Why and Whom to Communicate? A Dual-Objective, Cost–Benefit Framework for Multi-Agent Communication

Yi-Yu Lin
The University of Manchester
Greater Manchester, United Kingdom
yi-yu.lin@manchester.ac.uk

Jiahao Zhang
Huawei Technologies Ltd.
Shanghai, China
zjiahao228@gmail.com

Xiao-Jun Zeng
The University of Manchester
Greater Manchester, United Kingdom
x.zeng@manchester.ac.uk

ABSTRACT

Effective communication in Multi-Agent Systems hinges on a fundamental trade-off between task performance and communication cost. Prevailing methods often rely on heuristics or fixed penalties that require manual tuning to enforce this balance, treating communication cost as a secondary concern and leaving the rationale for *why* to communicate under-specified and *whom* to communicate with lacks a principled basis. To address this gap, we formalise Multi-Agent Communication as a dual-objective problem that jointly maximises long-term reward and minimises communication usage. We introduce **COBE** (**CO**st-**BE**nefit Communication), a framework in which agents proactively share intended actions only when the anticipated gain in long-term value justifies the inherent cost. COBE's core innovation is a data-driven process that constructs an empirical Pareto front over outcome profiles which map the trade-off between a stable, forward-looking critic value and communication usage. A curvature-based knee-point selection then identifies an operating point that encodes the steepest marginal value improvement per unit cost, serving as a dynamic target to guide policy learning. This mechanism eliminates the need for manual weight or budget tuning, allowing adaptive communication strategies to emerge. Experiments demonstrate that COBE outperforms state-of-the-art methods, learning efficient and context-aware communication patterns that achieve a better balance between collective performance and communication overhead.

KEYWORDS

Multi-Agent Communication; Multi-Agent System; Multi-Agent Deep Reinforcement Learning; Reinforcement Learning

ACM Reference Format:
Yi-Yu Lin, Jiahao Zhang, and Xiao-Jun Zeng. 2026. Why and Whom to Communicate? A Dual-Objective, Cost–Benefit Framework for Multi-Agent Communication. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 9 pages. https://doi.org/10.65109/TAYO7266

1 INTRODUCTION

Effective communication is a cornerstone of cooperation in Multi-Agent Systems (MAS), enabling agents to coordinate under partial observability and adapt to dynamic environments. Yet, practical constraints like bandwidth and energy create a fundamental trade-off between task performance and communication overhead. Navigating this dynamic to learn efficient and effective communication strategies remains a key challenge in multi-agent learning.

Recent works in Multi-Agent Communication (MAC) have sought to suppress unnecessary messaging, but often rely on heuristics or fixed penalties that require manual tuning [32, 43, 45, 49]. This yields inflexible strategies where *whom* an agent contacts is dictated by a fixed topology, *when* it communicates is governed by context-poor thresholds, and *what* is transmitted lacks foresight. Crucially, such approaches lack a principled account of *why* communication should be initiated. They fail to balance strategic value against inherent cost, treating communication as a secondary concern rather than a primary objective to be optimised.

To address these limitations, we introduce **COBE** (**CO**st-**BE**nefit Communication), a framework that formalises the communication decision as a cost-benefit analysis grounded in social exchange theory [3, 4], with the goal of maximising reward while minimising communication cost. In this framework, agents learn a proactive policy to share *what* is most crucial, their intended action, only when its anticipated benefit justifies the inherent cost. The decision of *when* and *why* to communicate is informed by the principles of marginal utility theory. This is realised through a novel data-driven mechanism that learns the empirical Pareto front of the value-cost trade-off and algorithmically selects its knee point via curvature analysis. This optimal operating point then guides the policy to dynamically select *whom* to contact, resulting in an efficient and adaptive communication topology that emerges holistically without manual tuning. Our primary contributions are:

- **Dual-Objective Formulation.** We formalise MAC as a principled bi-criteria problem, treating task performance and communication cost as competing primary objectives. This moves beyond prevailing single-objective methods that rely on manually-tuned penalties or budgets.
- **Data-Driven Learning.** We introduce a method to learn the empirical Pareto front of the cost-benefit trade-off. Unlike prior work which presumes a fixed preference, our approach discovers the landscape of optimal, non-dominated solutions.
- **Preference-Free Policy Guidance.** We propose an automated mechanism for policy guidance via curvature-based knee-point selection. This algorithmic approach identifies the point of greatest marginal return, eliminating the need for brittle hyperparameter tuning.
- **Emergent and Proactive Communication.** Our framework cultivates a dynamic policy where the principled *why* informs the adaptive *whom* and *when* of communication, overcoming the limitations of fixed topologies or heuristic-based gating.

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/TAYO7266

- **Comprehensive Empirical Validation.** We provide extensive evidence that our approach achieves a more efficient performance-cost trade-off than state-of-the-art baselines and is more robust than tuned single-objective methods, demonstrating a superior and more stable learning process.

2 RELATED WORK

2.1 *Whom* and *When*: Learning the Structure

A significant body of research addresses the structural aspects of communication, moving beyond the excessive overhead of early broadcasting protocols such as RIAL and DIAL [11]. Gating mechanisms like IC3Net [35], MBC [16], and CACOM [22] determine when agents should communicate. Scheduling-based approaches such as SchedNet [20], MAGIC [30], and SeqComm [6] train modules to select which agents may communicate at each timestep. Attention-based methods provide a granular answer to who to communicate with. ATOC [19] and TarMAC [5] use attention to weigh peer importance, a principle also seen in the relevance heuristics of FlowComm [10]. A prominent approach involves modelling the agent network explicitly using Graph Neural Networks, as seen in DGN [18], GA-Comm [25], and CommFormer [17]. More sophisticated structural learning includes the hierarchical teams in TeamComm [39], the personalised patterns of PMAC [28], and the structured request-reply protocol of I2C [8]. Theoretical work such as GDN [29] further characterises the expressiveness of such topology-constrained systems. While powerful, the criteria for forming a connection often rely on local heuristics or on impractical centralised controllers, rather than the principle of cost-benefit efficiency.

2.2 *What*: Optimising the Content

A parallel line of research focuses on optimising what agents communicate, recognising that message content is as important as its destination. Information-theoretic compression offers a principled way to improve content. MAGI [7] uses a graph information bottleneck to learn maximally informative yet concise message representations. Other methods focus on the semantic content of messages. IDEAL [9] enhances messages with agent identities. LToS [40] enables the sharing of local goals. HAMMER [15] uses symbolic plans to guide coordination. Similarly, NeurComm [2] and GAXNet [44] embed environment-specific priors into message content. A further dimension is message quality and processing. ADMAC [42] and T2MAC [37] introduce estimators for reliability and trustworthiness. TMC [47] applies temporal smoothing to reduce redundancy over time. Other techniques like Variable-Length Coding [13] and the permutation invariance of MASIA [14] also contribute to content efficiency. However, these approaches to content optimisation are often decoupled from a principled optimisation of the overall communication budget.

2.3 *Why*: The Missing Principled Answer

Despite these advances, a fundamental question often remains implicit. Why should a communication act be performed at all? Answering this requires a framework that formally balances the potential gain against the communication overhead. A few reward-guided methods have begun to address this. RGIC [23], VBC [46], and CAROC [24] encourage messaging only when it is predicted to improve an agent's expected return. RGMComm [1] takes a similar path and aims to communicate in a way that minimises the return gap. These methods identify the link, yet they still treat the decision as a heuristic trigger within a single-objective paradigm, rather than the trade-off between two competing objectives.

3 METHODOLOGY

3.1 Problem Formulation

We consider a cooperative Multi-Agent Deep Reinforcement Learning (MADRL) setting under partial observability, represented as a Decentralised Partially Observable Markov Decision Process (Dec-POMDP) [31]. The Dec-POMDP is formally described by the tuple $\mathcal{M} = (\mathcal{I}, \mathcal{S}, \mathcal{A}, \mathcal{O}, \mathrm{Tr}, R, \gamma)$, where $\mathcal{I}$ is the set of N agents, $\mathcal{S}$ is the state space, $\mathcal{A} = \prod_{i\in\mathcal{I}} \mathcal{A}_i$ is the joint action space, $\mathcal{O} = \prod_{i\in\mathcal{I}} \mathcal{O}_i$ is the joint observation space, and $\gamma \in [0, 1)$ is the discount factor. At each timestep t, each agent $i \in \mathcal{I}$ receives a local observation $o_i^t \sim O_i(\cdot \mid s^t)$ from the global state $s^t \in \mathcal{S}$. Agents select actions according to decentralised policies $\pi_{\theta_i}(a_i^t \mid h_i^t)$, where h_i^t is the agent's internal state, a function of its observation–action history and any received communication messages, m_i^t. The joint policy factorises as $\pi_\theta(a^t \mid o^t, m^t) = \prod_{i\in\mathcal{I}} \pi_{\theta_i}(a_i^t \mid h_i^t)$. The joint action $a^t = \{a_1^t, \ldots, a_N^t\}$ yields a transition $s^{t+1} \sim \mathrm{Tr}(\cdot \mid s^t, a^t)$ and a shared global reward $r^t = R(s^t, a^t)$. The objective is to maximise the expected discounted return over a finite horizon T, $\mathbb{E}\left[\sum_{t=0}^{T-1} \gamma^t r^t\right]$.

To improve stability, we adopt the Centralised Training with Decentralised Execution (CTDE) paradigm [26, 36] and employ a multi-agent actor-critic framework [12, 21]. Each agent learns its policy π_{θ_i}, while a centralised value critic, parameterised by ψ, learns a state–value function $V_\psi(s^t)$. In practice, where the true state s^t is unavailable, we overload s^t to denote the CTDE surrogate formed from joint observations [27, 41].

3.2 Cost-Benefit Objective Formulation

Effective MAC presents a fundamental trade-off between maximising task performance and minimising communication overhead. The two predominant paradigms for managing this can both be expressed as single-objective problems. Scalarisation seeks to maximise a weighted combination of reward and cost using a preference weight λ, $\max_{\pi_\phi} \left[\hat{r}(\pi_\phi) - \lambda\hat{\zeta}(\pi_\phi)\right]$, where $\hat{r}(\pi_\phi)$ and $\hat{\zeta}(\pi_\phi) \in [0, 1]$ are the expected mean return and communication usage, respectively, achieved when using a given communication policy π_ϕ, which is shorthand for the peer-wise decision function $\pi_\phi(h^t, j)$ that determines whether to communicate with a specific neighbour j. Constrained optimisation can be solved via a Lagrangian, which uses a learnable parameter μ to enforce a specified budget ϵ, $\max_{\pi_\phi} \left[\hat{r}(\pi_\phi) - \mu(\hat{\zeta}(\pi_\phi) - \epsilon)\right]$. Both methods, however, are problematic as they rely on a manually specified hyperparameter, that imposes a static preference and obscures the true underlying dynamics of the cost-benefit trade-off.

To address this, our framework learns a dedicated communication policy, π_ϕ, using a formulation that preserves the independence of the two competing goals. We explicitly seek a communication policy that optimises the trade-off between reward and cost, which

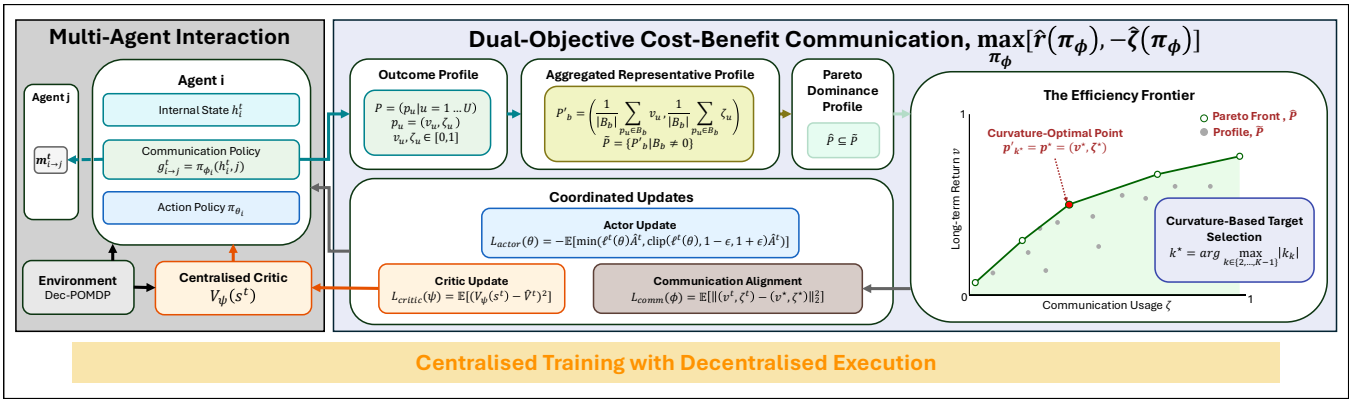

Figure 1: Overview of the COBE architecture.

we express with the following concise, vector-valued objective:

$$\max_{\pi_\phi} \left[\hat{r}(\pi_\phi), -\hat{\zeta}(\pi_\phi)\right]. \tag{1}$$

This approach enables a preference-free operating-point selection, as it learns the entire cost-benefit frontier before algorithmically identifying a balanced solution, thereby avoiding the need to specify either a weight λ or a budget ϵ a priori.

Definition 1 (Communication Usage). *For a decentralised communication policy, where each agent $i \in \mathcal{I}$ has a policy π_{ϕ_i} that produces a per-link send probability, $g_{i\to j}^t \in [0,1]$, for each potential communication channel from agent i to agent $j \in \mathcal{I}, (i \neq j)$. The communication usage, $\zeta^t \in [0,1]$, is defined as the average of these send probabilities over all potential communication links at timestep t. The expected communication probability, $\hat{\zeta}(\pi_\phi)$, where $\phi = \{\phi_i\}_{i\in\mathcal{I}}$, represents the mean probability over all channels and timesteps.*

To guide the learning of the communication policy, a stable, forward-looking signal for performance is required, as the instantaneous reward r^t is often sparse and noisy. COBE resolves this by employing the centralised critic's value estimate, $V_\psi(s^t)$, as a robust proxy for long-term benefit. While a naive critic can suffer from instability and arbitrary scaling, which would corrupt the Pareto analysis, our framework incorporates a principled normalisation protocol to address this challenge. We employ PopArt standardisation to stabilise the critic's learning and adaptively rescale its outputs, thereby preventing erratic fluctuations. As communication usage, ζ^t, is inherently on a $[0,1]$ scale, we then map the critic's standardised value to the same bounded range using a sigmoid function. This step ensures both objectives are on a commensurate scale, which is essential to prevent the value estimate from disproportionately influencing policy updates. For notational simplicity, we denote this bounded, forward-looking benefit as v^t.

Definition 2 (Outcome Profile). *An outcome profile is a tuple defined at each timestep as $\mathbf{p}^t = (v^t, \zeta^t)$. The first element, $v^t \in [0,1]$, is the forward-looking benefit represented by the bounded critic value, while the second, $\zeta^t \in [0,1]$, is the associated communication cost captured by the mean communication usage.*

During each training iteration, we populate a rollout buffer by executing the current policy across multiple parallel environments. This process yields an outcome profile $\mathbf{p}_e^t = (v_e^t, \zeta_e^t)$ for each environment e at each timestep t. For the purpose of analysis, we aggregate all of these profiles into a single collection, formally expressed as:

$$\mathcal{P} = \{\mathbf{p}_u = (v_u, \zeta_u) \mid u = 1, \ldots, U\}, \tag{2}$$

where each $\mathbf{p}_u$ corresponds to a unique (e,t) pair, and the total number of profiles, U, is the product of the number of parallel environments and the rollout length. The aggregated set $\mathcal{P}$ provides a comprehensive sample of the system's learning dynamics, revealing the evolving cost-benefit landscape as agent strategies are refined.

To ensure this dual-objective trade-off is explored broadly, we anneal the entropy bonus for the communication policy. A higher entropy coefficient at the start of training encourages stochasticity, leading to a wider exploration of communication probabilities and a well-populated set of outcome profiles. For consistency, we couple the straight-through Gumbel–Sigmoid temperature τ to this same coefficient so that, as the exploration bonus decays, τ decreases in lockstep and gate sharpness increases without a separate schedule. This joint annealing gradually drives the policy toward a more deterministic strategy.

3.3 Learning the Efficiency Frontier

In decentralised MAS the distribution $\mathcal{P}$ of outcome profiles observed during training exhibits inherent high variability due to exploration noise, partial observability, and non-stationary learning dynamics. As a result, many profiles are noisy, transient, or suboptimal. Directly reasoning over these raw samples can misrepresent the cost-benefit trade-off surface.

To address this challenge, we employ a principled, two-stage process to derive a stable estimate of the efficiency frontier. The first stage utilises a non-parametric smoothing technique based on adaptive data aggregation. Let U be the number of outcome profiles in the current on-policy minibatch. We dynamically set the number of equal-width interval bins, B, for the $\zeta \in [0,1]$ communication axis according to the square-root rule, a common heuristic in histogram construction:

$$B = \text{clip}(\lfloor\sqrt{U}\rfloor, B_{\min}, B_{\max}). \tag{3}$$

This approach scales the resolution sub-linearly with the number of data points, balancing the trade-off between statistical variance and bias. The bounds $B_{\min}$ and $B_{\max}$ help ensure stability of the frontier estimation. The lower bound, $B_{\min} = 6$, guarantees a sufficient number of representative points to compute a meaningful curvature. A coarser granularity would risk high bias, potentially obscuring the true trade-off landscape. Conversely, the upper bound, $B_{\max} = 20$, prevents an excessively large number of bins that would exhibit high statistical variance due to low sample counts. This cap ensures the aggregated profiles remain stable across training iterations.

The outcome profiles, $\mathbf{p}_u = (v_u, \zeta_u)$, are then grouped into their corresponding bins. $\mathcal{B}_b$ denotes the set of profiles assigned to bin b:

$$\mathcal{B}_b = \{\mathbf{p}_u \in \mathcal{P} \mid \zeta_u \in I_b\}, \tag{4}$$

where $[0, 1]$ is partitioned into equal-width, disjoint intervals $I_b = [\frac{b-1}{B}, \frac{b}{B})$ for $b = 1, \ldots, B-1$ and $I_B = [\frac{B-1}{B}, 1]$, ensuring complete coverage of ζ without boundary duplication. To ensure statistical robustness, we discard any bin with fewer than $\max(3, \lfloor 0.05U \rfloor)$ outcome profiles. For each remaining bin, we compute an aggregated representative profile:

$$\mathbf{p}'_b = \left(\frac{1}{|\mathcal{B}_b|} \sum_{\mathbf{p}_u \in \mathcal{B}_b} v_u, \quad \frac{1}{|\mathcal{B}_b|} \sum_{\mathbf{p}_u \in \mathcal{B}_b} \zeta_u \right). \tag{5}$$

This adaptive aggregated process produces a smoothed and robust set of representative profiles, forming a reliable foundation for the subsequent Pareto analysis.

The second stage applies the principle of Pareto dominance to the set of smoothed profiles, $\tilde{\mathcal{P}} = \{\mathbf{p}'_b \mid \mathcal{B}_b \neq \emptyset\}$.

Definition 3 (Pareto Dominance and Pareto Front). *An outcome $\mathbf{p}'_i = (v_i, \zeta_i)$ Pareto-dominates another outcome $\mathbf{p}'_j = (v_j, \zeta_j)$ if $v_i \geq v_j$ and $\zeta_i \leq \zeta_j$, with at least one inequality being strict. The Pareto front, denoted $\hat{\mathcal{P}} \subseteq \tilde{\mathcal{P}}$, is the set of all profiles in $\tilde{\mathcal{P}}$ that are not dominated by any other profile [38].*

This two-stage process confers several critical benefits. First, the aggregation step is computationally efficient, reducing the complexity from a quadratic operation on the raw samples, $O(|\mathcal{P}|^2)$, to a more tractable process involving a linear-time aggregation pass and a subsequent dominance check on a small, bounded set of representative profiles. It also suppresses the variance induced by stochastic training updates. Second, the subsequent application of Pareto optimisation ensures that only statistically meaningful, non-dominated strategies are selected for consideration. This is a key distinction from prior methods, as it circumvents the need for hyperparameters that encode a static preference, such as a fixed scalarisation weight or a predetermined communication budget, on the cost-benefit trade-off. Collectively, this construction yields a robust, interpretable, and computationally tractable framework for characterising the cost-benefit geometry of the MAS.

3.4 Policy Optimisation with Pareto Guidance

Policy optimisation is guided by the established cost-benefit frontier. The process involves selecting a dynamic target from the front via curvature analysis, aligning the communication policy with this target using a specialised loss, and jointly updating the actor and critic to realise an efficient and coordinated strategy.

3.4.1 Curvature-Based Target Selection. The Pareto front, $\hat{\mathcal{P}}$, represents a set of optimal, non-dominated solutions. To provide a single, concrete objective for policy optimisation, we must select one profile from this set to act as a balanced solution and dynamic target, $\mathbf{p}^\star$. A principled, preference-free approach is to identify the *knee* or *elbow* of the front. This point operationalises the principle of diminishing returns by identifying where the steepest marginal gain in performance occurs per unit of communication cost. It is found by maximising the discrete curvature, thereby providing a robust, data-driven method for trade-off selection without prescribing a manual scalarisation weight λ or communication budget ϵ.

Criterion 1 (Curvature-Based Target Selection). *Let $\hat{\mathcal{P}} = \{\mathbf{p}'_1, \ldots, \mathbf{p}'_K\}$ be non-dominated profiles on the Pareto front, sorted by ascending communication usage. Both v and ζ axes are bounded in $[0, 1]$ to prevents axis-scale bias, ensuring neither dimension disproportionately influences the curvature computation or target selection.*

In alignment with standard knee-detection literature, we select the target profile from the interior of the front, as the knee reflects a change in behaviour within the series of trade-offs, not at its extremes [33, 34, 48]. For fronts with three or more points, we compute the discrete curvature for all interior indices $k \in \{2, \ldots, K-1\}$:

$$\kappa_k = \frac{(v_{k+1} - v_k)(\zeta_k - \zeta_{k-1}) - (v_k - v_{k-1})(\zeta_{k+1} - \zeta_k)}{[(v_{k+1} - v_{k-1})^2 + (\zeta_{k+1} - \zeta_{k-1})^2]^{3/2}}. \tag{6}$$

The selected target profile, $\mathbf{p}^\star = (v^\star, \zeta^\star)$, corresponds to the interior point with the highest absolute curvature:

$$\mathbf{p}^\star = \mathbf{p}'_{k^\star} \quad \textit{where} \quad k^\star = \arg \max_{k \in \{2, \ldots, K-1\}} |\kappa_k|. \tag{7}$$

In the edge case where the front has fewer than three points, preventing the calculation of an interior knee, we fall back to a deterministic lexicographic rule: select the profile that maximises the value v, breaking ties by minimising the mean communication usage ζ. This curvature-based criterion follows classical knee detection practice, adapted for a smoothed, discrete Pareto front.

The efficacy of this criterion relies on the smoothed approximation of the cost-benefit landscape derived in the previous section. By reducing sensitivity to the noisy and irregular profiles generated during training, our preprocessing allows the curvature-based selection to remain a robust and well-defined method. This provides a stable reference target for the policy optimisation that follows.

3.4.2 Pareto-Guided Policy Alignment. With a dynamic target profile, $\mathbf{p}^\star = (v^\star, \zeta^\star)$, selected from the Pareto front, we operationalise the bi-criteria goal. We introduce a loss function that provides a strategic learning signal to the communication policy, π_ϕ, guiding it to produce outcomes that align with the efficient target. This is accomplished by minimising the squared Euclidean distance between the outcome profile observed at a given timestep and the target profile. The Pareto-guided alignment loss is defined as:

$$\mathcal{L}_{\text{comm}}(\phi) = \mathbb{E}\left[\|(v^t, \zeta^t) - (v^\star, \zeta^\star)\|_2^2 \right], \tag{8}$$

where $\mathbb{E}[\cdot]$ denotes the empirical average over the on-policy mini-batch. The outcome profile is formed by the critic's bounded value $v^t \in [0, 1]$ and the mean communication usage $\zeta^t \in [0, 1]$. To ensure the loss provides a direct learning signal only to the communication policy, we treat the value-based components as fixed

targets in this calculation. The alignment loss is therefore only differentiable with respect to the communication policy parameters, ϕ, via the ζ^t term; the gradient paths through the critic's value estimate v^t and the target profile $(v^\star, \zeta^\star)$ are detached using a stop-gradient operation. This guides the communication policy to produce outcomes that align with the target's cost-benefit profile without interfering with the critic's learning objective. To ensure time-scale consistency, the Pareto front $\hat{\mathcal{P}}$ and target $\mathbf{p}^\star$ are recomputed from the same on-policy minibatch, ensuring the alignment signal reflects the current batch's statistics.

By doing so, the learned communication behaviours are not fixed but dynamically adapt to the agents' local context and evolving environmental conditions, yielding communication structures that are optimised in response to changing situational demands.

3.4.3 Coordinated Actor-Critic Updates. The final stage is a coordinated training process that jointly optimises the communication policy, actor network, and centralised critic. These components are updated in concert to guide the MAS towards the efficient behaviour identified by the dynamic target profile, $\mathbf{p}^\star$.

The actor policy, π_θ, overcomes partial observability by conditioning its decisions on a message-augmented observation history. This history is maintained in the actor's internal state, h^t, by a Gated Recurrent Unit that recurrently processes feature embeddings of the agent's sensory input and received messages. These messages contain encoded representations of the sender's intended action, generated using local observation, and are aggregated by the receiver via mean aggregation. To solve the credit assignment problem for these discrete communication events, we employ a straight-through Gumbel–Sigmoid estimator. This creates a differentiable communication channel, and its annealed temperature τ is coupled to the policy's entropy coefficient, which avoids introducing an additional hyperparameter.

To optimise the actor's parameters θ and ensure stable training updates, we employ the clipped surrogate objective from Proximal Policy Optimisation (PPO). Denoting the importance ratio as $\ell^t(\theta) = \frac{\pi_\theta(a^t|h^t)}{\pi_{\theta_{\text{old}}}(a^t|h^t)}$, where $\pi_{\theta_{\text{old}}}$ is the policy before the update, the actor objective is:

$$\mathcal{L}_{\text{actor}}(\theta) = -\mathbb{E}\left[\min\left(\ell^t(\theta)\hat{A}^t, \text{clip}(\ell^t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}^t\right)\right], \quad (9)$$

where $\hat{A}^t$ is the GAE-based advantage given in eq. (11) and ϵ is the PPO clipping parameter.

The centralised critic, parameterised by ψ, is a value function $V_\psi(s^t)$ trained under the CTDE paradigm. Its role is to provide the low-variance advantage estimates required for the PPO actor update. To this end, it is trained by minimising a bootstrapped value loss derived from GAE. First, we compute the temporal-difference (TD) residuals, making the terminal state handling explicit:

$$\delta^t = r^t + \gamma(1-d^t)V_\psi(s^{t+1}) - V_\psi(s^t), \quad (10)$$

where d^t is an indicator that is 1 if s^{t+1} is a terminal state and 0 otherwise. The advantages are then calculated as the discounted sum of these residuals, truncated at the end of the episode:

$$\hat{A}^t = \sum_{l=0}^{T-1-t} (\gamma\lambda_{\text{GAE}})^l \delta^{t+l}, \quad (11)$$

where $\lambda_{\text{GAE}} \in [0,1]$. We form fixed value targets, $\hat{V}^t := \text{stopgrad}(\hat{A}^t + V_\psi(s^t))$, and minimise the mean-squared error over the on-policy minibatch $\mathcal{D}$:

$$\mathcal{L}_{\text{critic}}(\psi) = \mathbb{E}\left[(V_\psi(s^t) - \hat{V}^t)^2\right]. \quad (12)$$

For the outcome profiles in our Pareto analysis, we use a normalised value $v^t \in [0,1]$ derived from $V_\psi(s^t)$, while the critic itself is trained on the unnormalised targets. As the critic's input state s^t includes all agents' observations and message information, its value estimates are communication-aware and consistently inform the value component, $v^t = V_\psi(s^t)$, of the outcome profiles.

The communication policy, actor, and critic are optimised jointly, creating a coupled learning dynamic. The communication policy π_ϕ is updated using $\mathcal{L}_{\text{comm}}$, learning to establish an efficient communication topology that leads to outcomes near the target $\mathbf{p}^\star$. The actor and critic are updated using their respective losses, learning to execute actions that maximise rewards within the communication structure established by π_ϕ. This coordinated process, anchored by the dynamically evolving Pareto front, allows the system to discover complex, adaptive, and efficient coordination strategies.

4 EXPERIMENTS

We conduct a series of experiments to validate COBE's effectiveness and analyse its learned behaviour. Our evaluation is guided by four central research questions (RQ), which we will address in turn: **RQ1**. How does COBE compare to state-of-the-art MAC baselines in terms of task performance and communication cost?; **RQ2**. What is the qualitative nature of the communication strategies learned by COBE, and how do they differ from other methods?; **RQ3**. To what extent is COBE's curvature-based target selection crucial for learning a balanced and efficient communication policy compared to simpler heuristics?; and **RQ4**. Can COBE's dual-objective formulation achieve a more effective reward-cost trade-off than tuned single-objective approaches using reward scalarisation and constrained optimisation?

4.1 Experimental Setup

4.1.1 Environments. We evaluate COBE in three multi-agent environments from existing literature [26, 36]. These tasks were intentionally chosen for their structural simplicity, which allows for a clear and direct analysis of emergent communication and coordination strategies. Unlike complex domains with fine-grained actuation dynamics, these environments are well-suited for isolating and analysing the learned cost-benefit trade-offs in decision-making. We use extended variants to increase task difficulty and encourage meaningful agent interaction.

Cooperative Navigation (CN). Seven agents coordinate to cover seven landmarks while avoiding collisions. The reward function penalises agents for distance to the nearest landmark, with additional penalties of -1 per collision and -1 per unoccupied landmark.

Predator-Prey (PP). Seven predators cooperate to capture three faster, evasive prey, making solo capture infeasible. The reward is based on the negative sum of distances to the nearest prey, with penalties for collisions and for each uncaptured prey.

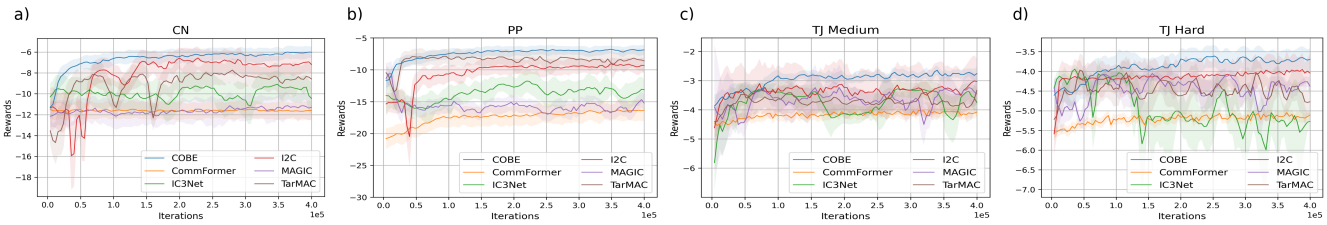

Figure 2: Performance analysis across methods

Table 1: Quantitative performance across methods, results are shown as mean ± standard deviation

Task	Metric	COBE	CommFormer	IC3Net	I2C	MAGIC	TarMAC
CN	Reward	−6.2 ± 0.6	−11.6 ± 0.5	−10.1 ± 1.1	−7.3 ± 1.2	−11.0 ± 0.7	−8.5 ± 1.3
	Comm%	39% ± 1.9%	81% ± 5.1%	89% ± 3.3%	66% ± 5.7%	57% ± 2.1%	100% ± 0.0%
PP	Reward	−7.2 ± 0.7	−16.2 ± 0.5	−13.5 ± 0.6	−9.5 ± 1.0	−16.5 ± 0.6	−8.7 ± 1.4
	Comm%	38% ± 2.1%	78% ± 5.8%	91% ± 5.3%	72% ± 2.5%	44% ± 2.7%	100% ± 0.0%
TJ Medium	Reward	−2.8 ± 0.3	−4.3 ± 0.2	−4.1 ± 0.7	−3.2 ± 0.6	−3.6 ± 0.5	−3.1 ± 0.4
	Comm%	33% ± 3.2%	65% ± 5.5%	70% ± 6.7%	48% ± 3.6%	41% ± 4.3%	100% ± 0.0%
TJ Hard	Reward	−3.7 ± 0.3	−5.1 ± 0.2	−5.3 ± 0.5	−4.1 ± 0.8	−4.6 ± 0.6	−4.7 ± 0.5
	Comm%	36% ± 3.6%	92% ± 7.1%	98% ± 6.6%	55% ± 4.7%	45% ± 5.2%	100% ± 0.0%

Traffic Junction (TJ). Vehicles navigate intersections to minimise travel time and avoid collisions. The medium setting features ten vehicles and one junction; the hard setting includes twenty and four interconnected junctions. The team reward includes penalties of -10 per collision and -0.5t_{entry} for prolonged presence, where t_{entry} is the time since entry.

4.1.2 Baselines. To benchmark COBE's performance, we compare it against five open-source MAC baselines that represent a diverse set of communication paradigms. These include CommFormer [17], which induces sparse graphs through attention-based connectivity; IC3Net [35], which employs a binary gating mechanism to decide whether to broadcast; I2C [8], which initiates targeted request-reply communication based on causal inference; MAGIC [30], which uses a graph-attention network to determine if and with whom to communicate; and TarMAC [5], a full-communication approach where agents broadcast unconditionally and rely on receiver-side attention to assess relevance.

4.1.3 Evaluation Metrics. A persistent challenge in MAC research is the inconsistent nature of its evaluation practices. This creates a gap in the literature for a standardised measure of communication volume, as most methods report task reward and message frequency as separate quantities. Such reporting makes a fair comparison of efficiency across different strategies, including broadcasting versus targeted messaging, difficult to achieve. To address this, we define a standardised metric, normalised communication usage Comm%, which provides a principled and consistent basis for comparison irrespective of the underlying communication paradigm.

Definition 4 (Experimental Communication Efficiency). *The normalised communication usage, denoted as Comm%, is the fraction of active directed communication channels. This value is determined by a set of binary transmission gates, $d_{i\to j} \in \{0, 1\}$. The usage is formally defined as:*

$$Comm\% = \frac{1}{N(N-1)} \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{I}, i \neq j} d_{i\to j}. \tag{13}$$

This metric normalises communication relative to a full broadcast, enabling consistent comparisons across different agent populations. The numerator counts the number of active directed links initiated by the sender. For instance, a single broadcast event from one agent contributes $N - 1$ links. In request–reply schemes, the request and the reply are counted as two distinct links. For attention-based receivers, a transmission still counts as a link if it is initiated by the sender, regardless of the receiver's attention weight.

4.2 RQ1. Comparative Performance Analysis

This section compares COBE against established baseline methods, with results summarised in Table 1 and Figure 2.

In CN, COBE demonstrates stable learning, achieving a strong reward of −6.2 with an efficient 39% communication. This performance contrasts with methods relying on indiscriminate messaging, such as TarMAC and IC3Net, whose noisy learning signals lead to significant instability, reflected in their high reward standard deviations of ±1.3 and ±1.1 respectively. Similarly, the topology-learning methods, CommFormer and MAGIC, stagnate at suboptimal rewards of −11.6 and −11.0. While I2C is the most competitive baseline with a reward of −7.3, its performance still trails COBE's and requires a substantially higher communication of 66%.

This pattern of superior efficiency continues in PP. COBE again secures the highest reward −7.2 with a low communication cost

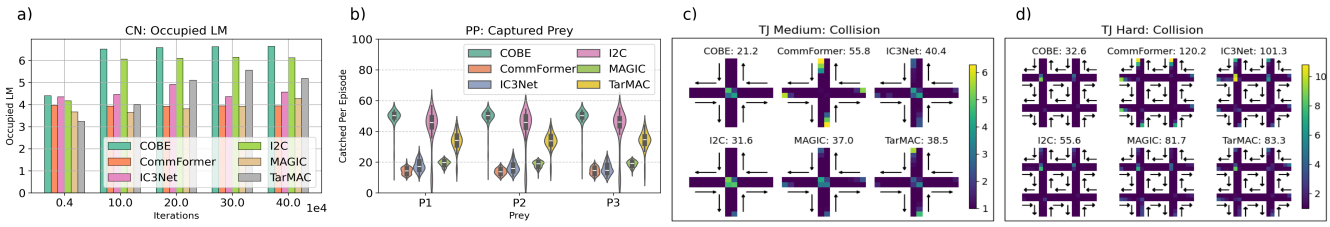

Figure 3: Qualitative comparison of key task-specific metrics a) Average number of landmarks occupied in CN; b) Distribution of captured prey in PP; c) & d) Collision density heatmaps in the TJ Medium and Hard settings respectively

of 38%. Although TarMAC's global broadcast is reasonably effective for this task's encirclement strategy, achieving a competitive reward of −8.7, it requires maximum communication cost. Meanwhile, MAGIC's trajectory shows a notable collapse, suggesting its learned topology cannot adapt to the dynamic chase, while IC3Net's performance remains volatile.

The advantages of COBE's targeted approach are most pronounced in TJ. Its learning trajectory remains stable and converges to the highest reward in both settings. Most baseline communication paradigms struggle to manage the precise, timed coordination required at intersections, leading to highly volatile learning curves. This is particularly true for IC3Net in the Hard setting, where its mechanism proves inadequate, resulting in a poor −5.3 reward despite a near-maximal 98% communication rate. I2C re-emerges as the strongest baseline, but its final reward in the Hard setting is still limited to −4.1 for a 55% communication cost.

4.3 RQ2. Qualitative Strategy Analysis

Figure 3a) shows COBE achieves high and consistent landmark coverage early in CN, as its cost-benefit framework allows agents to efficiently establish an allocation, after which communication is rightly suppressed. I2C also performs well due to its structured protocol, though its reactive nature is less efficient. The performance of TarMAC and IC3Net fluctuates, a result of their indiscriminate communication which respectively cause information overload and an unreliable learning signal. Conversely, topology-learning methods like CommFormer and MAGIC stagnate, as their mechanisms struggle to adapt to the fluid nature of the task.

The violin plots in Figure 3b) illustrate coordination consistency. COBE's dense and balanced distribution stems from its ability to identify moments of high marginal utility for communication, such as when predators are poised to form an encirclement. This contrasts with I2C, which, despite achieving high peak performance as shown by its upper whisker, exhibits a broader distribution skewed towards lower outcomes, suggesting less consistent success. TarMAC's performance is also noteworthy; its broadcast, while inefficient, provides shared awareness for predators to coordinate.

In TJ, the heatmaps in Figure 3c) and d) show COBE registers the lowest collision count, as its cost-benefit analysis correctly pinpoints the intersection as the highest-risk area where communication is most valuable. Similarly, I2C focuses its communication on the junction, as its causal inference model deduces high inter-agent influence in this zone. However, its higher collision count suggests this reactive protocol is less effective at pre-empting incidents. A key distinction is found with CommFormer and TarMAC, which suffer numerous collisions at lane entries. This behaviour indicates that CommFormer's attention likely defaults to a simple proximity link, while TarMAC's broadcast overwhelms agents, causing fatal indecision. MAGIC and IC3Net also exhibit high collision rates, primarily within the junction, suggesting their mechanisms identify the correct location but lack the sophistication to determine the precise timing or content required for prevention.

4.4 RQ3. Impact of Curvature-Based Selection

To isolate the impact of our selection mechanism, we compare COBE against three ablation variants that employ simpler, heuristic-based target selection in PP. These heuristics target the Pareto front endpoint with the lowest communication cost abl:$\mathbf{p}_1$, the endpoint with the highest reward abl:$\mathbf{p}_K$, and the median point abl:$\mathbf{p}_m$.

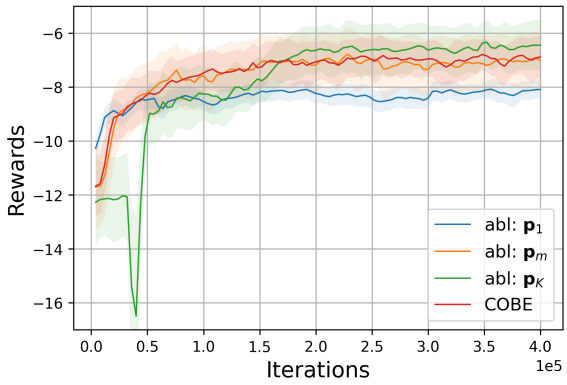

Figure 4: Performance analysis of target selections in PP

The learning trajectories in Figure 4 reveal important differences in learning stability. The abl:$\mathbf{p}_1$ variant, which minimises communication, quickly plateaus at a low reward. Conversely, the abl:$\mathbf{p}_K$ variant, targeting maximal reward, exhibits significant initial instability; its performance collapses before recovering, which suggests that encouraging maximal communication before a coherent policy is learned is actively detrimental. While both COBE and the median heuristic, abl:$\mathbf{p}_m$, converge to a comparable final reward, COBE's learning trajectory is visibly more stable. This is likely because the median target of abl:$\mathbf{p}_m$ is sensitive to the distribution of points on the empirical Pareto front, creating a more volatile objective. In

contrast, COBE's curvature-based target is a more robust geometric feature of the front, providing a more consistent learning signal.

Quantitatively, the results in Table 2 confirm these trade-offs and highlight the efficiency of COBE's selection criterion. The $\mathbf{p}_1$ and $\mathbf{p}_K$ heuristics define the performance bounds, showing the respective costs of under- and over-communication. The most telling comparison is with the median heuristic, abl:$\mathbf{p}_m$. While COBE's reward is statistically on par with abl:$\mathbf{p}_m$, it achieves this performance with a communication cost of 38%, a marked improvement over the 47% required by the simpler heuristic. This demonstrates that curvature-based selection successfully identifies a more efficient operating point on the cost-benefit frontier, promoting not only a superior final trade-off but also a more stable learning process.

Table 2: Quantitative comparison of target selections in PP, results are shown as mean ± standard deviation

Method	Reward	Comm%
abl:$\mathbf{p}_1$	-8.1 ± 0.4	$28\% \pm 1.6\%$
abl:$\mathbf{p}_m$	-7.3 ± 1.0	$47\% \pm 2.5\%$
abl:$\mathbf{p}_K$	-6.5 ± 0.9	$71\% \pm 1.9\%$
COBE	-7.2 ± 0.7	$38\% \pm 2.1\%$

4.5 RQ4. Dual-Objective Effectiveness

This experiment evaluates the two single-objective paradigms outlined in Section 3.2, constrained optimisation and reward scalarisation, against COBE's dual-objective approach in CN. As shown in Figure 5 and Table 3, this experiment highlights the advantages of COBE's dual-objective approach in navigating the cost-benefit landscape. COBE successfully identifies a point on the Pareto-optimal frontier that is demonstrably more efficient than the outcomes produced by any of the tuned single-objective methods, securing both a higher reward and lower communication usage.

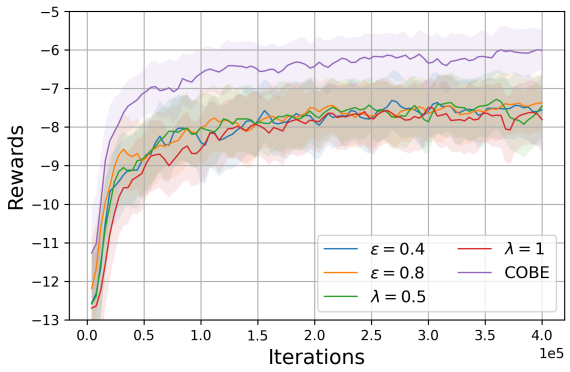

Figure 5: Performance analysis of dual- and single-objective approaches in CN

This performance gap is explained by the learning trajectories in Figure 5, which show the single-objective variants becoming confined to a suboptimal performance plateau. This suggests that enforcing a static communication budget or applying a fixed penalty fails to produce a meaningful communication strategy, instead yielding a simplistic, heuristic-based policy unable to discover the superior policy space that COBE accesses. The quantitative results in Table 3 further detail these inefficiencies. The constrained optimisation approach ϵ exhibits clear diminishing returns, as doubling the communication budget from 40% to 80% yields only a marginal reward improvement. This implies that without a mechanism to assess the marginal utility of each communication act, much of the enforced additional communication is of low value. The reward scalarisation method λ proves even more brittle; a moderate penalty is too permissive and results in excessive communication, while a high penalty is too restrictive and harms performance.

Table 3: Quantitative comparison of dual- and single-objective approaches in CN, results are shown as mean ± standard deviation

Method	Reward	Comm%
ϵ=0.4	-7.8 ± 0.7	$40\% \pm 0.3\%$
ϵ=0.8	-7.4 ± 0.9	$80\% \pm 0.5\%$
$\lambda = 0.5$	-7.5 ± 1.0	$89\% \pm 5.1\%$
$\lambda = 1$	-7.9 ± 1.1	$47\% \pm 3.2\%$
COBE	-6.2 ± 0.6	$39\% \pm 1.9\%$

These single-objective methods are limited by their reliance on a global preference. In contrast, COBE's dual-objective formulation enables it to dynamically assess the value of a message, allowing it to achieve a trade-off without brittle hyperparameter tuning.

5 CONCLUSION

This paper presents COBE, a novel framework that conceptualises MAC as a dual-objective, cost-benefit problem. The methodology leverages a data-driven process to construct an empirical Pareto front, which maps the trade-off between a forward-looking benefit and the associated communication cost. This provides a principled answer to *why* communication should occur, ensuring a message is sent only when its cost is justifiable. By operationalising the principle of diminishing returns, a curvature-based analysis of the front identifies the point of greatest marginal return, which informs *whom* to contact, guiding the policy towards maximising reward while minimising communication without manual tuning.

The empirical evaluation validates this principled approach, demonstrating that COBE consistently surpasses state-of-the-art baselines by achieving a superior cost-benefit trade-off. This advantage is not merely quantitative, it stems from the cultivation of qualitatively more robust strategies, wherein agents learn adaptive and context-aware communication behaviours. The dual-objective formulation and curvature-based selection proves more robust and efficient than brittle single-objective paradigms and simpler heuristics. Future work will address the scalability of these principles under realistic conditions, including heterogeneous costs and uncertainty-aware value estimates.

REFERENCES

[1] Jingdi Chen, Tian Lan, and Carlee Joe-Wong. 2024. Rgmcomm: Return gap minimization via discrete communications in multi-agent reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 17327–17336.

[2] Tianshu Chu, Sandeep Chinchali, and Sachin Katti. 2020. Multi-agent reinforcement learning for networked system control. *arXiv preprint arXiv:2004.01339* (2020).

[3] Karen S Cook, Coye Cheshire, Eric RW Rice, and Sandra Nakagawa. 2013. Social exchange theory. In *Handbook of social psychology*. Springer, 61–88.

[4] Russell Cropanzano and Marie S Mitchell. 2005. Social exchange theory: An interdisciplinary review. *Journal of management* 31, 6 (2005), 874–900.

[5] 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.

[6] Gang Ding, Zeyuan Liu, Zhirui Fang, Kefan Su, Liwen Zhu, and Zongqing Lu. 2024. Multi-Agent Coordination via Multi-Level Communication. *Advances in Neural Information Processing Systems* 37 (2024), 118513–118539.

[7] Shifei Ding, Wei Du, Ling Ding, Lili Guo, and Jian Zhang. 2024. Learning efficient and robust multi-agent communication via graph information bottleneck. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 17346–17353.

[8] Ziluo Ding, Tiejun Huang, and Zongqing Lu. 2020. Learning individually inferred communication for multi-agent cooperation. *Advances in neural information processing systems* 33 (2020), 22069–22079.

[9] Wei Du, Shifei Ding, Lili Guo, Jian Zhang, and Ling Ding. 2024. Expressive multi-agent communication via identity-aware learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 17354–17361.

[10] 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*. 456–464.

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

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

[13] Benjamin Freed, Rohan James, Guillaume Sartoretti, and Howie Choset. 2020. Sparse discrete communication learning for multi-agent cooperation through backpropagation. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 7993–7998.

[14] 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.

[15] Nikunj Gupta, G Srinivasaraghavan, Swarup Mohalik, Nishant Kumar, and Matthew E Taylor. 2023. Hammer: Multi-level coordination of reinforcement learning agents via learned messaging. *Neural Computing and Applications* (2023), 1–16.

[16] Shuai Han, Mehdi Dastani, and Shihan Wang. 2023. Model-based sparse communication in multi-agent reinforcement learning. In *Proceedings of the 2023 international conference on autonomous agents and multiagent systems*. International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS), 439–447.

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

[18] Jiechuan Jiang, Chen Dun, Tiejun Huang, and Zongqing Lu. 2018. Graph convolutional reinforcement learning. *arXiv preprint arXiv:1810.09202* (2018).

[19] Jiechuan Jiang and Zongqing Lu. 2018. Learning attentional communication for multi-agent cooperation. *Advances in neural information processing systems* 31 (2018).

[20] Daewoo Kim, Sangwoo Moon, David Hostallero, Wan Ju Kang, Taeyoung Lee, Kyunghwan Son, and Yung Yi. 2019. Learning to schedule communication in multi-agent reinforcement learning. *arXiv preprint arXiv:1902.01554* (2019).

[21] Vijay Konda and John Tsitsiklis. 1999. Actor-critic algorithms. *Advances in neural information processing systems* 12 (1999).

[22] Xinran Li and Jun Zhang. 2023. Context-aware communication for multi-agent reinforcement learning. *arXiv preprint arXiv:2312.15600* (2023).

[23] Yi-Yu Lin and Xiao-Jun Zeng. 2023. Reward-Guided Individualised Communication for Deep Reinforcement Learning in Multi-Agent Systems. In *UK Workshop on Computational Intelligence*. Springer, 79–94.

[24] Yi-Yu Lin and Xiao-Jun Zeng. 2025. Enhancing multi-agent communication through credibility and reward-based optimisation. *International Journal of General Systems* (2025), 1–25.

[25] Yong Liu, Weixun Wang, Yujing Hu, Jianye Hao, Xingguo Chen, and Yang Gao. 2020. Multi-agent game abstraction via graph attention neural network. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 7211–7218.

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

[27] Xueguang Lyu, Yuchen Xiao, Brett Daley, and Christopher Amato. 2021. Contrasting centralized and decentralized critics in multi-agent reinforcement learning. *arXiv preprint arXiv:2102.04402* (2021).

[28] Xiangrui Meng and Ying Tan. 2024. Pmac: Personalized multi-agent communication. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 17505–17513.

[29] Matthew Morris, Thomas D Barrett, and Arnu Pretorius. 2022. Universally expressive communication in multi-agent reinforcement learning. *Advances in neural information processing systems* 35 (2022), 33508–33522.

[30] Yaru Niu, Rohan R Paleja, and Matthew C Gombolay. 2021. Multi-Agent Graph-Attention Communication and Teaming.. In *AAMAS*, Vol. 21. 20th.

[31] Frans A Oliehoek and Christopher Amato. 2016. *A concise introduction to decentralized POMDPs*. Springer.

[32] Afshin Oroojlooy and Davood Hajinezhad. 2023. A review of cooperative multi-agent deep reinforcement learning. *Applied Intelligence* 53, 11 (2023), 13677–13722.

[33] Stan Salvador and Philip Chan. 2004. Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. In *16th IEEE international conference on tools with artificial intelligence*. IEEE, 576–584.

[34] Ville Satopaa, Jeannie Albrecht, David Irwin, and Barath Raghavan. 2011. Finding a" kneedle" in a haystack: Detecting knee points in system behavior. In *2011 31st international conference on distributed computing systems workshops*. IEEE, 166–171.

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

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

[37] Chuxiong Sun, Zehua Zang, Jiabao Li, Jiangmeng Li, Xiao Xu, Rui Wang, and Changwen Zheng. 2024. T2mac: Targeted and trusted multi-agent communication through selective engagement and evidence-driven integration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 15154–15163.

[38] Jie Xu, Yunsheng Tian, Pingchuan Ma, Daniela Rus, Shinjiro Sueda, and Wojciech Matusik. 2020. Prediction-guided multi-objective reinforcement learning for continuous robot control. In *International conference on machine learning*. PMLR, 10607–10616.

[39] Ming Yang, Kaiyan Zhao, Yiming Wang, Renzhi Dong, Yali Du, Furui Liu, Mingliang Zhou, and Leong Hou U. 2024. Team-wise effective communication in multi-agent reinforcement learning. *Autonomous Agents and Multi-Agent Systems* 38, 2 (2024), 36.

[40] Yuxuan Yi, Ge Li, Yaowei Wang, and Zongqing Lu. 2022. Learning to share in networked multi-agent reinforcement learning. *Advances in Neural Information Processing Systems* 35 (2022), 15119–15131.

[41] Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. 2022. The surprising effectiveness of ppo in cooperative multi-agent games. *Advances in neural information processing systems* 35 (2022), 24611–24624.

[42] Lebin Yu, Yunbo Qiu, Quanming Yao, Yuan Shen, Xudong Zhang, and Jian Wang. 2024. Robust communicative multi-agent reinforcement learning with active defense. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 17575–17582.

[43] Lei Yuan, Jianhao Wang, Fuxiang Zhang, Chenghe Wang, Zongzhang Zhang, Yang Yu, and Chongjie Zhang. 2022. Multi-agent incentive communication via decentralized teammate modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 9466–9474.

[44] Won Joon Yun, Byungju Lim, Soyi Jung, Young-Chai Ko, Jihong Park, Joongheon Kim, and Mehdi Bennis. 2021. Attention-based reinforcement learning for real-time UAV semantic communication. In *2021 17th International Symposium on Wireless Communication Systems (ISWCS)*. IEEE, 1–6.

[45] Mohamed Salah Zaïem and Etienne Bennequin. 2019. Learning to communicate in multi-agent reinforcement learning: A review. *arXiv preprint arXiv:1911.05438* (2019).

[46] Sai Qian Zhang, Qi Zhang, and Jieyu Lin. 2019. Efficient communication in multi-agent reinforcement learning via variance based control. *Advances in neural information processing systems* 32 (2019).

[47] Sai Qian Zhang, Qi Zhang, and Jieyu Lin. 2020. Succinct and robust multi-agent communication with temporal message control. *Advances in neural information processing systems* 33 (2020), 17271–17282.

[48] Xingyi Zhang, Ye Tian, and Yaochu Jin. 2014. A knee point-driven evolutionary algorithm for many-objective optimization. *IEEE transactions on evolutionary computation* 19, 6 (2014), 761–776.

[49] Changxi Zhu, Mehdi Dastani, and Shihan Wang. 2024. A survey of multi-agent deep reinforcement learning with communication. *Autonomous Agents and Multi-Agent Systems* 38, 1 (2024), 4.