SCoUT：通过效用引导的时间分组实现多智能体强化学习中的可扩展通信

Manav Vora、Gokul Puthumanaillam、Hiroyasu Tsukamoto、Melkior Ornik

关键词：多智能体强化学习、通信调度、大规模多智能体系统。

arXiv:2603.04833v1 [cs.MA] 5 Mar 2026

摘要

学习式通信能够改善部分可观测多智能体强化学习中的协调，但扩展到大型团队并不容易，因为（i）决定何时以及与谁通信会带来大量相互耦合的离散选择；（ii）单条消息的影响难以单独识别并用于学习。我们提出了**SCoUT**（**S**可扩展**Co**通信：通过**U**效用引导的**T**时间分组），这是一种可扩展的学习式通信方法，通过引入时间与智能体抽象来实现学习式通信。SCoUT 每隔 K 步重新采样软智能体分组，以形成可微的亲和度，并据此调整接收方选择；同时，分组感知评论家预测群组级值，并将其映射为逐智能体基线，从而降低评论家复杂度。为了获得精确的通信学习信号，SCoUT 使用基于反事实邮箱的优势，分别隔离单条消息对发送和接收方决策的贡献。基准测试实验表明，随着智能体种群增长至数百个，SCoUT 仍然有效，而以往的学习式通信基线方法则会随着规模扩大而性能下降。

贡献

1. **SCoUT** 是一种可扩展的学习式通信方法，每经过 K 个环境步数便重新采样软智能体分组，并将由此产生的亲和度作为可微偏置，用于逐步选择接收方。这种时间扩展的 软分组机制，将每步在 $2^{N(N-1)}$ 种可能的有向通信图中进行的组合式链路选择，转化为由组亲和度在训练期间引导的结构化可微路由。

   背景： 以往的选择性通信方法通常通过调度（Kim 等人，2019）、基于注意力的路由（Das 等人，2019）或施加全局拓扑（Li 等人，2025），在每个步数重新决定通信结构；而 SCoUT 引入了一种缓慢变化的潜在结构，在不施加硬性约束的情况下引导路由。

2. 一种分组感知评论家，利用全局信息预测组级价值，并通过软分组分配将其映射为逐智能体价值基线，从而降低评论器输出的复杂度，并稳定大规模群体中的集中式训练与分散式执行训练。

   背景： 这与标准集中式训练与分散式执行基线中的单体集中式评论家不同（Yu 等人，2022年），并且与价值分解方法互为补充（Rashid 等人，2020），因为在我们的设置中，分解与学习得到的通信结构紧密关联。

3. 一种反事实邮箱信用分配规则：通过比较实际邮箱与留一法（反事实）邮箱下接收方的预测价值，为二元发送动作和接收方选择都提供学习信号。

   背景： 本方法受到多智能体演员-评论家学习中的反事实基线启发（Foerster 等人，2018年），但针对通信进行了专门设计，通过隔离单条消息的边际贡献来实现。

4. 在大规模群体基准测试上进行的实证评估表明，SCoUT 能够在拥有数百个智能体的场景中高效训练，并且随着种群规模增大，其性能优于学习式通信基线方法。

   背景： 我们考虑的场景中，受控智能体数量分别比先前工作在 Battle 和 Pursuit 中多约 1.6× 和 12× 个。

SCoUT：多智能体强化学习中经由效用引导的时间分组的可扩展通信

Manav Vora, Gokul Puthumanaillam，Hiroyasu Tsukamoto,
Melkior Ornikmkvora2, gokulp2, hiroyasu, mornik@illinois.edu

伊利诺伊大学厄巴纳-香槟分校

摘要

通信能够改善部分可观测多智能体强化学习中的协调，但学习何时以及与谁进行通信，需要在众多可能的发送方-接收方对之间做出选择，而且任何单条消息对未来奖励的影响都难以分离。我们提出**SCoUT** （**S**可扩展**Co**通信，经由**U**效用引导的**T**时间分组），通过传统多智能体强化学习中的时间与智能体抽象来应对这两项挑战。在训练过程中，SCoUT每隔 K 个环境步数（宏步）借助Gumbel-Softmax重新采样软智能体组；这些组是潜在簇，由此诱导出的亲和度被用作接收方上的可微先验。基于相同的分配，分组感知评论家预测每个智能体组的值，并通过相同的软分配将其映射为逐智能体基线，从而降低评论家复杂度与方差。每个智能体采用三头策略进行训练：环境动作、发送决策与接收方选择。为获得精确的通信学习信号，我们通过以解析方式从接收方的聚合消息中移除各发送方的贡献，推导反事实通信优势。该反事实计算能够对发送决策与接收方选择决策进行精确的信用分配。执行时，所有集中式训练组件均被舍弃，仅运行逐智能体策略，从而保持分散执行。大规模基准上的实验表明，SCoUT能够学习定向通信，并在大规模场景中保持有效，而先前方法会随着种群规模的增长而变差。最后，消融实验证实，时间分组与反事实通信信用对于可扩展性均至关重要。

项目网站、视频和代码：https://scout-comm.github.io/

1 引言

通信是部分可观测的多智能体强化学习（MARL）中一种自然的协调机制：交换局部信息或意图，可以缓解个体视野受限的问题，并帮助各智能体同步行为，以实现共同目标（Bernstein 等人，2002年；Hernandez-Leal 等人，2019；Buşoniu 等人，2008年）。因此，近期研究通过端到端方式，在学习智能体策略的同时学习通信协议和路由决策，而非预先设定固定拓扑或手工设计消息（Niu 等人，2021年；Li 等人，2021年）。然而，当学习式通信扩展到规模更大的团队时，仍面临困难：随着智能体种群扩大，计算成本不断增加，训练不稳定性也随之加剧（Li 等人，2021年）。

学习式通信在扩展性方面的一个关键瓶颈在于，决定何时以及与谁通信本质上是一个组合问题：在每个步骤中，团队都会隐式地选择一个通信图

该通信图涵盖 N 个智能体，会产生 $2^{N(N-1)}$ 种可能的连接模式。全连接架构无需显式进行路由决策，但每个步骤都需要进行 $O(N^2)$ 次消息交换和聚合；随着 N 规模增大，这会带来高昂成本，并产生大量无关通信（Sukhbaatar 等人，2016；Foerster 等人，2016）。选择性通信方法通过门控或调度传输（Singh 等人，2019；Kim 等人，2019），或通过注意力学习定向路由（Jiang 和 Lu，2018；Das 等，2019；Liu 等，2020），降低每个步骤的通信和计算开销。然而，这些方法通常采用演员-评论家目标进行集中式训练、分散式执行训练，其中集中式价值函数为环境动作和通信决策提供学习信号（Lowe 等人，2017；Yu 等，2022）。随着智能体数量增加，训练时往往仍面临两项挑战：（i）训练时可扩展性——评论家和通信模块必须处理高维多智能体信息；（ii）信用分配——当许多通信共同影响结果时，如何将下游回报归因于特定消息和接收方选择（Foerster 等，2018）。从经验上看，学习式通信的研究通常只在 $\mathcal{O}$(10-100)个智能体的规模下进行演示（Li 等，2025）。

为了让包含数百个智能体的大型团队能够实现学习式通信，我们提出了**SCoUT** (**S**可扩展**CO**通信，通过**U**效用引导的**T**时间分组），这是一种引入时间和智能体抽象的训练框架。其核心思想是将通信结构视为缓慢变化的潜变量：每经过 K 个环境步数（一个宏步），SCoUT 就会重新采样一组软智能体分组——即将智能体可微地聚类到少数几个潜在群组中——并采用可微分类采样（Jang 等人，2017年；Maddison 等人，2017年）。这些组亲和度为接收方选择和消息聚合提供可微先验，并应用于具有一步延迟的邮箱中；邮箱即每个智能体的消息缓冲区，用于存放上一步收到的消息。相同的分组分配还通过分组感知评论家对价值估计进行分解：该评论家预测组级价值，并将其映射回逐智能体基线，从而降低评论家复杂度和方差。我们使用三头近端策略优化策略（环境动作、发送、接收方选择）训练每个智能体（Schulman 等人，2017年；2016），并通过解析地从接收方的聚合邮箱中移除发送方的贡献，得出反事实通信优势。这一做法借鉴了用于多智能体信用分配的反事实基线（Foerster 等人，2018年）。

贡献。

(i) 我们为多智能体强化学习通信引入一种时间扩展的软分组机制，该机制诱导出关于接收方的学习分布。该机制以由组亲和度引导的结构化、可微路由，取代了在 $2^{N(N-1)}$ 个可能的有向通信图上进行的逐步组合链路选择。(ii) 我们提出一种分组感知评论家，它通过软分组对值预测进行分解，并将组值映射回逐智能体基线，从而提升大规模种群中集中式训练与分散式执行训练的可扩展性与稳定性。(iii) 我们开发了一种反事实邮箱信用分配规则，通过解析地从接收方的聚合消息中移除发送方的贡献，为发送与接收方选择提供直接的学习信号。

在执行时，SCoUT 丢弃集中式组件（群体采样器、评论器以及反事实计算），仅运行学习到的逐智能体策略，从而保持分散执行。

我们在来自 PettingZoo（Terry 等人，2021年）和 MAgent（Zheng 等人，2018年）的大规模合作基准测试上评估 SCoUT，它在扩展到数百个智能体的同时学习定向通信，并随着种群规模增长优于最先进基线方法；Pursuit 中的消融实验分离了时间分组与反事实信用分配的影响。

2 相关工作

基础理论与集中式训练、分散式执行训练。部分可观察多智能体决策可通过去中心化控制形式体系来描述，例如去中心化部分可观测马尔可夫决策过程（Bernstein 等人，2002年）以及马尔可夫博弈（Littman，1994）。深度多智能体强化学习通常采用集中式训练、分散式执行，利用全局信息训练评论家或进行价值分解，同时执行分散式策略。

（CTDE），利用全局信息训练评论家或进行价值分解，同时执行分散式策略（Lowe 等人，2017年；Foerster 等，2018）。代表性的集中式训练与分散式执行方法包括集中式评论家（MADDPG）（Lowe 等人，2017年）、反事实演员-评论家学习（COMA）（Foerster 等，2018），以及价值分解（VDN/QMIX、QTRAN、QPLEX）（Sunehag 等人，2018年；Rashid 等，2020；Son 等人，2019；Wang 等人，2021年）；MAPPO 等 PPO 风格的 CTDE 基线方法也具有竞争力（Schulman 等，2017；Yu 等，2022）。然而，随着智能体数量增加，集中式评论家可能成为瓶颈：它们必须概括高维交互，并提供低方差的学习信号。SCoUT 通过软分组分解价值估计，并将群组值映射为逐智能体基线，以此解决这一问题。

多智能体强化学习中的学习式通信。大量研究采用端到端方式学习可微通信，使智能体能够交换消息，从而在部分可观测条件下改善协调。CommNet 聚合广播式连续消息（Sukhbaatar 等人，2016），而 DIAL/RIAL 则在学习去中心化策略的同时，通过可微和离散信道学习通信（Foerster 等人，2016）。对于大规模群体，密集通信效率低下且噪声较多，也难以区分消息的效用。选择性方案通过学习何时进行通信（IC3Net）（Singh 等人，2019），或在带宽约束下调度部分发言者（Kim 等人，2019），以减少传输。然而，这些方法通常每一步都要重新决定通信，并依赖集中式训练信号，因此随着规模扩大，通信的信用分配变得愈发困难。为缓解这一可扩展性问题，SCoUT 通过软分组引入时间抽象，使其持续 K 步数，并利用反事实邮箱信号高效训练发送和接收方选择。

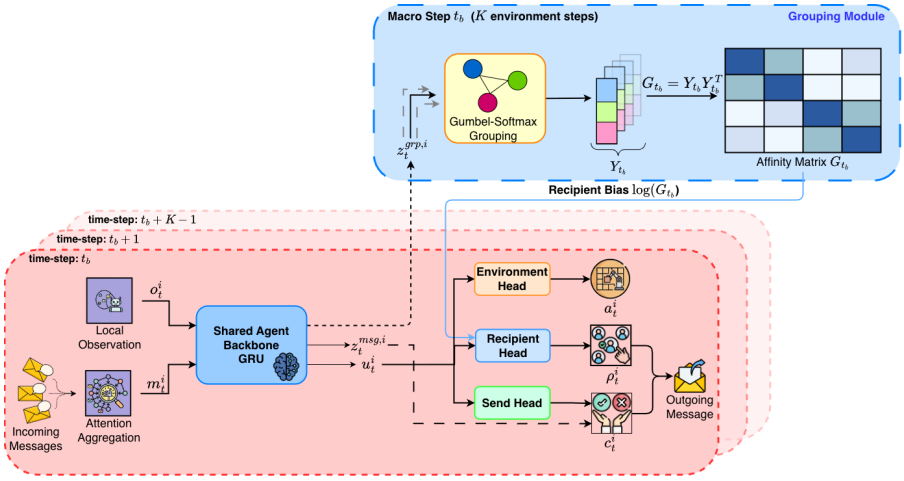

图 1：SCoUT 在两个时间尺度上的前向传播概览。在每个宏步边界 t_b，分组模块会采样软分配 Y_{t_b} 并构造亲和矩阵 $G_{t_b} = Y_{t_b} Y_{t_b}^\top$；该矩阵在接下来的 K 原始步内保持固定，并作为对数偏置 $\log(G_{t_b})$ 用于接收方选择。在每个原始步 t，每个智能体都会嵌入其局部观测 o_t^i 和邮箱输入m_t^i，更新共享的门控循环单元主干网络，并输出一个三头策略：环境动作 a_t^i、发送决策 c_t^i和接收方 ρ_t^i。如果满足 $c_t^i = 1$，智能体便会向选定的接收方传输消息内容 $x_t^i = z^{\text{msg}}{}_{,t}{}^{i}$；接收方会将传入消息汇总到下一步的邮箱中。

定向路由、注意力与结构。除了“发言或保持沉默”之外，定向路由方法学习与谁进行通信。ATOC 通过注意力形成临时协调组（Jiang 和 Lu，2018年），TarMAC 学习基于注意力的寻址（Das 等人，2019），G2ANet 利用多阶段注意力对关系进行建模（Liu 等人，2020）；ToM2C 纳入了心智理论式推理（Wang 等人，2022年）。这些方法提供了比广播更丰富的结构，但往往需要在众多潜在接收方上评估路由，并且在大规模 N下每一步重新计算时可能代价高昂或不稳定。与此同时，拓扑设计方法主张固定全局结构；

例如，ExpoComm 利用指数图拓扑实现小直径传播（Li 等人，2025）。SCoUT 介于这些方向之间：它学习一种随时间缓慢变化的任务自适应分组结构，并将其作为接收方选择上的结构化偏置，而无需固定拓扑或逐步全局重路由。

可扩展性与多智能体基准测试。人们还通过平均场近似来提升多智能体强化学习的可扩展性，即用聚合效应取代多体相互作用（Yang 等人，2018年）。标准化基准测试能够在不同场景下进行评估，包括 SMAC（Samvelyan 等人，2019）以及 MAgent 中的多智能体网格世界（Zheng 等人，2018年）。SCoUT 面向大规模有限群体而设计，这些群体具有异质的局部观测和结构化的协调需求；在此类场景中，密集通信和整体式评论家并不切实际，通信决策的信用分配问题也尤为突出。

3 问题设定

我们考虑部分可观测条件下的合作式多智能体强化学习，并将其建模为去中心化部分可观测马尔可夫决策过程（Dec-POMDP）（Bernstein 等人，2002年）。一个 Dec-POMDP 定义为 $\langle \mathcal{S}, \{\mathcal{A}_i\}_{i=1}^N, \mathcal{P}, r, \{\mathcal{O}_i\}_{i=1}^N, \mathcal{Z}, \gamma \rangle$，其中 $s_t \in \mathcal{S}$ 是环境状态，每个智能体 $i \in \{1, \dots, N\}$ 选择一个环境动作 $a_t^i \in \mathcal{A}_i$，转移核为 $\mathcal{P}(s_{t+1} \mid s_t, a_t^1, \dots, a_t^N)$，且所有智能体共享合作奖励 $r(s_t, a_t^1, \dots, a_t^N)$，折扣因子为 $\gamma \in (0, 1)$。每个智能体接收由 $\mathcal{Z}(o_t^1, \dots, o_t^N \mid s_t)$ 生成的局部观测 $o_t^i \in \mathcal{O}_i$，并根据其局部历史 $h_t^i = (o_{1:t}^i, a_{1:t-1}^i)$ 采取行动。

通信接口。除环境动作外，智能体还可以通过有向点对点信道进行通信。在时间 t，智能体 i 选择：（i）二元发送决策 $c_t^i \in \{0, 1\}$；（ii）如果 $c_t^i = 1$，则选择接收方索引 $\rho_t^i \in \{1, \dots, N\}$；以及（iii）具有固定维度 d 的消息向量 $x_t^i \in \mathcal{X} \subseteq \mathbb{R}^d$。我们假设采用同步的单步延迟：在时间 t 发送的消息将在时间 $t+1$ 对接收方可用。由于接收方可能在同一时间步接收多条消息，我们通过聚合算子将其传入通信表示为固定维度的邮箱 向量 $m_{t+1}^j \in \mathbb{R}^{d_m}$

$$m_{t+1}^j \;=\; g\big(\{x_t^i :\; c_t^i = 1,\; \rho_t^i = j\}\big)\,,$$

其中，g 将传入消息多重集映射为单个向量；如果没有消息发送给 j，则 $m_{t+1}^j = \mathbf{0}$。

策略与目标。智能体 i 的随机去中心化策略是在给定其局部信息的情况下，对其决策进行建模的条件分布，即 $\pi^i(\cdot \mid h_t^i,\ m_t^i)$。我们假设采用参数共享，并为所有智能体使用同一个策略 π_θ（Christianos 等人，2021年；Terry 等人，2020）：

$$(a_t^i, c_t^i, x_t^i, \rho_t^i) \sim \pi_\theta(\cdot \mid h_t^i, m_t^i), \qquad \forall i \in \{1, \dots, N\}.$$

这会产生一个在给定各智能体局部输入的条件下进行因子分解的联合策略，

$$\boldsymbol{\pi}_\theta(\mathbf{a}_t, \mathbf{c}_t, \mathbf{x}_t, \boldsymbol{\rho}_t \mid \mathbf{h}_t, \mathbf{m}_t) = \prod_{i=1}^{N} \pi_\theta(a_t^i, c_t^i, x_t^i, \rho_t^i \mid h_t^i, m_t^i).$$

我们的目标是最大化期望折扣团队回报

$$J(\boldsymbol{\pi}_\theta) = \mathbb{E}_{\boldsymbol{\pi}_\theta}\left[\sum_{t=0}^{\infty} \gamma^t\, r(s_t, a_t^1, \dots, a_t^N)\right]. \tag{1}$$

学习设置（集中式训练、分散式执行）。我们采用集中式训练、分散式执行（CTDE）：执行时，每个智能体仅使用 $(h_t^i$、$m_t^i)$ 行动；训练期间，我们则可以利用额外的全局信息（在我们的情形中即全局状态）来学习评论家或辅助组件（Lowe 等人，2017年；Foerster 等人，2018年）。

4 SCoUT

现在，我们介绍用于合作多智能体强化学习的可扩展效用引导通信框架。SCoUT：（i）通过软分组引入缓慢变化的潜在通信结构，而非每一步都重新确定稠密通信图；（ii）利用该结构，通过分组感知评论家简化并分解价值估计；（iii）训练定向通信决策，包括二元发送动作 $c_t^i \in \{0,1\}$ 和接收方选择 ρ_t^i，并使用反事实优势将效用归因于单条消息。图 1 展示了 SCoUT 训练过程的前向传播概览，重点呈现宏步分组以及逐步执行的三头通信策略。

4.1 智能体主干网络与描述符

每个智能体运行共享的循环主干网络，以在部分可观测性下维护记忆。令$\varphi(\cdot)$ 嵌入拼接后的局部观测和邮箱输入。我们使用门控循环单元（ u_t^i， Cho 等人，2014年）：

$$u_t^i = \mathrm{GRU}\big(\varphi([o_t^i,\, m_t^i]),\, u_{t-1}^i\big).$$

根据 u_t^i，我们通过两个可学习的投影计算轻量级描述符：(i) 分组描述符 $z^{\mathrm{grp}}{}_t{}^i$，仅每隔 K 步（在宏步边界处）计算一次；(ii) 消息描述符$z^{\mathrm{msg}}{}_t{}^i$，在智能体发送消息时用作消息内容。所有表征参数均通过来自策略与价值目标，以及下文介绍的分组和通信组件的梯度进行端到端训练。

4.2 时间软分组与亲和性先验

大型团队通信中的一个核心难题是，每一步都选择有向通信模式会引入大量相互耦合的离散决策；在奖励稀疏或延迟的情况下，这可能导致梯度方差过高，并使信用分配不稳定。SCoUT 通过引入缓慢变化的潜在结构来缓解这一问题：每个宏步 b (每隔 K 个原始步)，我们采样软分组，将其用作 (i) 粗粒度且持久的通信邻域概念，以及 (ii) 下游组件可重复使用的结构（见图1上方面板）。

令 t_b 表示宏步 b 的第一个原始步。我们使用智能体 i 的分组描述符 $z^{\mathrm{grp}}{}_{t_b}{}^i$ 以及一组可学习的群组原型 $\{p_g\}_{g=1}^M$，将每个智能体分配到 $M \ll N$ 个潜在群组之一。直观而言，原型充当锚点，类似于软 k-均值质心。我们通过余弦相似度计算分配 logits：

$$\ell_{t_b}^{i,g} = \alpha \cdot \frac{\langle z_{t_b}^{\mathrm{grp},i},\, p_g \rangle}{\|z_{t_b}^{\mathrm{grp},i}\|\,\|p_g\|}, \qquad g \in \{1,\dots,M\},$$

其中 $\alpha > 0$是一个可学习的规模。随后，我们使用 Gumbel–Softmax 重参数化（软分配进行采样（Jang 等人，2017年；Maddison 等人，2017年）：

$$Y_{t_b}^{i,g} = \mathrm{softmax}\left(\frac{\ell_{t_b}^{i,\cdot} + \epsilon^{i,\cdot}}{\tau}\right)_g, \quad \epsilon^{i,g} \sim \mathrm{Gumbel}(0,1),$$

温度为 τ时，得到 $Y_{t_b} \in \mathbb{R}^{N\times M}$。遵循可微池化领域的常见做法（Ying 等人，2018年），我们构建亲和矩阵

$$G_{t_b} \;=\; Y_{t_b} Y_{t_b}^{\top} \in \mathbb{R}^{N\times N}.$$

其中，当智能体 i 和 j 被分配到相似的组时，$G_{t_b}^{ij}$ 较大。该亲和度在整个宏步内保持固定，并作为通信伙伴的可微先验。

分组目标。 由于分组是一种随机潜在决策，我们使用得分函数目标对其进行优化（Williams，1992），并采用宏步优势信号。令 $P^{i,\cdot}_{t_b,\tau}{}^{g} = \mathrm{softmax}(\ell^{i,\cdot}_{t_b}/\tau)^g$ 表示分配概率，令 $A^{\mathrm{grp},i}_{b}$ 表示分组感知评论家（第 4.4 节）产生的宏步优势。我们最小化

$$\mathcal{L}_{\mathrm{grp}} = \mathcal{L}^{\mathrm{PG}}_{\mathrm{grp}} + \lambda_{\mathrm{bal}}\mathcal{L}^{\mathrm{bal}}_{\mathrm{grp}} + \lambda_{\mathrm{ent}}\mathcal{L}^{\mathrm{ent}}_{\mathrm{grp}} + \lambda_{\mathrm{edge}}\mathcal{L}^{\mathrm{edge}}_{\mathrm{grp}},$$

其中，λ_{bal}、λ_{ent}、$\lambda_{\mathrm{edge}} \geq 0$ 是由用户选择的系数，用于控制各正则项的强度。策略梯度项为

$$\mathcal{L}^{\mathrm{PG}}_{\mathrm{grp}} = -\sum_{b}\sum_{i=1}^{N} A^{\mathrm{grp},i}_{b} \sum_{g=1}^{M} Y^{i,g}_{t_b} \log P^{i,g}_{\tau},$$

该项将得分函数权重应用于采样分配。我们还通过正则化避免退化或过早坍塌：

$$\mathcal{L}^{\mathrm{bal}}_{\mathrm{grp}} = \left\| \frac{1}{N}\sum_{i=1}^{N} P^{i}_{\tau} - \frac{1}{M}\mathbf{1} \right\|_2^2, \qquad \mathcal{L}^{\mathrm{ent}}_{\mathrm{grp}} = -\frac{1}{N}\sum_{i=1}^{N} H(P^{i}_{\tau}),$$

其中，$\mathcal{L}^{\mathrm{bal}}_{\mathrm{grp}}$ 鼓励群组使用大致均衡，$\mathcal{L}^{\mathrm{ent}}_{\mathrm{grp}}$ 鼓励在训练初期进行探索。最后，$\mathcal{L}^{\mathrm{edge}}_{\mathrm{grp}}$鼓励学习得到的亲和度 G_{t_b} 为那些根据我们学习得到的成对效用信号（第4.5节）预测通信有用的智能体对赋予更高权重。

4.3 三头策略与邮箱通信

每个智能体使用一个共享策略 $\pi_\theta(\cdot \mid u^i_t)$，该策略包含三个头：(i) 环境动作头，(ii) 二元发送头 $c^i_t \in \{0,1\}$，以及 (iii) 一个针对 $\rho^i_t \in \{1,\dots,N\}$的接收方头。主干网络生成两个描述向量（第4.1节）：$z^{\mathrm{msg}}_{t}{}^{i} \in \mathbb{R}^d$，用作智能体的消息表示；以及 $z^{\mathrm{grp}}_{t_b}{}^{i}$，用于在宏步边界进行分组。当智能体 i 发送消息时（$c^i_t = 1$），它会传输消息向量 $x^i_t \triangleq z^{\mathrm{msg}}_{t}{}^{i} \in \mathbb{R}^d$，并从接收方头中采样一个接收方 ρ^i_t（见图1底部面板）。

基于组亲和度的接收方偏置。令 $\phi^{\mathrm{recv}}_{t}{}^{i} \in \mathbb{R}^N$ 表示智能体 i的策略头在时刻 t生成的接收方逻辑值。为在不施加硬性约束的情况下鼓励群组内通信，我们添加一个依赖于宏步、由亲和矩阵 G_{t_b}导出的对数偏置：

$$\tilde{\phi}^{\mathrm{recv},i}_t(j) = \phi^{\mathrm{recv},i}_t(j) + \log\!\big(G^{ij}_{t_b} + \varepsilon\big), \quad j \in \{1,\dots,N\},$$

其中，t_b 是包含 t 的宏步，$\varepsilon > 0$ 是一个较小的常数。随后，我们从 $\mathrm{softmax}(\tilde{\phi}^{\mathrm{recv}}_{t}{}^{i})$中采样 ρ^i_t。

邮箱聚合（选择 g）。如第3节所述，接收方可能会收到数量不定的传入消息；因此，我们使用聚合算子 $g(\cdot)$将收到的消息向量多重集映射为固定维度的邮箱 m^j_{t+1}。这为下一步的策略提供了一个置换不变、大小固定的输入。在我们的实现中，我们选择了一种简单的无参数缩放点积注意力规则（Vaswani 等人，2017年）：

$$m^j_{t+1} = \sum_{i:\,\rho^i_t = j} w_{ij}\, x^i_t, \quad w_{ij} = \frac{\exp\!\big(\langle x^j_t, x^i_t\rangle/\sqrt{d}\big)}{\sum_{i':\,\rho^{i'}_t = j} \exp\!\big(\langle x^j_t, x^{i'}_t\rangle/\sqrt{d}\big)}, \tag{2}$$

其中 d 是消息维度。直观而言，接收方自身的消息描述符 x^j_t 充当查询，对作为键和值的传入消息描述符进行加权。

4.4 分组感知评论家与分组优势

分组感知评论家。 在集中式训练、分散式执行下，我们学习一个集中式价值函数，用于根据全局状态估计期望的折扣团队回报。为降低大规模群体的输出复杂度，SCOUT在群组 层面预测值，并利用软分配将其转换为用于优势估计的逐智能体价值基线。具体而言，给定 s_t，评论器输出$\mathbf{v}_t = [v_t^1, \dots, v_t^M]^\top = V_\psi(s_t) \in \mathbb{R}^M$，而逐智能体基线是按分配加权的混合

$$V_t^i = \sum_{g=1}^{M} P_\tau^{i,g}\, v_t^g = \big(P_\tau \mathbf{v}_t\big)_i.$$

我们训练 V_ψ，采用标准的近端策略优化/广义优势估计价值回归目标 $\mathcal{L}_V$：我们根据广义优势估计构造自举目标（Schulman 等人，2016），并最小化平方误差损失（Schulman 等人，2017年）。

分组优势。在每个宏步边界 t_b，我们通过将学习到的分配下预测的逐智能体价值与置换基线进行比较，为分组模块导出优势信号。令 σ 为 $\{1, \dots, M\}$ 的一个随机置换，并定义置换后的群组价值向量$(\mathbf{v}_{t_b}^\sigma)_g \triangleq \mathbf{v}_{t_b,\ \sigma(g)}$。我们设定

$$A_b^{\mathrm{grp},i} = \underbrace{\big(P_\tau \mathbf{v}_{t_b}\big)_i}_{\text{value under learned grouping}} - \underbrace{\big(P_\tau \mathbf{v}_{t_b}^\sigma\big)_i}_{\text{permutation baseline}}.$$

减去第二项起到REINFORCE式基线的作用。它在保持预测值规模的同时独立于特定的群组标注，从而降低学习分组分布时的梯度方差（Williams，1992；Greensmith等人，2004年）。

4.5 通信评论器与反事实邮箱信用分配

学习式通信的一大核心难题是信用分配：多个同时发出的消息可能共同影响未来的奖励。SCoUT 学习一个带有两个头的辅助通信评论家：(i) 用于接收者评估的消息条件价值头 $V_{\mathrm{msg}\omega}$，以及 (ii) 用于计算发送者—接收者效用的成对效用头 Q_ω。我们通过最小化一个批次中 Sutton、1988 所定义的 T 个原始步上的一步时序差分（TD）误差平方来训练该评论器，并采用 Polyak 平均（Polyak 和 Juditsky，1992；Lillicrap 等人，2016年）。

反事实发送优势。对于发送方 i 向接收方 $j = \rho_t^i$ 发送消息的情况，我们计算一个反事实邮箱：移除 i 的贡献，同时保持其他所有传入消息不变；这一做法遵循 Foerster 等人（2018年）讨论的反事实信用分配范式。对于发送方 i 向接收方 $j = \rho_t^i$ 发送消息的情况，令 $m_{t+1}^j = g(\{x_t^{i'} :\ c_t^{i'} = 1,\ \rho_t^{i'} = j\})$ 如第 3 节所定义。我们通过移除 i 的消息，并重新应用相同的聚合器 $m_{t+1}^{j\setminus i} = g\Big(\{x_t^{i'} :\ c_t^{i'} = 1,\ \rho_t^{i'} = j,\ i' \neq i\}\Big)$.，来定义留一反事实邮箱。在我们的实现中，$g$ 是式（2）中的无参数注意力规则，因此这会简化为归一化留一注意力聚合：

$$m_{t+1}^{j\setminus i} = \sum_{i' \in \mathcal{S}_t(j)\setminus\{i\}} \frac{w_{i'j}}{\sum_{k \in \mathcal{S}_t(j)\setminus\{i\}} w_{kj}}\, x_t^{i'}.$$

随后，我们将发送优势定义为

$$A_t^{\mathrm{send},i} = V_\omega^{\mathrm{msg}}\big(s_{t+1},\, f_{t+1}^j(m_{t+1}^j)\big) - V_\omega^{\mathrm{msg}}\big(s_{t+1},\, f_{t+1}^j(m_{t+1}^{j\setminus i})\big),$$

其中 $f_{t+1}^j(\cdot)$ 表示使用指定邮箱输入的接收方特征构造。

R接收方优势与边效用对齐。令时刻 t b e

$$U_t^{ij} := Q_\omega\big(s_{t+1},\, f_t^i,\, f_{t+1}^j\big),$$

对于有效的发送方-接收方对（$i,\ j$）。对于发送方 i 及其选定的接收方 ρ_t^i，我们构造一个中心化的接收方优势

$$A_t^{\mathrm{recv},i} = U_t^{i\rho_t^i} - \frac{1}{|\mathcal{R}_t^i|}\sum_{j\in\mathcal{R}_t^i} U_t^{ij},$$

其中，$\mathcal{R}_t^i$ 是 i 的可能接收方集合。为使分组与通信效用保持一致，我们将 $\tilde{G}^i_{\cdot,\cdot}$ 和 $\tilde{U}^i_{\cdot,\cdot}$ 分别定义为亲和度行和效用行经过行中心化及 ℓ_2 归一化后的版本，并设定

$$\mathcal{L}_{\mathrm{grp}}^{\mathrm{edge}} = -\frac{1}{N}\sum_{i=1}^{N}\frac{\langle \tilde{G}^{i,\cdot}_{t_b},\, \tilde{U}^{i,\cdot}_{t_b}\rangle}{\|\tilde{G}^{i,\cdot}_{t_b}\|\,\|\tilde{U}^{i,\cdot}_{t_b}\|}.$$

这类似于计算两个矩阵的行向量之间的余弦相似度（Brockmeier 等人，2017年），并促使群组保持通信一致性：当学习到的效用预测智能体彼此是有用的通信伙伴时，就将它们分到同一组；否则则促使它们彼此远离。

4.6 整体优化与分散执行

我们使用近端策略优化（Schulman 等人，2017年）训练三头策略，其中使用：（i）针对环境动作的标准广义优势估计优势（Schulman 等人，2016）；以及（ii）上文针对发送方头和接收方头的通信优势 $A^{\mathrm{send},i}_t$ 和 $A^{\mathrm{recv},i}_t$。令 $\mathcal{L}^{\mathrm{env}}_{\mathrm{PPO}}$，$\mathcal{L}^{\mathrm{send}}_{\mathrm{PPO}}$，$\mathcal{L}^{\mathrm{recv}}_{\mathrm{PPO}}$ 表示各个头对应优势的标准裁剪近端策略优化替代目标。完整目标结合了策略、环境评论器、通信评论器和分组损失：

$$\mathcal{L} = \mathcal{L}^{\mathrm{env}}_{\mathrm{PPO}} + \mathcal{L}^{\mathrm{send}}_{\mathrm{PPO}} + \mathcal{L}^{\mathrm{recv}}_{\mathrm{PPO}} + c_V\mathcal{L}_V + c_\omega\mathcal{L}_\omega + c_{\mathrm{grp}}\mathcal{L}_{\mathrm{grp}},$$

其中 $\mathcal{L}_\omega$ 表示第 4.5 节所述的通信评论家损失。系数c_V、c_ω、$c_{\mathrm{grp}} \geq 0$ 是由用户设定的标量权重，用于权衡辅助损失与近端策略优化策略损失；我们将其视为超参数。在测试时，SCoUT 会舍弃集中式训练组件，仅由每个智能体执行共享的分散式策略。

5 实验

我们在两个大种群规模的多智能体基准测试上评估 SCOUT，目标有两个：(i) 展示 SCOUT 能够直接在数百个智能体 的规模下训练，同时保持强大的协调能力；(ii) 分析哪些设计选择带来了这些提升。

5.1 基准测试

MAgentBattle。 我们考察 MAGENT 中的 BATTLE (Zhengetal.，2018年) 场景。这是一个竞争性的双团队网格世界，每个团队必须在固定回合时限内协同作战，消灭对方团队。每一步中，每个智能体都可以选择移动或攻击；攻击会降低目标的生命值，当智能体的生命值降至零时便会被淘汰（有关 BATTLE的更多细节见附录 A.1）。我们评估四种种群规模： 20对20、64对64、81对81和100对100。值得注意的是，该基准测试中以往的结构化通信结果通常只报告到64对64；我们通过 81对81 和 100对100 设置，明确对超越这一范围的可扩展性进行压力测试。

追捕（SISL）。我们还在Pursuit （PettingZoo SISL 套件中的环境，Terry 等人，2021年）上进行评估。该环境由P 个追捕者和 E 个逃避者组成（E<P）。当逃避者周围所有空闲相邻格均被追捕者占据时，它就会被捕获；因此，在开放空间中需要 4 个追捕者，沿墙壁或障碍物时需要 3 个，在角落则需要 2 个（有关 Pursuit 的更多详情，请参见附录A.2）。我们

主要使用Pursuit 进行针对性消融实验，以了解 SCoUT 的哪些组成部分对于实现强协调是必要的。我们评估了五种规模：20P-8E、40P-16E、60P-24E、80P-32E和100P-40E。不同于默认的 Pursuit 配置（8 个追捕者、30 个逃避者）（Terry 等人，2021年；Gerstgrasser 等人，2023年），我们研究了以下设置 $E < P$，以对智能体数量方面的可扩展性进行压力测试，同时在固定的 500 步时域内确保捕获里程碑仍具有信息量。

在所有基准测试中，我们使用固定的宏步长度 $K{=}10$，并按照简单的、依赖于规模的规则设置组数。对于 Battle： $M = \lfloor \sqrt{N} \rfloor$，每队有 N 个智能体；对于 Pursuit： $M = \lfloor E/2 \rfloor$，有 E 个逃避者。附录B 汇总了实验中使用的所有超参数，附录 D 则研究了 M 和 K的选择敏感性。

计算资源。 所有实验均在一台配备 NVIDIA L40 GPU 和 AMD CPU（32 个核心，4.1GHz）的机器上运行。

5.2 MAgent Battle 上的可扩展竞争性协调

设置与待检验主张。 本实验评估随着种群规模增大，SCOUT 是否仍能保持稳健且果断的竞争性协调。与以往强调从小规模进行N 训练后实现零样本迁移的工作不同，我们在目标规模（最高可达 100对100）下直接训练每种方法，以考验大规模实例场景下的优化稳定性。所有方法均训练了 4M 步数。

基线方法。 我们将一种非通信基线方法与若干学习式通信基线方法进行比较：

(i) **独立深度 Q 学习（IDQN）**(Rashid 等人，2020；Mnih 等人，2015），每个智能体根据局部观测进行学习，不进行显式消息交换。(ii) **CommFormer**(Hu 等人，2024）采用 Transformer 风格消息编码器，通过注意力机制处理其他智能体的通信令牌。与固定拓扑方案相比，这种方法支持基于内容的路由，并能更丰富地建模交互。我们将其作为具有注意力机制的学习式通信基线代表。(iii) **ExpoComm**(Li 等人，2025），旨在通过指数拓扑邻域实现多智能体通信。我们在相同的每智能体接收者预算下评估 对等方-$n{=}7$ （时变邻域）和 静态-$n{=}7$ （固定邻域）。

指标与评估流程。我们在每种设置下使用 20 个独立评估随机种子评估每个训练好的策略。我们报告：(i) 胜率（红队存活智能体数量多于蓝队的回合所占百分比）；(ii) 淘汰率（蓝队被淘汰的比例，即蓝队击杀数除以 N）；以及 (iii) 通过到达里程碑的时间衡量决胜能力：达到 50% / 75% 蓝队淘汰率所需的 TT_{50} / TT_{75} =步数（时域=200）。此外，我们报告达成率 R_{50}/R_{75}, ，即在时域内达到各里程碑的评估回合所占百分比。仅当 $\text{R}_k \geq 50\%$ 时，我们才报告 TT_k （否则为不适用）。此外，我们绘制训练曲线（回合回报与环境步数的关系；图2），以评估在各目标规模下直接训练时的优化稳定性和样本效率。最后，为了刻画学习式通信与发现的交互结构之间的契合程度，我们在所有规模下报告组内消息比例（附录 C.3（表8））。我们还展示了 100 对 100 场景的定性策略展开（附录 C.1）。

结果与讨论。 表 1 与图 2显示出两种一致的模式[1]。 **（1）SCoUT在所有规模上均能可靠训练，且高度可重复**。 SCoUT对所有 N均实现100%胜率，并维持近乎完全歼灭（95–99%），歼灭时间与里程碑时间在20个评估随机种子上表现出极低方差。训练同样稳定：在全部四个规模上，SCoUT呈现出清晰的提升阶段，随后稳定收敛，且跨随机种子离散程度显著低于基线方法（图 2）；值得注意的是，这种稳定性即使在最大

[1]较大规模下的 CommFormer 训练曲线被截断：81对81 达到了计划 400 万步数的 ≈50%，而 100对100 达到了 ≈35%（两者均训练了 6 天， ∼4× 与 SCOUT 使用的实际运行时间预算相同）。

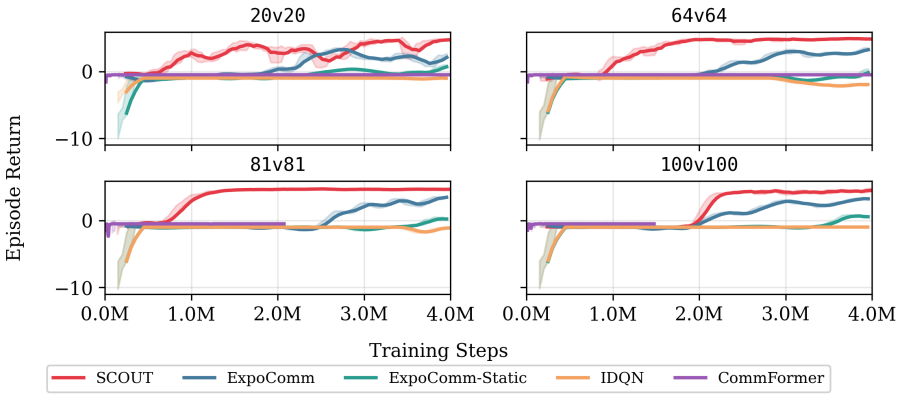

图2：**不同规模下的 Battle 训练曲线**。回合回报与环境步数的关系（阴影区域表示不同随机种子之间的差异）。SCOUT 在不同规模下均能快速且稳定地学习，而基线方法表现出显著的规模和拓扑敏感性。

方法	20对20	64对64	81对81	100对100	20对20	64对64	81对81	100对100
胜率 *(%)*					淘汰率 *(%)*（蓝色被淘汰 / N）			
SCOUT	**100.0**±0.0				**95**±6	**98**±4	**99**±3	**99**±3
ExpoComm	0.0±0.0	95.0±4.9	96.0±4.2	96.4±2.7	11±16	94±12	72±16	75±21
ExpoComm-Static	0.0±0.0	0.0±0.0	60.0±11.0	65.0±10.7	3±8	1±2	57±30	36±10
IDQN	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0	0	7±2	0
CommFormer	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0	0	0	0
里程碑达成率 *(%)* (R_{50} / R_{75})					TT_{50} /TT_{75}（步数；时域=*200*）			
SCOUT	**100/ 100**					**24**±2 / **29**±1	**27**±3 / **32**±2	**31**±4 / **39**±2
ExpoComm	0 / 0	95 / 95	90 / 55	80 / 60	N/A	59±20/ 81±24	100±44/ 182±14	101±21/ 160±23
ExpoComm-Static	0 / 0	0 / 0	60 / 60	10 / 0	N/A	N/A	85±7 / 153±20	N/A
IDQN	0 / 0	0 / 0	0 / 0	0 / 0	N/A	N/A	N/A	N/A
CommFormer	0 / 0	0 / 0	0 / 0	0 / 0	N/A	N/A	N/A	N/A

表1：**不同规模下的 MAgent Battle 结果（20 个评估随机种子）**。我们报告胜率、淘汰率、**里程碑达成率**R_{50}/R_{75},以及**决胜能力**，通过TT_{50}/TT_{75}。

群体（81对81 与100对100）时依然保持，其中SCOUT相比基线方法收敛得更平滑、更迅速。在规模谱系的另一端（20对20），SCOUT仍能可靠学习并显著优于基线方法，而EXPOCOMM 表现始终较差，且未出现持续的学习进展。在各个规模上，静态拓扑变体均逊于其自适应对应方法，这表明在有效协调伙伴随时间变化的动态战斗中，固定邻域过于受限。 **（2）SCOUT能更一致、更迅速地将胜利转化为决定性歼灭**。 在更大的 N，ExpoComm基线方法可以取得较高胜率，但往往留下大量敌方剩余数量，且无法在时域内达到75%歼灭；基线方法的里程碑达成率随规模下降，表明许多获胜回合并不能始终将优势转化为高歼灭结果。即使达到里程碑，SCOUT达到这些里程碑的速度也显著更快（表 1），这与其更平滑、更稳定的学习动态相一致（图 2）。值得注意的是，该方法是唯一取得R_{50}=R_{75}=100% 、且在每个规模上均如此的方法，从而证实其始终实现决定性歼灭，而非偶尔的快速胜利。

5.3 Pursuit 消融实验

设置与待检验主张。我们使用 PURSUIT 来分离 SCOUT 中哪些要素推动了合作捕获中的可扩展多智能体协调。所有方法均直接在各目标规模下训练 4M 步。

变体与基线。 我们将完整方法与针对性移除版本及一种可扩展通信基线进行比较：

方法	20P-8E									
捕获率（捕获率）						完成率（完成率）				
SCOUT	**94±10**	**93±8**	**89±5**	**90±6**	**88±5**	**70**	**50**	**0**	**5**	**0**
无反事实	72±21	18±20	28±27	29±13	58±26	20	0	0	0	0
不分组	57±18	10±9	14±6	37±10	41±10	0	0	0	0	0
ExpoComm（对等方-n=7）	86±14	58±14	62±12	72±13	69±10	40	0	0	0	0

方法	20P-8E	40P-16E	60P-24E	80P-32E	100P-40E
TT_{50}（步数；时域=*500*）（括号内为 R_{50}%）					
SCOUT	**74±78 (100)**	**143±52 (100)**	**72±19 (100)**	**98±38 (100)**	**70±18 (100)**
无反事实	173±149 (90)	不适用 (15)	不适用 (20)	不适用 (5)	173±86 (55)
不分组	212±184 (80)	不适用 (0)	不适用 (0)	不适用 (15)	不适用 (20)
ExpoComm (对等方-n=7)	125±73(100)	191±114(80)	307±80 (85)	184±93 (100)	137±66 (95)
TT_{75}（步数；时域=*500*）（括号内为 R_{75}%）					
SCOUT	**138±82 (95)**	**227±65 (95)**	**139±64(100)**		
无反事实	214±136 (65)	不适用 (0)	不适用 (15)	不适用 (0)	不适用 (35)
不分组	不适用 (30)	不适用 (0)	不适用 (0)	不适用 (0)	不适用 (0)
ExpoComm（对等方-n=7）	234±114（80）	不适用（15）	不适用（20）	不适用（40）	425±59 （50）

表2：**不同规模下的 Pursuit 消融实验（20 个评估随机种子；时域=500）**。我们报告捕获率、完成率以及里程碑时间 TT_{50}/TT_{75}。对于里程碑行，每个条目显示为 TT_k （R_k），其中 R_k 表示达到 k% 捕获里程碑的回合所占百分比。

(i) **SCoUT：** 学习得到的分组 +分组感知评论家 + 反事实通信优势。(ii) **无反事实：** 移除用于通信决策的反事实分解（改用标准的广义优势估计（Schulman 等人，2016））。(iii) **不分组：** 移除学习得到的分组；因此，禁用分组感知评论家。(iv) **ExpoComm（对等方-n=7）：** 与第 5.2 节相同的学习式通信基线。

指标与评估流程。 我们在每种设置下，使用 20 个独立的评估随机种子评估每个训练好的策略。我们报告：(i) **捕获率** (逃避者被捕获的比例)；(ii) **完成率** (捕获所有逃避者的回合所占百分比)；以及 (iii)**决胜能力**，采用与第 5.2 节相同的里程碑指标：R_{50}/R_{75} (里程碑达成率)和TT_{50}/TT_{75} (到达里程碑的时间)，但里程碑依据 捕获比例 而非消灭情况来定义。此外，我们通过绘制捕获率和 TT_{50} 随种群规模变化的趋势（图3）来展示规模扩展趋势。最后，与 battle 类似，我们还在附录 PURSUIT 中报告群组内消息比例，见附录 C.3 (表8)，并提供`100P-40E`的定性策略展开（附录 C.2）。

结果与讨论。表2和图3揭示了两种一致的模式。（1）完整的 SCOUT 能够随规模扩展，同时保持高捕获率并可靠地达成里程碑。在全部五个规模上，SCOUT 均保持较高的捕获率，且 R_{50}=100% 表明每个评估回合都能在 500 步内达到 50% 捕获。在中等和大规模下，它也能可靠地达到 75% 捕获，里程碑时间始终较短，且跨随机种子的误差线稳定（图3）。我们将完成率报告为在固定的 500 步时间范围内实现完全捕获；完成率会随规模自然下降（常常降至 0%），因为逃避者数量增加，而完全捕获要求同时包围每一个剩余逃避者，因此即使部分协同捕获很强，在 500 步内完成全部捕获也越来越不可能。（2）隔离实验表明，反事实信用分配与时间分组驱动了可扩展性。移除反事实优势会导致可扩展性突然失效：超过 20P-8E 后，捕获率急剧下降且高度不稳定，里程碑达成率也大幅下降——R_{50} 在 40P-16E/60P-24E 上降至 15-20%，在 80P-32E 上降至 5%），使得 TT_{50}/TT_{75}无法定义。在 100P-40E 上，它仅部分恢复至 R_{50}=55%，但仍无法可靠达到 75%。移除分组在规模扩大时同样脆弱：从 40P-16E 到 100P-40E，捕获率降至 10-41%，在这些更大规模的场景中始终为 $R_{50} \leq 20$% 与 R_{75}=0%，表明-

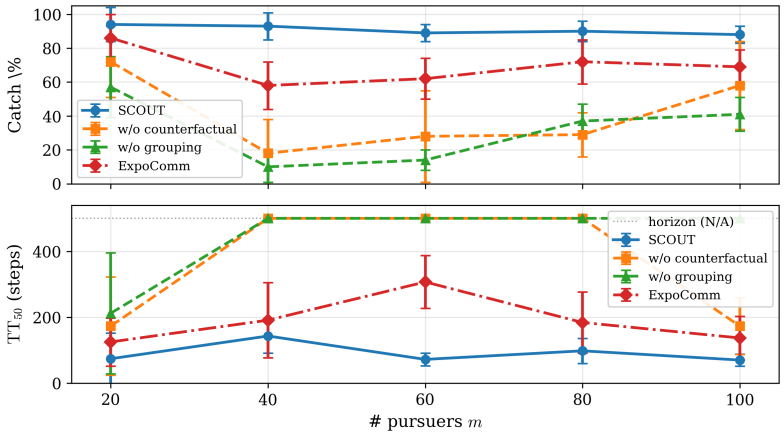

图3：**Pursuit 规模扩展总结**。捕获率（上图）和TT_{50}（下图）随追捕者数量 m的变化（误差线：20 个评估随机种子的标准差）。在下图中，我们绘制了TT_{50}=500 当 R_{50}<50%（表2将这些情况报告为不适用）。

持续无法形成有效的捕获联盟。综合来看，这些消融实验支持如下结论：随着种群规模增长，SCOUT 的反事实通信学习与时间分组对于维持高捕获率以及较早达成里程碑都是必要的。

6 结论

将学习式通信扩展到大型团队并非易事，因为发送方—接收方交互会呈组合式增长，消息级信用也会变得嘈杂。我们提出了 SCOUT，这种方法通过学习缓慢变化的潜在交互结构来实现可扩展通信：它以宏步时间尺度重新采样软分组，将由此得到的亲和度用作接收方的可微先验，并在分组感知评论家中复用同一结构，从而在大规模N下实现可处理的价值学习。SCOUT 还引入了反事实邮箱机制，用于隔离每个发送方对接收方聚合消息的边际贡献，从而为发送方选择和接收方选择提供细粒度的学习信号。在大规模群体基准测试中，SCOUT 可直接在目标规模下训练，并在 N 增长时保持强劲、可靠的协调能力。在 MAGENT BATTLE 中，它取得了完美胜率、近乎彻底的淘汰效果，并持续较早达到里程碑；而最先进基线方法的表现则不够果断，也不够可靠。在 PURSUIT 中，SCOUT 在不同规模下均保持较高的捕获率，并较早达到里程碑。消融实验证实，时间分组和反事实通信信用缺一不可：移除其中任何一项，都会导致中等和大规模下的性能大幅下降。

局限性与未来工作。 SCOUT 引入了训练时结构，尤其是宏步K 和组数 M，目前我们将它们设为固定超参数。尽管同一组设置在我们的基准测试中表现良好，但性能可能取决于 K 与交互时间尺度的匹配程度，以及 M 与智能体行为异质性的匹配程度。此外，我们的留一法邮箱反事实在固定聚合规则下为发送方提供了明确的边际信用分配，但无法捕获多消息交互效应。未来工作包括利用接收方熵或效用漂移等信号在线调整 K 和 M，将信用分配扩展到消息交互建模，并在连续控制、异质性以及明确的通信/能耗预算条件下评估 SCOUT。

参考文献

Daniel S. Bernstein、Robert Givan、Neil Immerman 和 Shlomo Zilberstein。马尔可夫决策过程去中心化控制的复杂性。载于第 *18* 届人工智能不确定性会议（*UAI*），2002年。

Austin J. Brockmeier 等。量化相似度测量的信息量。《机器学习研究杂志》，18(1):1–41，2017年。

Lucian Bu¸soniu、Robert Babuška 和 Bart De Schutter。多智能体强化学习综述。*IEEE*系统、人类与控制论汇刊*C*辑（应用与评论），38(2):156–172，2008年。

Kyunghyun Cho、Bart van Merriënboer、Ça˘glar Gülçehre、Dzmitry Bahdanau、Fethi Bougares、Holger Schwenk 和 Yoshua Bengio。使用循环神经网络编码器—解码器学习用于统计机器翻译的短语表示。载于自然语言处理实证方法会议（*EMNLP*），第1724–1734页，2014年。

Filippos Christianos、Georgios Papoudakis、Muhammad A.Rahman 和 Stefano V. Albrecht。通过选择性参数共享扩展多智能体强化学习。载于第*38*届国际机器学习会议（*ICML*），第1989–1998页，2021年。

Abhishek Das、Théophile Gervet、Joshua Romoff、Dhruv Batra、Devi Parikh、Mike Rabbat 和 Joelle Pineau。TarMAC：定向多智能体通信。发表于第*36*届国际机器学习会议（*ICML*），第1538–1546页，2019年。

Jakob N. Foerster、Yannis M. Assael、Nando deFreitas 和 ShimonWhiteson。通过深度多智能体强化学习学习通信。发表于第*30*届神经信息处理系统国际会议（*NIPS*），第2145–2153页，2016年。

Jakob N. Foerster、Gregory Farquhar、Triantafyllos Afouras、Nantas Nardelli 和 Shimon Whiteson。反事实多智能体策略梯度。发表于第*32*届*AAAI*人工智能会议、第三十届人工智能创新应用会议和第八届*AAAI*人工智能教育进展研讨会，2018年。

Matthias Gerstgrasser、Tom Danino 和 Sarah Keren。选择性地共享经验可改善多智能体强化学习。发表于第*37*届神经信息处理系统大会（*NIPS*），2023年。

Evan Greensmith、Peter L Bartlett 和 Jonathan Baxter。强化学习中梯度估计的方差缩减技术。《机器学习研究杂志》，5:1471–1530，2004年。

Pablo Hernandez-Leal、Bilal Kartal 和 Matthew E. Taylor。多智能体深度强化学习综述与评析。《自主智能体与多智能体系统》，33(6):750–797，2019。

Shengchao Hu、Li Shen、Ya Zhang 和 Dacheng Tao。从图建模视角学习多智能体通信。发表于第*12*届国际学习表征会议（*ICLR*），2024年。

Eric Jang、Shixiang Gu 和 Ben Poole。基于 Gumbel-softmax 的类别重参数化。发表于第*5*届国际学习表征会议（*ICLR*），2017年。

Jiechuan Jiang 和 Zongqing Lu。学习多智能体协作中的注意力通信。载于第*32*届神经信息处理系统国际会议（*NIPS*），第 7265–7275 页，2018年。

Daewoo Kim、Sangwoo Moon、David Hostallero、Wan Ju Kang、Taeyoung Lee、Kyunghwan Son 和 Yung Yi。《在多智能体强化学习中学习调度通信》。载于第七届国际学习表征会议（*ICLR*），2019年。

Sheng Li、Jayesh K. Gupta、Peter Morales、Ross Allen 和 Mykel J. Kochenderfer。《用于多智能体强化学习的深度隐式协调图》。载于第二十届自主智能体与多智能体系统国际会议（*AAMAS*），第764–772页，2021年。Xinran Li、Xiaolu Wang、Chenjia Bai 和 Jun Zhang。《多智能体强化学习中由指数拓扑实现的可扩展通信》。载于第十三届国际学习表征会议（*ICLR*），2025年。Timothy P. Lillicrap、Jonathan J. Hunt、Alexander Pritzel、Nicolas Heess、Tom Erez、Yuval Tassa、David Silver 和 Daan Wierstra。《利用深度强化学习实现连续控制》。载于第四届国际学习表征会议（*ICLR*），2016年。Michael L. Littman。《作为多智能体强化学习框架的马尔可夫博弈》。载于第十一届国际机器学习会议（*ICML*），1994年。Yong Liu、Weixun Wang、Yujing Hu、Jianye Hao、Xingguo Chen 和 Yang Gao。《通过图注意力神经网络实现多智能体博弈抽象》。载于第*34*届*AAA*人工智能会议，第7211–7218页，2020年。Ryan Lowe、Yi Wu、Aviv Tamar、Jean Harb、Pieter Abbeel 和 Igor Mordatch。《用于混合合作竞争环境的多智能体演员-评论器》。载于第*31*届神经信息处理系统国际会议（*NIPS*），第6382–6393页，2017年。Chris J. Maddison、Andriy Mnih 和 Yee Whye Teh。《Concrete分布：离散随机变量的连续松弛》。载于第五届国际学习表征会议（*ICLR*），2017年。Volodymyr Mnih 等人。《通过深度强化学习实现人类水平的控制》。《自然》，518（7540）：529–533，2015年。Yaru Niu、Rohan R Paleja 和 Matthew C Gombolay。《多智能体图注意力通信与协同组队》。载于第二十届自主智能体与多智能体系统国际会议（*AAMAS*），第964–973页，2021年。Boris T. Polyak 和 Anatoli B. Juditsky。《通过平均加速随机逼近》。《*SIAM*控制与优化杂志》，30(4)：838–855，1992年。Tabish Rashid、Mikayel Samvelyan、Christian Schroeder De Witt、Gregory Farquhar、Jakob Foerster 和 Shimon Whiteson。《用于深度多智能体强化学习的单调价值函数分解》。《机器学习研究杂志》，21(178)：1–51，2020年。Mikayel Samvelyan、Tabish Rashid、Christian Schroeder de Witt、Gregory Farquhar、Nantas Nardelli、Tim G. J. Rudner、Chia-Man Hung、Philip H. S. Torr、Jakob Foerster 和 Shimon White- son。《星际争霸多智能体挑战赛》。载于第十八届自主智能体与多智能体系统国际会议（*AAMAS*），第2186–2188页，2019年。John Schulman、Philipp Moritz、Sergey Levine、Michael Jordan 和 Pieter Abbeel。《利用广义优势估计实现高维连续控制》。载于第四届国际学习表征会议（*ICLR*），2016年。John Schulman、Filip Wolski、Prafulla Dhariwal、Alec Radford 和 Oleg Klimov。《近端策略优化算法》。*arXiv*预印本 *arXiv:1707.06347*，2017年。Amanpreet Singh、Tushar Jain 和 Sainbayar Sukhbaatar。《学习在多智能体合作与竞争任务中如何实现规模化通信》。载于第七届国际学习表征会议（*ICLR*），2019年。Kyunghwan Son、Daewoo Kim、Wonseok Kang、David Hostallero 和 Yung Yi。《QTRAN：通过变换学习分解合作式多智能体强化学习》。载于第*36*届国际机器学习会议（*ICML*），第5887–5896页，2019年。

Sainbayar Sukhbaatar、Arthur Szlam 和 Rob Fergus。通过反向传播学习多智能体通信。载于第*30*届神经信息处理系统国际会议（*NIPS*），第2252–2260页，2016年。Peter Sunehag、Guy Lever、Audrunas Gruslys、Wojciech Marian Czarnecki、Vinicius Zambaldi、Max Jaderberg、Marc Lanctot、Nicolas Sonnerat、Joel Z. Leibo、Karl Tuyls 和 Thore Graepel。基于团队奖励的合作式多智能体学习价值分解网络。载于第*17*届自主智能体与多智能体系统国际会议（*AAMAS*），2018年。Richard S Sutton。通过时序差分方法学习预测。机器学习，3（1）：9–44，1988年。Jordan Terry、Benjamin Black、Nathaniel Grammel、Mario Jayakumar、Ananth Hari、Ryan Sullivan、Luis S Santos、Clemens Dieffendahl、Caroline Horsch、Rodrigo Perez-Vicente 等。Petting-zoo：用于多智能体强化学习的 Gym。载于第*35*届神经信息处理系统国际会议（*NIPS*），第15032–15043页，2021年。

Justin K Terry、Nathaniel Grammel、Sanghyun Son、Benjamin Black 和 Aakriti Agrawal。《重新审视多智能体深度强化学习中的参数共享》。*arXiv* 预印本 *arXiv:2005.13625*，2020。

Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N. Gomez、Łukasz Kaiser 和 Illia Polosukhin。《注意力就是你所需要的一切》。发表于第*31*届神经信息处理系统大会，2017年。

Jianhao Wang、Zhizhou Ren、Terry Liu、Yang Yu 和 Chongjie Zhang。QPLEX：双重决斗多智能体Q学习。发表于第*9*届国际学习表征会议（*ICLR*），2021年。

Yuanfei Wang、Fangwei Zhong、Jing Xu 和 Yizhou Wang。ToM2C：基于心智理论的面向目标的多智能体通信与协作。发表于第*10*届国际学习表征会议（*ICLR*），2022。

Ronald J. Williams. 面向联结主义强化学习的简单统计梯度跟踪算法。机器学习，8(3–4):229–256，1992。

Yaodong Yang、Rui Luo、Minne Li、Ming Zhou、Weinan Zhang 和 Jun Wang。均值场多智能体强化学习。载于第*35*届国际机器学习会议（*ICML*），第5571–5580页，2018年。

Rex Ying、Jiaxuan You、Christopher Morris、Xiang Ren、William L. Hamilton与Jure Leskovec。基于可微池化的层次化图表示学习。载于第*32*届神经信息处理系统国际会议（*NIPS*），第4805–4815页，2018年。Chao Yu、Akash Velu、Eugene Vinitsky、Jiaxuan Gao、Yu Wang、Alexandre Bayen与Yi Wu。近端策略优化在合作型多智能体博弈中的惊人有效性。载于 第*36*届神经信息处理系统国际会议（*NIPS*），2022年。Lianmin Zheng、Jiacheng Yang、Han Cai、Ming Zhou、Weinan Zhang、Jun Wang与Yong Yu。MAgent：面向人工集体智能的多智能体强化学习平台。载于第*32*届*AAAI*人工智能会议，2018年。

补充材料

以下内容未必经过同行评审。

A 环境详情

A.1 MAgentBattle（battle_v3）

地图尺寸与种群规模。 在 MAgent Battle中，种群规模由（正方形）地图尺寸决定。对于尺寸为 $S \times S$ 的地图，环境会生成 N 个智能体 每个团队，如表 3所示（例如， S=40 $\Rightarrow$ 64 对 64）。

S （地图尺寸）	25	40	45	50
N （智能体/团队）	20	64	81	100

表 3：Battle规模缩放：地图尺寸 S 至人口 N （每队智能体数）。

智能体动力学与交互半径。 每个智能体占据一个网格单元（1×1），生命值为 =10，生命恢复率为每步_恢复 $= 0.1$，速度为 speed $= 2$，视野半径为 view_范围 =圆形范围(3)，攻击半径为 attack_范围 = 圆形范围(1。5)，攻击伤害为 damage $= 2$。

奖励。 环境提供以下默认塑形项：步进奖励步进_奖励 = -0.005、死亡惩罚 死亡_惩罚 = -0.1、攻击惩罚攻击_惩罚 $= -0.1$，以及每次成功攻击事件的额外奖励攻击_对手_奖励 $= 0.2$。此外，消灭对手后，智能体还会获得击杀奖励 击杀_奖励 $= 5$。回合将持续到所有 蓝色 智能体被击杀，或最多运行最大_循环 $= 200$步。

受控队伍与对手策略。 在所有 Battle实验中，我们控制 红色 队伍，而 蓝色 队伍作为对手，使用预训练IDQN策略，并遵循 Li 等人 (2025)采用的相同评估协议。

生成布局。 重置时，每个团队都会在地图的相对两侧初始化为密集方阵。设 S 为地图尺寸。环境设置 init_num $= 0.04\,S^2$，并形成边长为 side $= 2\lfloor\sqrt{\text{init_num}}\rfloor$的方阵，将智能体置于隔格晶格（步长为 2）上，以避免重叠。两个方阵以中线为中心，相隔 gap $= 3$ 个单元格的水平间隙。

参考可视化。 图 4展示了一个具有代表性的 64 对 64 配置。

A.2 Pursuit（pursuit_v3）

场景规模。 在 Pursuit 中，世界是大小为 $S \times S$ 的方形网格，其中 S 取决于种群规模（表 4）。

场景	20P–8E	40P–16E	60P–24E	80P–32E	100P–40E
地图尺寸 S	40	45	50	55	60

表 4：追捕规模：追捕者/逃避者（P/E）对应的地图尺寸 $S \times S$。

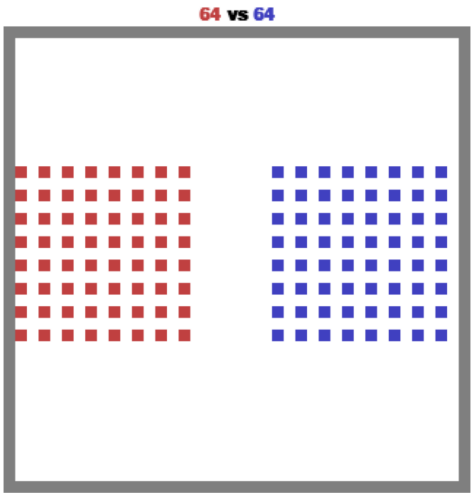

图4：对战参考地图，64对阵64（地图尺寸 40×40）。红色 方块表示受控智能体，蓝色方块表示对手智能体。

状态与观测。 环境维护一个由以下内容组成的 3 通道网格状态：(i) 静态障碍物/墙壁，(ii) 追捕者占用状态，以及 (iii) 逃逸者占用状态。每个追捕者观测以自身为中心的局部$R \times R$区块，观测_范围为 $= R$（默认 $R = 7$），并以 3 通道张量返回。追捕者在与基本网格移动和标记动作相对应的离散动作空间中行动。

捕获规则（包围）。我们始终使用包围 $=$ True。当追捕者占据位置（x，y）周围的所有开放四邻域单元格（上/下/左/右），并标记逃避者时，位于该位置的逃避者就会被移除。所需的包围单元格数量会考虑地图边界和障碍物（也就是说，若部分相邻单元格被阻挡或位于边界之外，所需数量会少于 4）。参与包围事件的每个追捕者都会被记为一名捕获者。

奖励与终止。默认奖励参数包括捕获_奖励 $= 5.0$（逃避者被移除时发放给捕获者）、标记_奖励 $= 0.0$和紧迫度_奖励 $= 0.0$。因此，这是一个非常稀疏的奖励环境。当所有逃避者都被移除，或达到时间限制 max_周期 $= 500$ 步时，回合即告终止。

受控队伍和对手策略。 在所有 PURSUIT 实验中，我们控制追捕者，而逃避者则在地图周围随机移动。

初始化。 重置时，追捕者和逃避者会在地图的一个（可能受约束的）窗口内随机初始化。在我们的运行中，我们使用约束_窗口 $= 1.0$，这对应于在整张地图范围内进行采样。

参考可视化。 图 5 展示了一个具有代表性的 20P–8E 配置。

B 超参数和架构

本节从可复现性的角度总结了 (i) 模型架构、(ii) 近端策略优化（PPO）的优化设置，以及 (iii) 通信和分组专用超参数。除非另有说明，两个基准测试均采用相同的设置。

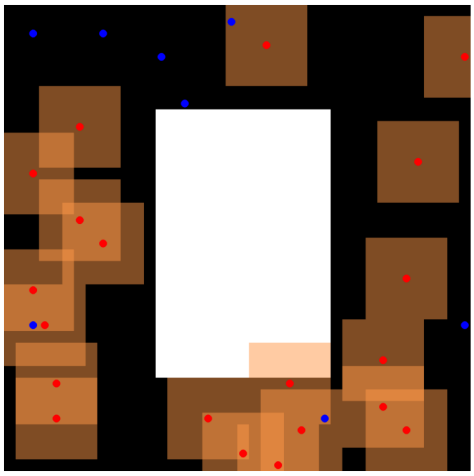

Figure 5：20P–8E 的追击参考地图（地图尺寸 40×40）。红色 圆圈表示追捕者，蓝 e circ表示逃避者。 橙色半透明方块表示追捕者的局部观测窗口 s.

B.1 模型架构

演员。 每个智能体都采用循环策略，门控循环单元的隐藏层大小为 $h = 64$。在每个环境步中，首先通过线性层和 ReLU 对智能体观测进行嵌入，然后输入门控循环单元进行处理：

$$\text{obs}_t \xrightarrow{\text{Linear+ReLU}} e_t \in \mathbb{R}^h \xrightarrow{\text{GRU}(h,h)} u_t \in \mathbb{R}^h.$$

根据 u_t，演员输出三个分支：(i) 环境动作对数几率，(ii) 二元“发送”决策对数几率，以及 (iii) 针对 N 候选接收者的接收者选择对数几率。

描述符网络。 描述符网络将每个智能体的输入编码为：(i) 用于软分组分配的分组描述符 $z_{\text{grp}} \in \mathbb{R}^{64}$，以及 (ii) 用作通信内容表示和通信评论器输入的消息特征 $z_{\text{msg}} \in \mathbb{R}^{96}$。描述符多层感知机在输出端应用层归一化，观测则使用运行归一化。

邮箱聚合（无参数注意力）。 传入消息通过消息特征空间中的单头点积注意力进行聚合：对于每个接收方，注意力权重是在发送方之间进行 softmax 计算得到的（缩放因子为 $1/\sqrt{d}$），邮箱则是发送方消息向量的加权和。所有智能体均可作为接收方。

Group感知评论器。 分组评论家使用一个2层M将全局状态映射到分组价值
线性规划（ReL
U，无层归一化），

$$V_{\text{grp}}: \quad s \to \mathbb{R}^h \to \mathbb{R}^h \to \mathbb{R}^M,$$

并通过（软）群组分配矩阵 P_τ生成各智能体的值。

通信评论器。 我们使用两个小型多层感知机（ReLU，无层归一化）：（i）价值网络 $V(\cdot)$，其输入为全局状态与智能体特征向量的拼接；（ii）边价值网络 $Q(i \to j)$，其输入为全局状态与发送方和接收方特征的拼接。智能体特征包括z_{msg}、输入邮箱嵌入、发送位以及 M维软群组分配。

B.2 优化与训练

近端策略优化。 我们使用近端策略优化，每次更新的采样轨迹长度为2048个环境步数，更新轮数为8，迷你批次大小为4096，裁剪参数为 $\varepsilon = 0.2$。折扣因子使用 $\gamma = 0.99$，广义优势估计使用

$\lambda = 0.95$。价值损失系数为0.5，最大梯度范数为1.0。我们使用 Adam，学习率为 3×10^{-4}。熵正则化系数分别为：环境动作 0.02、发送决策 0.01，以及接收方选择 0.01。

分组目标与调度。 软分组分配采用 Gumbel–Softmax 温度调度，在训练过程中从 $\tau = 10$ 线性退火至 $\tau = 0.5$ 。分组分配基于原型，使用与 M 个可学习原型的余弦相似度，Logit 缩放为 3.0。辅助分组损失使用：平衡权重 0.1、行熵权重 0.01，以及边效用权重在训练过程中从 0 线性调度至 1。

宏步。 分组分配及相应的分组诱导的接收方先验每 $K = 10$ 个环境步重新计算。

为方便起见，表 5 汇总了上文所述的共享架构与训练超参数。

类别	**设置**
网络	
演员隐藏层大小 h	64（门控循环单元）
描述符维度	z_{grp} : 64, z_{msg} : 96
分组评论家	多层感知机: $s \to 64 \to 64 \to M$ (ReLU)
通信评论器	多层感知机：$(s\|\|\cdot) \to 64 \to 64 \to 1$(ReLU)
归一化	运行观测归一化；描述符输出处采用层归一化
邮箱聚合	单头点积注意力，规模 $1/\sqrt{d}$
PPO	
采样轨迹长度	2048
小批量大小	4096
每次更新的轮数	8
裁剪 ε	0.2
γ，广义优势估计 λ	0.99，0.95
价值损失系数	0.5
熵系数（动作 / 发送 / 接收）	0.02 / 0.01 / 0.01
优化器	Adam，学习率 3×10^{-4}
最大梯度范数	1.0
分组／通信	
宏步 K	10
Gumbel–Softmax τ	线性 $10 \to 0.5$
Logit 缩放	3.0
辅助损失（平衡／行熵／边效用）	0.1 / 0.01 / 线性 $0 \to 1$
邮箱延迟	一步（$t \to t{+}1$）

表 5：Battle 和 Pursuit 中使用的共享超参数。

B.3 基准测试特定设置

分组数量 M。 我们使用简单的、依赖于规模的启发式方法来设置群组数量 ：

$$\textbf{Battle: } M = \left\lfloor \sqrt{N} \right\rfloor, \qquad \textbf{Pursuit: } M = \left\lfloor \frac{E}{2} \right\rfloor,$$

其中 N 表示 Battle 中每个团队的智能体数量 ，而 E 表示 Pursuit 中逃避者的数量。表6和7列出了 M 主论文中报告的种群规模。附录D中的敏感性研究进一步论证了选择这些依赖规模的启发式方法的合理性。

场景	# 组 M
20对20	4
64对64	8
81对81	9
100对100	10

表6：Battle：各规模对应的组数。

场景	# 组数 M
20P-8E	4
40P-16E	8
60P-24E	12
80P-32E	16
100P-40E	20

表7：追逐任务：各规模对应的组数。

C 补充结果

我们展示了 SCOUT 在每个基准测试的代表性规模下，单个示例性评估回合中的展开过程快照：**100对100战斗**（地图尺寸为 50，使用预训练对手）和**100P–40E追逐**。对于每个回合，我们报告进程 步数 $t \in \{0, \lfloor T/4 \rfloor, \lfloor T/2 \rfloor, T\}$处的四个帧，其中 T是所示回合的最终步数。我们还报告了同一设置下事后计算的组内消息比例，以支持以下观点：习得交互结构将通信集中在组内。

C.1 Battle（100对100）

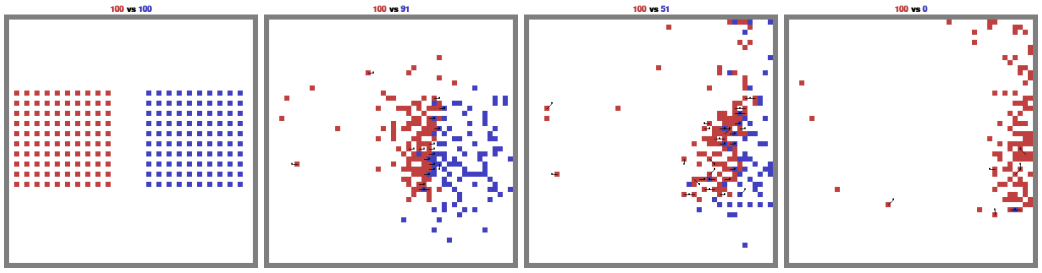

图6：Battle 100对100：SCoUT 在进度 $t = 0$，$\lfloor T/4 \rfloor$，$\lfloor T/2 \rfloor$，T时的展开过程（本回合：0，15，30，60，其中 $T = 60$）。红色智能体由SCoUT控制；蓝色智能体是预训练的IDQN对手。

关键事件（图 6）。在 $t = 0$，两支团队从对称的出生区域开始。随着双方接触，SCoUT 的接收方选择变得有章可循：消息优先在同一（软）群组内传递，这通常对应于彼此邻近的智能体以及局部协同攻击。到回合中期，受控的红色智能体形成多个连贯的局部交战，而非单一、分散的战线。到最后一帧，剩余的红色智能体往往呈现为空间上连贯的集群，这与局部化的群组驱动的协同相符。

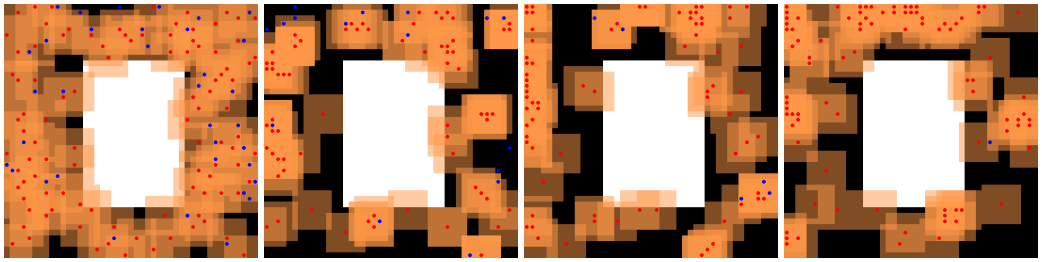

图7：追捕 `100P-40E`：SCoUT 在进度 $t = 0$，$\lfloor T/4 \rfloor$，$\lfloor T/2 \rfloor$，T（本回合：0、35、70、105，包含 $T = 135$）。红色圆圈表示追捕者；蓝色圆圈表示逃避者。白色单元格是障碍物/墙壁；黑色单元格是自由空间。

指标	20对20	64对64	81对81	100对100	20P-8E	40P-16E	60P-24E	80P-32E	100P-40E
战斗					Pursuit				
组内比例	0.343±0.046	0.257±0.009	0.488±0.032	0.406±0.010	0.297±0.024	0.319±0.010	0.376±0.056	0.238±0.029	0.397±0.003
随机基线（$1/M$）	0.250	0.125	0.111	0.100	0.250	0.125	0.083	0.062	0.050
群组数（M）	4	8	9	10	4	8	12	16	20

表8：**不同规模下的组内消息比例（10 个评估随机种子）**。我们报告基于实际产生的发送者—接收者消息边计算得出的组内消息比例（上文公式）的均值 ± 标准差。随机基线对应于均匀随机组成员关系，其组内消息比例的期望值为 $1/M$。

C.2 追捕 (100P–40E)

关键事件（图7）。最初，追捕者分散在地图各处，并开始围绕附近的逃避者形成局部联盟。随着回合推进，子团队逐渐出现，并在地图的不同区域反复执行包围捕获行为（表现为成群的追捕者逐渐围紧逃避者）。这一行为与组诱导的接收者偏置一致：通信强调已作为局部捕获单元行动的智能体之间的协调，从而减少不必要的跨地图广播。

C.3 分组效应：组内消息比例

为量化通信是否与习得交互结构相一致，我们计算组内消息比例：即实际发送者—接收者消息边（$i \to j$）中，发送方与接收方属于同一群组的比例。具体而言，在每个宏步，我们得到软分配 $P_\tau \in [0,1]^{N \times M}$，并定义硬组标签$\hat{g}(i) = \arg\max_m P_\tau(i,m)$。令 $\mathcal{E}$表示一个回合中所有实际发送者—接收者消息边的集合。我们报告

$$\text{InGroupFrac} = \frac{1}{|\mathcal{E}|} \sum_{(i \to j) \in \mathcal{E}} \mathbb{I}[\hat{g}(i) = \hat{g}(j)] .$$

我们对这一指标在 10 个评估随机种子上取平均。该诊断指标仅用于分析，本身并未作为训练目标进行优化。表8 报告了所有规模下的群组内占比，并给出了随机基线 $1/M$。

在所有报告的规模下，组内消息比例都远高于 $1/M$ 基线，表明接收方选择与习得的分组结构高度一致。这与图 6 和图 7 中的定性策略展开相符，也印证了分组在 SCOUT 中的作用：它在训练期间形成一种缓慢变化的交互结构，从而减少弥散式广播并促进局部协调。

D 群组数量与宏步长度的敏感性

SCOUT 引入了两个面向环境的超参数，用于控制习得交互结构的粒度和时间持续性：组数 M 和宏步长度 K。本节评估性能对 (M, K) 的敏感程度，并考察主实验中采用的 $K{=}10$ 和依赖规模的 M 是否处于稳定的高性能区域。

设置。 我们在两个具有代表性且计算开销可控的规模上进行受控扫描：PURSUIT 40P-16E 和 BATTLE 64v64。这些设置规模足够大，能够呈现非平凡的协调动态，同时计算开销又远低于最大的压力测试案例。即使在这些规模下，每次 $4{\times}4$ 扫描也需要**4 天** 来完成每个基准测试，因此在更大规模上进行类似扫描的成本高得难以承受。

D.1 实验设置

在 (M, K) 上进行网格搜索。 对于每个基准测试，我们评估一个包含 4×4 个网格点的网格：

$$M \in \{4, 6, 8, 12\}, \qquad K \in \{5, 10, 15, 20\}.$$

该网格包含主论文中使用的默认配置 $(M{=}8, K{=}10)$，对应于 64对64 Battle ($M = \lfloor\sqrt{64}\rfloor = 8$) 和 40P-16E Pursuit ($M = \lfloor 16/2 \rfloor = 8$) 下的尺度规则。

训练协议与指标。 除 (M, K) (附录B) 外，所有运行均使用相同的超参数，并训练 2000 次迭代。我们报告每种配置的最终平均回合回报。图 8 和 9 展示了由此得到的性能曲面。

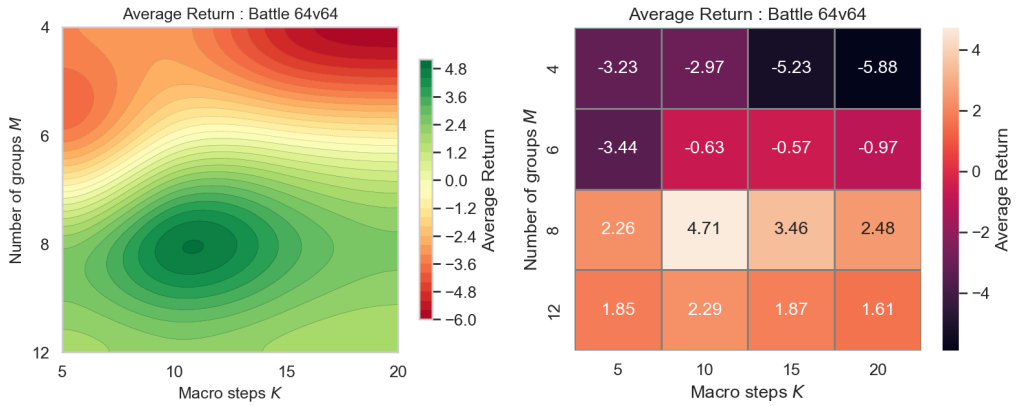

图 8：SCOUT 在 BATTLE 64v64 中对 (M, K) 的敏感性。**左侧：** 插值等高线图。**右侧：** 每个已评估的 (M, K) 组合的精确回报热力图。

D.2 结果与解读

Battle 64对64。 图 8 显示其表现强烈依赖于 M。极小的 M 在各个K下表现都很差，其中 $M{=}4$ 在所有 K 下都带来最低回报。相比之下，中等数量的群组表现明显更好：最佳配置为 $(M{=}8,\ K{=}10)$，相近的设置也保持了竞争力。对于较大的 M，平均奖励仍为正，但通常低于最佳区间。这一规律符合直觉：Battle 需要有意义的分组结构，同时也要保留足够的组内规模，以协调局部交战。

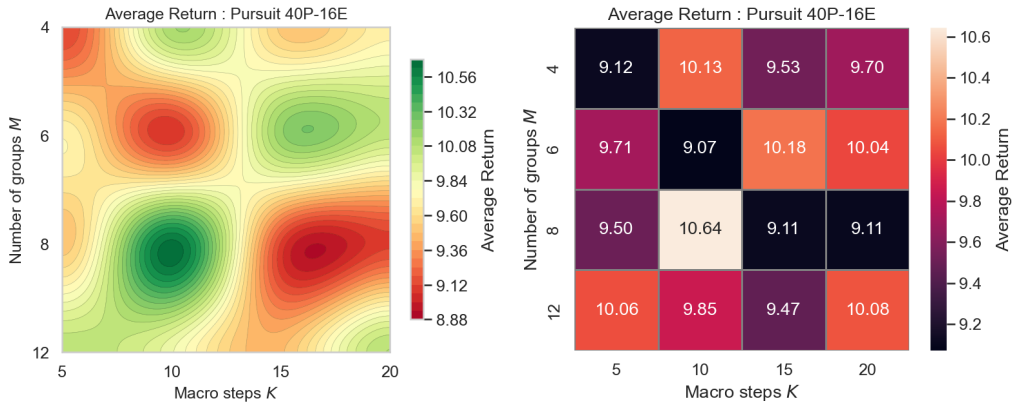

图 9：SCOUT 在 PURSUIT 40P-16E 上对 $(M,\ K)$ 的敏感性。**左图：**粗略趋势的插值等高线图（越高越好）。**右图：**热力图，展示每个经过评估的 $(M,\ K)$ 配对对应的精确回报。

Pursuit `40P-16E`。 图 9 显示，表现最佳的配置为（M=8，K=10）。此外，若干设置也取得了有竞争力的回报：（M=6，K=15）达到 10。18，（M=4，K=10）达到 10。13，以及（M=12，K=20）达到 10。08）。总体而言，这组参数扫描结果支持默认设置（8，10），同时表明，与 BATTLE 相比，SCOUT 在 PURSUIT 中对 $(M,\ K)$ 的选择不那么敏感；而 BATTLE 对 M 的依赖则更有规律且更显著。

要点。在两个基准测试中，参数扫描结果均表明，性能最佳区域位于主实验所用的默认选项 (M=8，K=10) 附近或与之重合。观察到的趋势与这两个超参数的作用相符：M 控制交互粒度——过小会降低选择性，并导致类广播通信；过大则可能使联盟碎片化。另一方面，K 控制持续性——过小可能引入不必要的变动；过大则可能使交互结构在动态场景中过时。这些结果进一步证明，在主论文中采用 K=10以及依赖规模的 M，且无需逐尺度调优，是合理的。