared to the original SMAC environment [17], SMACv2 incorporates increased stochasticity and meaningful partial observability,

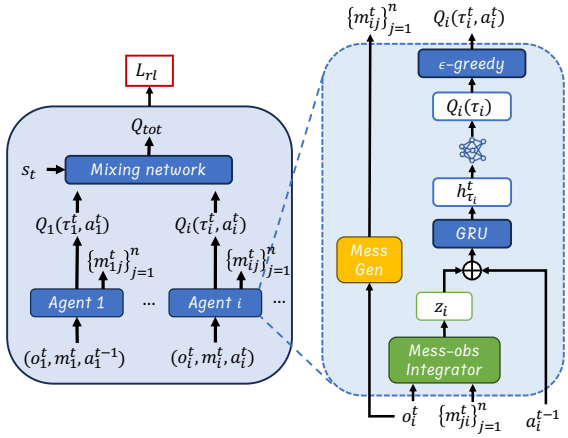

Figure 4: Online policy training pipeline of TACTIC, illustrating the integration of QMIX architecture with pre-trained communication components. The pre-trained message generator (Mess Gen) and message-observation integrator (Mess-obs Integrator) remain fixed during the policy training.

necessitating the development of complex closed-loop policies for effective agent coordination. We use three maps from SMACv2 for our experiments: Terran, Protoss, and Zerg. Each map features distinct unit types, with Terran units including Marines, Marauders, and Medivacs; Protoss units comprising Stalkers, Zealots, and Colossi; and Zerg units consisting of Zerglings, Hydralisks, and Banelings. Agents are generated procedurally, with varying numbers ranging from 5 to 20, and are assigned specific sight and attack ranges that enhance the complexity of the scenarios.

For each map, we consider three agent number configurations (5, 10, and 20 agents) and three sight-range ratios (0.2, 1, and 5). The sight-range ratios are applied by multiplying the agents' original sight ranges, which vary among different agents, to adjust their visibility accordingly. This allows us to assess the adaptability of the proposed communication mechanism under varying observability conditions.

For each environment setup specified by a combination of the map and the agent number configuration, we pre-train an offline communication mechanism (Section 4.1). This communication mechanism is then used for online policy training (Section 4.2) with a team of agents with fixed sight ranges. The learned policy is further evaluated on a variety of sight ranges (the sight range ratios are between 0.2 and 5) on the same environment setup.

To valid the effectiveness of TACTIC, the performance regarding the generalization across various sight ranges and the online training efficiency is compared to five baselines: QMIX [16], QMIX-Att [6], NDQ [28], TarMAC [2], and MASIA [5]. QMIX is one of the more commonly used CTDE MARL algorithms that does not use communication during execution. The others are four variants of QMIX with different communication mechanisms. We adopted the hyperparameters from the original implementations of the baselines and reused the hyperparameters of QMIX for TACTIC. All results in the following sections are from five independent runs with different random seeds.

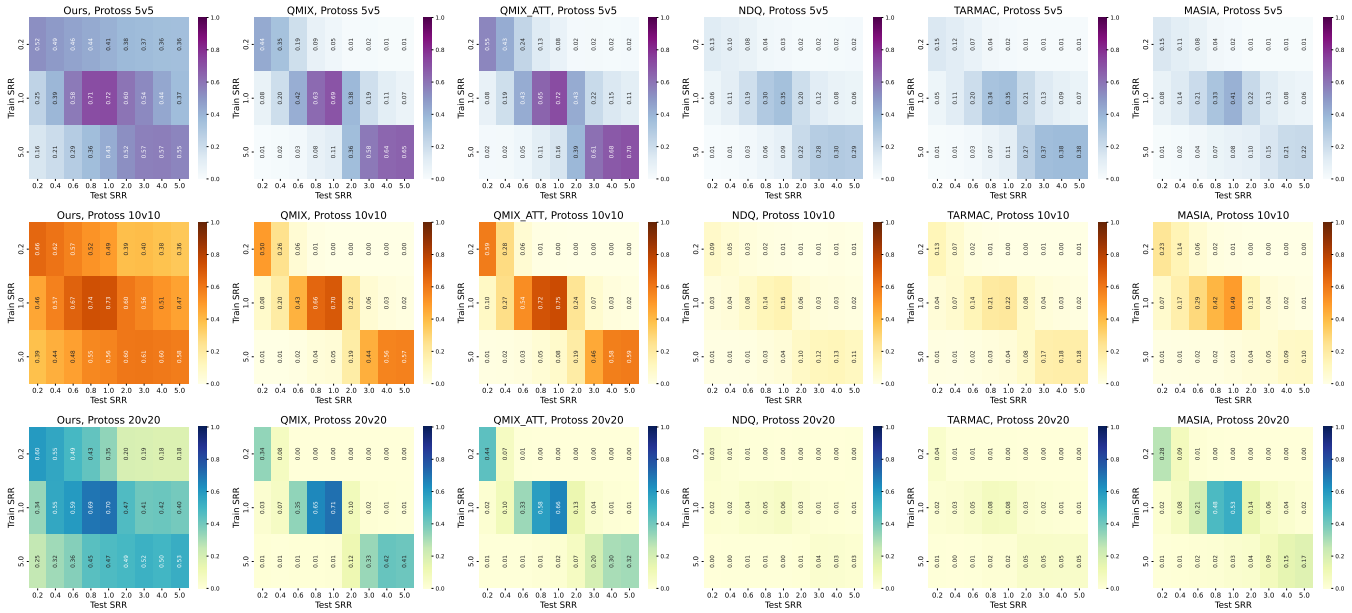

Figure 5: Performance of TACTIC and baseline models on policy generalizability across various sight ranges in the Protoss map.

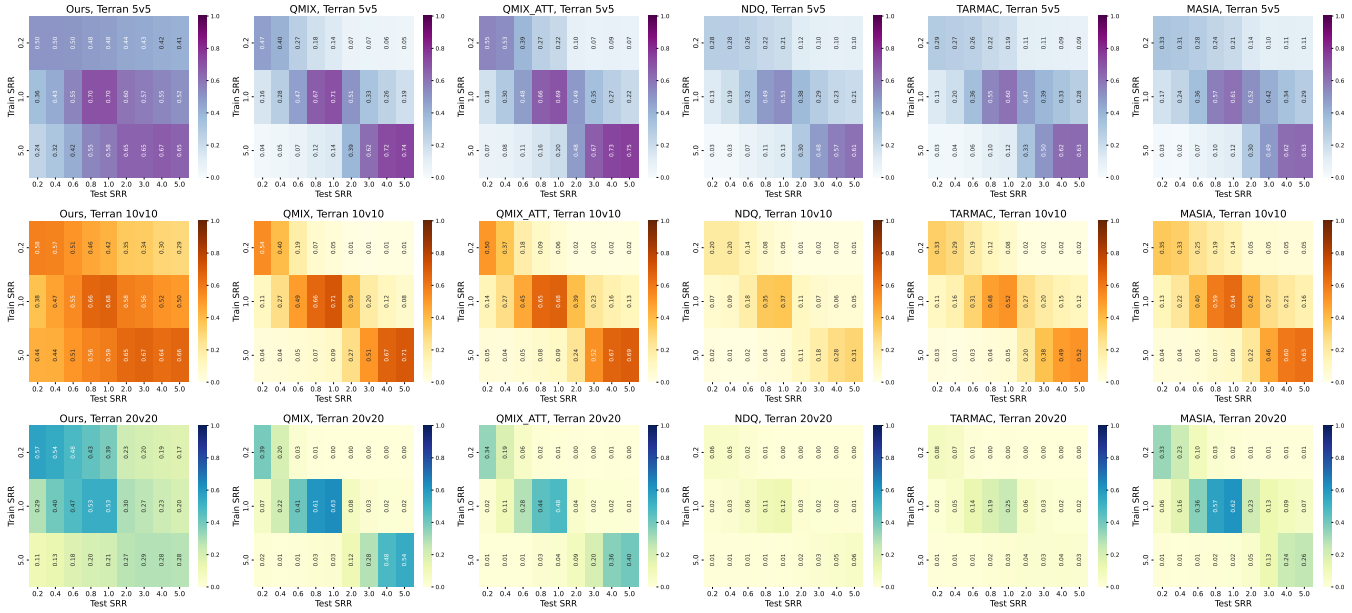

Figure 6: Performance of TACTIC and baseline models on policy generalizability across various sight ranges in the Terran map.

5.2 Policy Generalization Across Sight Ranges

To answer Q1, in this section, we report the generalization capability of our trained policies across various sight ranges. As introduced in Section 5.1, our policies are learned based on a pre-trained communication mechanism under the same environment setup with fixed sight ranges; they are then tested on a broader set of sight-range ratios valued from 0.2 to 5. Their performances (i.e., *mean battle won rate*) in all scenarios are recorded and visualized with heatmaps as shown in Figure 5, 6 and 7.

It can be clearly seen that our method TACTIC produces policies that demonstrate more robust performances in all the environment settings and across various sight-range ratios that were not seen during policy training. The five baselines can have reasonable performances only when the test and train sight ranges are

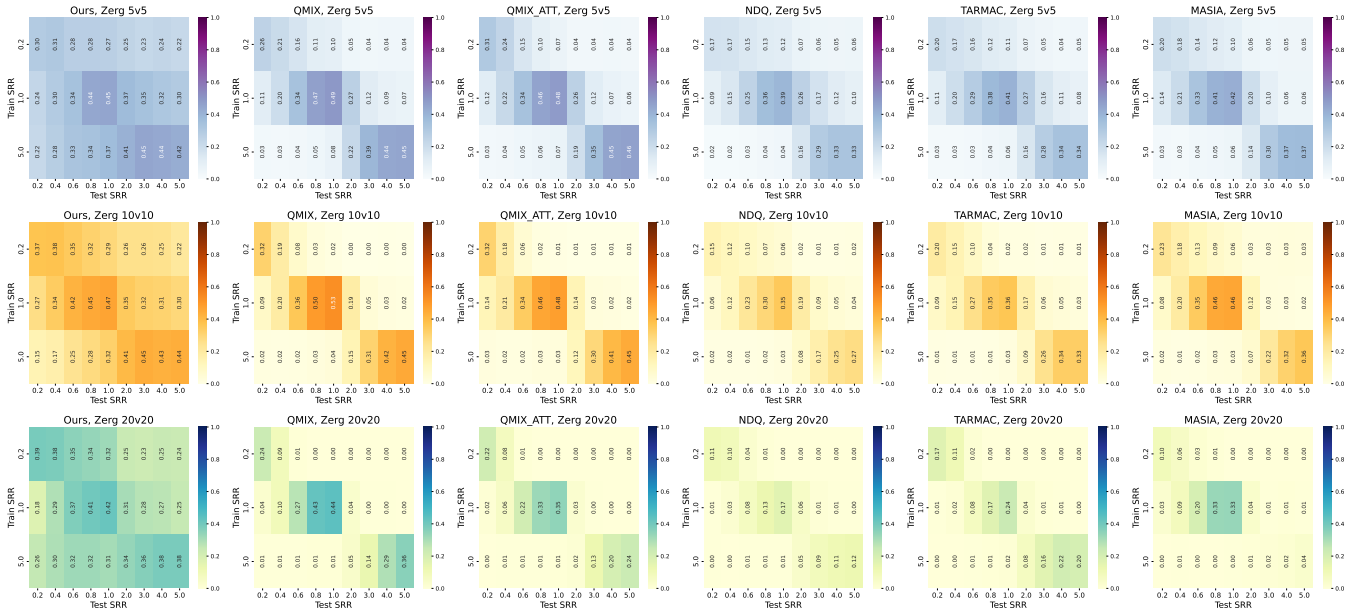

Figure 7: Performance of TACTIC and baseline models on policy generalizability across various sight ranges in the Zerg map.

close (referring to the diagonal cells of the heatmaps for the baselines), while policies learned by TACTIC are capable of maintaining satisfying performances even when the train and test sight ranges are different by large. For example, on the map Terran with 5 agents, policy trained with sight-range ratio 0.2 by TACTIC has a 0.41 *mean battle won rate* when tested with sight-range ratio 5, while the corresponding results for the five baselines are all under 0.10; it is also true if exchanging the train and test sight-range ratios or in other maps. With more agents introduced into the environment, baseline models become more strict on the difference between train and test sight ranges (i.e., their heatmaps become more sparse with more zero values), while our method TACTIC still generalizes well.

The observed phenomenon suggests that our adaptive communication mechanism that is learned offline helps agents better generate and interpret messages and develop flexible strategies to take actions to adapt to changing observability conditions, leading to **reusable policies** in environments with unseen sight ranges. However, for the five baseline approaches, one needs to *re-train* the model to get a usable policy if the observation conditions change.

5.3 Online Policy Training Efficiency

To answer Q2, we compare the online policy training efficiency of our method TACTIC against the five baselines. Figure 8 presents the learning curves (*mean battle won rate* versus *timesteps*) of all six algorithms on every combination of map type, number of agents, and the sight-range ratio.

It can be observed that in every environment setup, the convergence speed of TACTIC is superior to or at the same level as that of QMIX or QMIX-Att, and is always better than that of NDQ, TarMAC, and MASIA. Notably, when the sight-range ratio used during online policy training is small (i.e., 0.2), TACTIC demonstrates significantly better convergence speed compared to the baselines.

The results indicate that the offline-trained communication mechanism contributes positively to the training efficiency of online policy learning, especially in scenarios where the number of agents is large or the sight ranges of agents are small.

5.4 Ablation Study

In this section, we investigate Q3 by conducting ablation studies on the quality of the offline dataset $\mathcal{D}$ for training the communication mechanism and the impact of different loss terms in our objective function (Eq. 3). For this section, we focus on the Terran map and 5-agent setting.

Data quality. To examine the influence of data quality of $\mathcal{D}$ on the learned online policy, we train the communication mechanism twice on two types of offline datasets.

- `Exploratory`: Trajectories generated in the exploratory stage of QMIX training, with 6000 episodes per task.
- `Random`: Trajectories generated from random environment interactions (where actions are chosen uniformly at random for each agent), with 6000 episodes per task.

We then perform online policy learning twice using the two pre-trained communication mechanisms, obtaining two policies. They are evaluated in the same way as in Section 5.2. The performance is shown in Figure 9(a-b). It can be seen that there is a very slight performance decay when using the Random dataset compared to the Exploratory dataset, indicating that the offline training stage for the communication mechanism can still learn meaningful and informative representations given low-quality data.

Loss terms. We further examine the influence of the two auxiliary loss objectives, the reconstruction loss L_{recon} and the dynamic loss L_{dyn}, by training three communication mechanisms with the weighing factors set to $\alpha = 0, \beta = 1$, $\alpha = 1, \beta = 0$, and $\alpha = 0, \beta = 0$ in Eq. 3, respectively. The online policy learning and evaluation

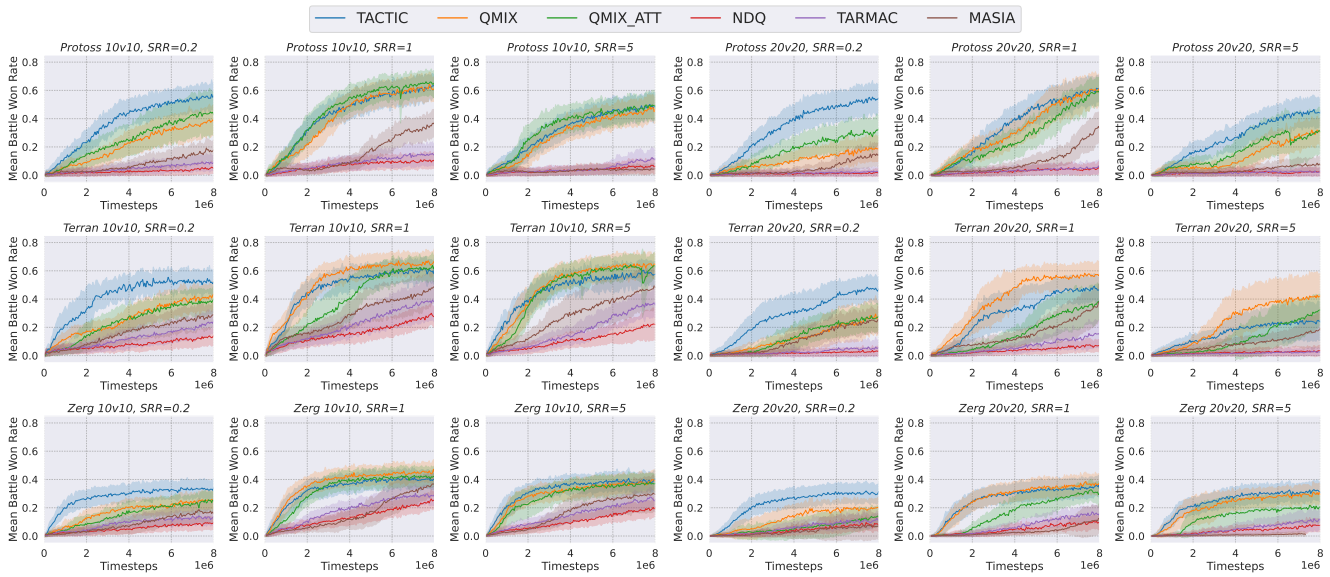

Figure 8: Training curves of the online policy learning stage in TACTIC and baseline models under different environment setups.

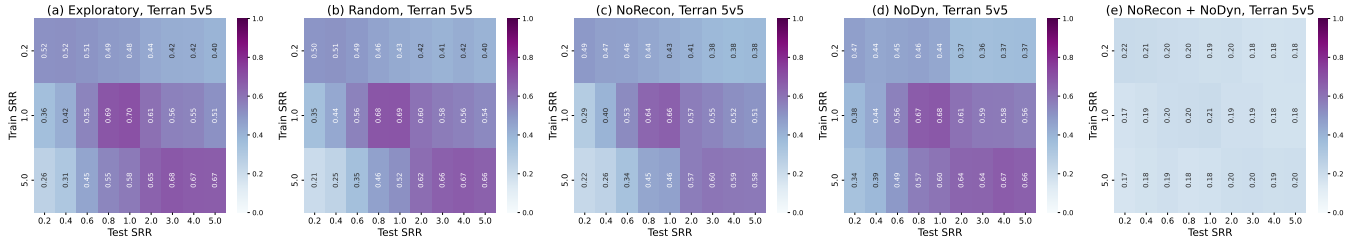

Figure 9: Performance of TACTIC on policy generalizability across various sight ranges under 5 different offline training schemas. The environment setup is the Terran map with a 5v5 agent configuration. (a) Offline training with a `Exploratory` dataset; (b) Offline training with a `Random` dataset; (c) Offline training without L_{recon}; (d) Offline training without L_{dyn}; (e) Offline training without L_{recon} and L_{dyn}.

remain the same as the data-quality ablation study. The results are shown in Figure 9(c-e).

In the first two cases where either the L_{recon} or L_{dyn} is dropped, there are no significant performance decays. Specifically, both terms can help the offline stage learn informative representations individually, while L_{recon} contributes slightly more to the overall performance than L_{dyn}. However, when both terms are dropped, a significant performance decay is observed as shown in Figure 9(e). This demonstrates the necessity of adopting at least one of the auxiliary loss objectives, and the combination of the two leads to the most effective representation learning in the offline stage.

6 CONCLUSION

In this paper, we introduce TACTIC, a communication mechanism for improving the generalizability of MARL systems. Specifically, we focus on generalizing the policy to scenarios where the sight range of the agents during execution may not be the same as that during training. The communication mechanism in TACTIC is trained offline and is task-agnostic. Utilizing contrastive loss allows agents to effectively align the integration of local observation and incoming messages with the egocentric state, leading to better situational awareness and improved coordination.

We show that standard benchmark MARL techniques with and without inter-agent communication generalize poorly when the sight range changes. In contrast, we show TACTIC generalizes better across sight ranges. Notably, we find that even random exploration trajectories can be leveraged to learn an effective communication strategy. Our findings suggest that TACTIC provides a robust framework to enhance inter-agent communication, leading towards more adaptive MARL systems.

ACKNOWLEDGMENTS

This work is supported in part by National Science Foundation Grant No. 1943368, Army Grant No. W911NF2120076, and UMD-Northrop Grumman seed grant.

REFERENCES

[1] Lucian Busoniu, Robert Babuska, and Bart De Schutter. 2008. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 38, 2 (2008), 156–172.

[2] Abhishek Das, Théophile Gervet, Joshua Romoff, Dhruv Batra, Devi Parikh, Mike Rabbat, and Joelle Pineau. 2019. Tarmac: Targeted multi-agent communication. In *International Conference on Machine Learning*. PMLR, 1538–1546.

[3] Yali Du, Bo Liu, Vincent Moens, Ziqi Liu, Zhicheng Ren, Jun Wang, Xu Chen, and Haifeng Zhang. 2021. Learning Correlated Communication Topology in Multi-Agent Reinforcement Learning. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 456–464.

[4] Benjamin Ellis, Jonathan Cook, Skander Moalla, Mikayel Samvelyan, Mingfei Sun, Anuj Mahajan, Jakob Foerster, and Shimon Whiteson. 2024. Smacv2: An improved benchmark for cooperative multi-agent reinforcement learning. *Advances in Neural Information Processing Systems* 36 (2024).

[5] Cong Guan, Feng Chen, Lei Yuan, Chenghe Wang, Hao Yin, Zongzhang Zhang, and Yang Yu. 2022. Efficient Multi-agent Communication via Self-supervised Information Aggregation. *Advances in Neural Information Processing Systems* 35 (2022), 1020–1033.

[6] Jian Hu, Siyang Jiang, Seth Austin Harding, Haibin Wu, and Shih wei Liao. 2021. Rethinking the Implementation Tricks and Monotonicity Constraint in Cooperative Multi-Agent Reinforcement Learning. (2021). arXiv:2102.03479 [cs.LG]

[7] Shengchao Hu, Li Shen, Ya Zhang, and Dacheng Tao. 2024. Learning multi-agent communication from graph modeling perspective. *arXiv preprint arXiv:2405.08550* (2024).

[8] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised Contrastive Learning. In *Advances in Neural Information Processing Systems*, Vol. 33. 18661–18673.

[9] Tianxu Li, Kun Zhu, Nguyen Cong Luong, Dusit Niyato, Qihui Wu, Yang Zhang, and Bing Chen. 2022. Applications of multi-agent reinforcement learning in future internet: A comprehensive survey. *IEEE Communications Surveys & Tutorials* 24, 2 (2022), 1240–1279.

[10] Toru Lin, Jacob Huh, Christopher Stauffer, Ser Nam Lim, and Phillip Isola. 2021. Learning to ground multi-agent communication with autoencoders. *Advances in Neural Information Processing Systems* 34 (2021), 15230–15242.

[11] Yen-Cheng Liu, Junjiao Tian, Chih-Yao Ma, Nathan Glaser, Chia-Wen Kuo, and Zsolt Kira. 2020. Who2com: Collaborative perception via learnable handshake communication. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 6876–6883.

[12] Y. L. Lo, B. Sengupta, J. Foerster, and M. Noukhovitch. 2024. Learning Multi-Agent Communication with Contrastive Learning. *arXiv preprint arXiv:2307.01403* (2024).

[13] Yaru Niu, Rohan R Paleja, and Matthew C Gombolay. 2021. Multi-Agent Graph-Attention Communication and Teaming. In *Proceedings of the AAMAS Conference*, Vol. 21. International Foundation for Autonomous Agents and Multiagent Systems.

[14] Frans A Oliehoek, Christopher Amato, et al. 2016. *A concise introduction to decentralized POMDPs*. Vol. 1. Springer.

[15] Kaige Qu, Weihua Zhuang, Qiang Ye, Wen Wu, and Xuemin Shen. 2024. Model-Assisted Learning for Adaptive Cooperative Perception of Connected Autonomous Vehicles. *IEEE Transactions on Wireless Communications* 23, 8 (2024), 8820–8835. https://doi.org/10.1109/TWC.2024.3354507

[16] Tabish Rashid, Mikayel Samvelyan, Christian Schroeder de Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. 2018. QMIX: Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning. arXiv e-prints, page. *arXiv preprint arXiv:1803.11485* (2018).

[17] Mikayel Samvelyan, Tabish Rashid, Christian Schroeder De Witt, Gregory Farquhar, Nantas Nardelli, Tim GJ Rudner, Chia-Man Hung, Philip HS Torr, Jakob Foerster, and Shimon Whiteson. 2019. The starcraft multi-agent challenge. *arXiv preprint arXiv:1902.04043* (2019).

[18] Esmaeil Seraj, Zheyuan Wang, Rohan Paleja, Matthew Sklar, Anirudh Patel, and Matthew Gombolay. 2021. Heterogeneous graph attention networks for learning diverse communication. *arXiv preprint arXiv:2108.09568* (2021).

[19] Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. 2016. Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv preprint arXiv:1610.03295* (2016).

[20] Jianzhun Shao, Hongchang Zhang, Yun Qu, Chang Liu, Shuncheng He, Yuhang Jiang, and Xiangyang Ji. 2023. Complementary Attention for Multi-Agent Reinforcement Learning. (2023).

[21] Jennifer She, Jayesh K Gupta, and Mykel J Kochenderfer. 2022. Agent-time attention for sparse rewards multi-agent reinforcement learning. *arXiv preprint arXiv:2210.17540* (2022).

[22] Siqi Shen, Yongquan Fu, Huayou Su, Hengyue Pan, Peng Qiao, Yong Dou, and Cheng Wang. 2021. Graphcomm: A graph neural network based method for multi-agent reinforcement learning. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 3510–3514.

[23] A. Singh, T. Jain, and S. Sukhbaatar. 2019. Learning When to Communicate at Scale in Multiagent Cooperative and Competitive Tasks. In *International Conference on Learning Representations*.

[24] Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus. 2016. Learning multiagent communication with backpropagation. In *Proceedings of the 30th International Conference on Neural Information Processing Systems* (Barcelona, Spain) *(NIPS'16)*. Curran Associates Inc., Red Hook, NY, USA, 2252–2260.

[25] Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus. 2016. Learning multiagent communication with backpropagation. In *Advances in Neural Information Processing Systems*. 2244–2252.

[26] Chuangchuang Sun, Macheng Shen, and Jonathan P How. 2020. Scaling up multiagent reinforcement learning for robotic systems: Learn an adaptive sparse communication graph. In *2020 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 11755–11762.

[27] A. van den Oord, Y. Li, and O. Vinyals. 2018. Representation Learning with Contrastive Predictive Coding. *arXiv preprint arXiv:1807.03748* (2018).

[28] Tonghan Wang, Jianhao Wang, Chongyi Zheng, and Chongjie Zhang. 2019. Learning nearly decomposable value functions via communication minimization. *arXiv preprint arXiv:1910.05366* (2019).

[29] Zhiwei Xu, Bin Zhang, Dapeng Li, Zeren Zhang, Guangchong Zhou, Hao Chen, and Guoliang Fan. 2023. Consensus Learning for Cooperative Multi-Agent Reinforcement Learning. *Proceedings of the AAAI Conference on Artificial Intelligence* 37, 10 (Jun. 2023), 11726–11734. https://doi.org/10.1609/aaai.v37i10.26385

[30] Erfu Yang and Dongbing Gu. 2004. *Multiagent reinforcement learning for multi-robot systems: A survey*. Technical Report. tech. rep.

[31] Kejun Zhang, Jayesh K Gupta, Alfred O Hero III, and Mykel J Kochenderfer. 2021. Succinct and robust multi-agent communication with temporal message control. *arXiv preprint arXiv:2107.06609* (2021).

[32] Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. 2021. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of Reinforcement Learning and Control* (2021), 321–384.

[33] S. Q. Zhang, Q. Zhang, and J. Lin. 2019. Efficient Communication in Multi-Agent Reinforcement Learning via Variance Based Control. In *Proceedings of the NeurIPS 32nd Conference on Neural Information Processing Systems*, Vol. 32. Curran Associates, Inc.

[34] Zhi Zhang, Jiachen Yang, and Hongyuan Zha. 2019. Integrating independent and centralized multi-agent reinforcement learning for traffic signal network optimization. *arXiv preprint arXiv:1909.10651* (2019).

[35] Ruijie Zheng, Xiyao Wang, Yanchao Sun, Shuang Ma, Jieyu Zhao, Huazhe Xu, Hal Daumé, and Furong Huang. 2024. TACO: temporal latent action-driven contrastive loss for visual reinforcement learning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA) *(NIPS '23)*. Curran Associates Inc., Red Hook, NY, USA, Article 2092, 23 pages.