arXiv:2305.05911v1 [cs.MA] 10 May 2023

突发策略变化下的快速队友适应

预印本

Ziqian Zhang[1, *], **Lei Yuan**[1, 2, *], **Lihe Li**[1], **KeXue**[1], **Chengxing Jia**[1,2], **Cong Guan**[1], **Chao Qian**[1], **Yang Yu**[1, 2, †]
[1] 新型软件技术国家重点实验室，南京大学
[2] Polixir.ai
{zhangzq,yuanl,lilh,xuek,jiacx,guanc}@lamda.nju.edu.cn, {qianc,yuy}@nju.edu.cn

2023年5月11日

摘要

在协作式多智能体强化学习（MARL）中，智能体为了实现共同目标而与队友协同合作时，可能会受到队友策略变化所引发的非平稳性影响。以往研究主要关注训练阶段中的策略变化，或队友在不同回合之间改变策略的情况，却忽略了队友可能在单个回合内突然发生策略变化这一事实，而这可能导致协调失误，进而造成性能下降。我们将这一问题建模为开放式分散式部分可观测马尔可夫决策过程（Dec-POMDP），其中，我们控制部分智能体，使其与不受控制的队友进行协作，而这些队友的策略可能在单个回合内发生变化。随后，我们提出一种新的框架 ***Fast teammates adaptation (Fastap)***，以解决这一问题。具体而言，我们首先训练具有多样性的队友策略，并通过中国餐馆过程（CRP）将其分配到不同的簇中。随后，我们训练受控智能体，使其通过捕捉被采样的非受控队友的身份特征，并将其作为上下文，实现快速适应。最后，每个智能体利用自身的局部信息预测队友的上下文，并据此进行决策。这一过程交替进行，最终形成一种鲁棒策略，使智能体能够在分散式执行阶段适应任意队友。我们在多个多智能体基准测试中证明，与多种基线方法相比，Fastap 在平稳和非平稳场景下均能取得更优的性能。

1 引言

近年来，协作式多智能体强化学习（多智能体强化学习）展现出巨大潜力，多个智能体协同完成具有共同目标 [Oroojlooy 和 Hajinezhad， 2022], 并在诸多领域取得显著进展（例如，寻路 [Sartoretti 等人， 2019], 主动电压控制 [Wang 等人， 2021], 以及动态算法配置 [Xue 等人， 2022]）。各种方法纷纷涌现，成为颇具前景的解决方案，包括基于策略的方法 [Lowe 等，2017，Yu 等人， 2022], 基于价值的方法系列 [Sunehag 等人，2018，Rashid 等人， 2018], 以及变换器等多种变体 [Wen 等人， 2022], 它们在星际争霸多智能体挑战（SMAC）、Google Research Football（GRF）等各类任务中展现出卓越的协作能力 [Gorsane 等人， 2022], 。其他研究则探讨了不同方面，包括智能体间的通信 [Zhu 等人， 2022], 模型学习 [Wang 等人，2022], 策略鲁棒性 [Guo 等人， 2022],临时团队协作 [Mirsky 等人， 2022], 等。

然而，多智能体强化学习（MARL）中可能出现的一个问题是非平稳性 [Papoudakis 等人， 2019] 这种情况是由队友策略的变化引起的。无论是在单智能体强化学习（SARL） [Padakandla 等人， 2019], 还是在多智能体强化学习 [Papoudakis 等人，2019] 中，非平稳性都是强化学习面临的一个危险问题：学习系统的环境动态（例如，转移函数或奖励函数）可能会随时间发生变化（跨回合或回合内）。针对这一问题，单智能体强化学习领域已提出许多解决方案，包括元强化学习 [Beck 等人，2023], 策略性撤退 [Dastider 和 Lin， 2022], 粘性层次狄利克雷过程（HDP）先验 [Ren 等人，

[*] 前两位作者贡献相同。
[†] 通讯作者

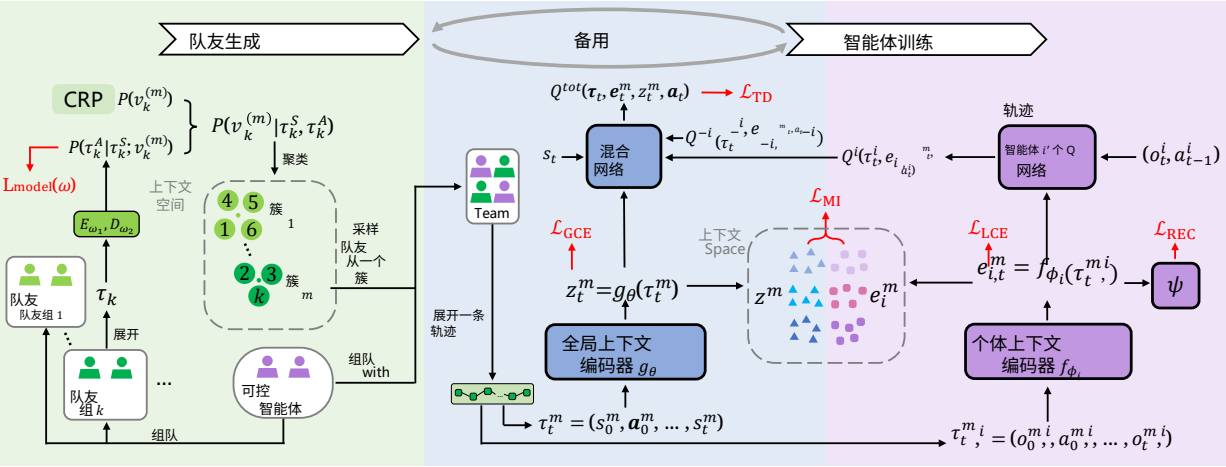


图 1：Fastap 的整体框架。

2022], 等等。然而，多智能体强化学习中的非平稳性要复杂得多，因为我们需要考虑多个队友引起的策略变化，而不是单智能体强化学习中单一的环境动态变化。多智能体强化学习领域的大多数研究主要关注训练阶段的非平稳性 [Albrecht 和 Stone，2018，Kim 等人， 2021],队友策略在不同回合之间的变化 [Qin 等人，2022，Hu 等人， 2020], 或发生扰动时的情况 [Guo 等人， 2022] （相关工作见附录 A）。然而，据我们所知，在部署过程中，队友在一个回合内突然改变策略这一问题尚未得到探索，无论是在问题表述方面，还是在高效算法设计方面。忽视这一问题会导致策略偏移，甚至造成灾难性协调失灵，因为在多智能体强化学习中，智能体的策略取决于其他队友 Zhang 等人 [2021]。另一方面，由于多智能体强化学习固有的特点（例如部分可观测性），单智能体强化学习中成功的方法并不适用于多智能体强化学习场景。这引出了一个问题：我们能否获得一种鲁棒策略，既能应对这种变化，又能迅速适应新的队友策略?

在本研究中，我们旨在针对上述问题开发一种鲁棒的协调策略。具体而言，我们将该问题表述为一个开放式分散式部分可观测马尔可夫决策过程，其中我们控制多个智能体与一些不受控制的队友进行协调，而这些队友的策略可能在一个回合内发生不可预测的变化。随后，我们开发了一种新的训练框架 Fastap，使智能体能够通过局部信息推断队友的身份。具体来说，由于相似的队友可能具有相似的身份特征，如果为每个队友学习特定的上下文，却忽略他们之间的关系，可能会导致编码过于简单。因此，我们通过中国餐馆过程（CRP）将他们分配到不同的簇中，以缩小上下文搜索空间。在训练受控协调策略时，我们选取具有代表性的队友进行协调，并将他们的身份特征编码为具有区分性的上下文，以在集中式训练阶段增强联合策略。随后，每个智能体利用其局部信息来近似全局上下文信息。上述过程交替进行，最终我们可以获得一种鲁棒策略，使其在分散式执行阶段逐步适应任意队友。

为了进行评估，我们在不同的多智能体强化学习基准测试上开展实验，其中队友的策略会在一个回合内发生变化，包括基于等级的觅食（LBF） [Papoudakis 等，2021b], 捕食者-猎物（PP）、来自多智能体粒子环境（MPE）的协作导航（CN） [Lowe 等人， 2017], 以及根据星际争霸多智能体挑战（SMAC） [Samvelyan 等人， 2019] 创建的地图。实验结果表明，所提出的 Fastap 能够将队友聚类为不同的群组，学习有意义的上下文以识别队友，并且与多种基线方法相比，在平稳和非平稳场景中均取得了出色的性能。

2 问题表述

本研究旨在训练多个可控智能体，使其能够与其他队友交互，而这些队友可能会在一个回合内的任意时间步突然改变其策略。因此，我们通过将 Dec-POMDP [Oliehoek 和 Amato, 2016] 的框架扩展为开放式分散式部分可观测马尔可夫决策过程 $\mathcal{M} = \langle \mathcal{N}, \bar{\mathcal{N}}, \mathcal{S}, \mathcal{A}, \bar{\mathcal{A}}, P, \Omega\ O,\ R,\ \mathcal{U},\ \gamma \rangle$，来形式化这一问题。这里， $\mathcal{N} = \{1, ..., n\}$和 $\bar{\mathcal{N}} = \{\bar{1}, ..., \bar{m}\}$ 分别是可控智能体和不可控队友的集合， $\mathcal{S}$

表示状态集合，$\mathcal{A}=\mathcal{A}^1\times...\times\mathcal{A}^n$ $\mathcal{N}$ 和 $\bar{\mathcal{A}}=\mathcal{A}^{\bar{1}}\times...\times\mathcal{A}^{\bar{m}}$ 分别表示相应的联合动作集合，$\mathcal{N}$、P、O和 R 分别表示相应的状态转移函数、观测函数和奖励函数，Ω表示观测集合，$\gamma\in[0,1)$ 表示折扣因子，而 $\mathcal{U}$ 表示用于控制突发变化发生频率的概率分布。

在每个回合开始时，将一开始参与协作的不可控队友集合记为 $\bar{\mathcal{N}}_0\in\mathcal{P}(\bar{\mathcal{N}})$，其中 $\mathcal{P}(\cdot)$ 表示幂集，等待时间则表示为 $u_0\sim\mathcal{U}$。在每个时间步，t、$u_t=u_{t-1}-1$ 和 $\bar{\mathcal{N}}_t=\bar{\mathcal{N}}_{t-1}$ 都会更新。如果 $u_t\leq 0$，则会从 $\mathcal{U}$中重新采样，并由一组全新的不可控队友 $\mathcal{N}_t\in\mathcal{P}(\mathcal{N})$ 取代之前的队友。同时，可控智能体 i 接收观测 $o^i=O(s,i)$ 并输出动作 $a^i\in\mathcal{A}^i$，不可控队友也执行相同操作。请注意，在一个回合中，不可控队友的数量可能会变化。联合动作 $(\boldsymbol{a},\bar{\boldsymbol{a}})$ 会导向下一个状态 $s'\sim P(\cdot|s,(\boldsymbol{a},\bar{\boldsymbol{a}}))$，并产生共享奖励 $R(s,(\boldsymbol{a},\bar{\boldsymbol{a}}))$，其中 $\boldsymbol{a}=(a^1,...,a^n)\in\mathcal{A}$ 和 $\bar{\boldsymbol{a}}\in\{(a^{\bar{i}})_{\bar{i}\in\bar{N}}|a^{\bar{i}}\in\mathcal{A}^{\bar{i}},\bar{N}\in\mathcal{P}(\bar{\mathcal{N}})\}$。为缓解部分可观测性，智能体 i 截至时间步t的轨迹历史 $(o_1^i,a_1^i,...o_{t-1}^i,a_{t-1}^i,o_t^i)$ 将通过门控循环单元（GRU） [编码为 τ_t^i（Cho 等， 2014]）。在开放式分散式部分可观测马尔可夫决策过程下，我们旨在求得一种最优策略，以应对不可控队友遭遇突发变化的情况。随后，$\boldsymbol{a}$给定 $\boldsymbol{\tau}_t=\langle\tau_t^1,...,\tau_t^n\rangle$，形式化目标是为可控智能体寻找联合策略 $\boldsymbol{\tau}\ \boldsymbol{a}\ =\mathbb{E}_{s\ \boldsymbol{a}\ \bar{\boldsymbol{a}}}[\sum_{t=0}^{\infty}\gamma^t R\ s\ \ \boldsymbol{a}\ \bar{\boldsymbol{a}}\ \ |s_0=s\ \boldsymbol{a}_0=\boldsymbol{a}\ \boldsymbol{\pi}\ \bar{\boldsymbol{\pi}}]$ $\bar{\boldsymbol{\pi}}$ (,)，使全局价值函数 $Q_{\text{tot}}^{\boldsymbol{\pi}}(,),,(,(,)),,,,$ 最大化，其中 表示不可控队友的未知联合策略。

3 方法

本节将详细介绍 Fastap（见图 1）的设计。这是一种新颖的多智能体策略学习方法，能够使可控智能体应对队友策略的突发变化，并迅速适应新队友。首先，我们基于中国餐馆过程（CRP） [Blei 和 Frazier， 2010] 设计一种无限混合模型，用于描述队友簇持续增加的分布（第 3.1 节和图 1(a)）。接下来，我们介绍用于快速适应的集中式上下文编码器学习目标（第 3.2 节和图 1(b)）。最后，考虑到合作式多智能体强化学习中广泛采用的集中式训练、分散式执行范式（CTDE），我们训练每个可控智能体根据其局部信息迅速识别并适应队友情形（第 3.3 节和图 1(c)）。

3.1 基于CRP的无限混合模型，用于动态队友生成

为了在评估期间迅速适应单个回合中行为各异的队友突发变化，我们希望维护一组多样化的策略，以模拟训练阶段可能遇到的队友。然而，将每个新生成的队友组都视为一种新类型，同时忽略它们之间的相似性，既不合理也缺乏效率。在队友增量生成的学习过程中，这种做法缺乏可扩展性；如果生成行为相似的队友，也可能降低训练效果。因此，我们希望通过应用行为检测模块，将行为相似的队友组分配到同一簇，从而明确区分队友的行为。为解决这一问题，可以采用无限狄利克雷过程混合模型（DPM） [Lee 等人， 2020] 因为该模型在簇的数量方面具有可扩展性和灵活性。具体而言，我们可以将队友生成过程表述为由不同轨迹批次构成的流 $\mathcal{D}_1$，$\mathcal{D}_2$.....，其中每个批次 $\mathcal{D}_k$ 是一组轨迹 $\tau=(s_0,\ \boldsymbol{a}_0...,\ s_T)$，这些轨迹是从 k^{th} 队友组与环境的交互中采样得到的，T为时域长度。考虑到轨迹维度较高、表征困难，我们使用由 ω_1 参数化的轨迹编码器 E_{ω_1}将 τ编码到潜在空间。具体而言，我们将轨迹 τ划分为 $\tau^S=(s_0,\ ...s_{T-1},\ s_T)$ 和 $\tau^A=(\boldsymbol{a}_0,\ .....,\ \boldsymbol{a}_{T-1})$，并应用 Transformer 架构从轨迹中提取特征，将其表示为 $v=E_{\omega_1}(\tau)$。对于迄今生成的第 k^{th} 个队友组，$v_k=\mathbb{E}_{\tau_k\sim\mathcal{D}_k}[E_{\omega_1}(\tau_k)]$ 将用于表示其行为类型，而 $\bar{v}^m$ 是该 m^{th} 簇的均值。

如果目前已实例化 M 个簇，则根据分配 $P(v_k^{(m)}|\tau_k)=P(v_k^{(m)}|\tau_k^S,\ \tau_k^A)$ $m=1,\ ...,\ M,\ M+1$ 推断 k^{th} 队友组所属的簇，其中 $v_k^{(m)}$ 表示根据其表征 v_k，该 k^{th} 组属于 m^{th} 簇。后验分布可写为：

$$P(v_k^{(m)}|\tau_k^S,\tau_k^A)\propto P(v_k^{(m)})P(\tau_k^A|\tau_k^S;v_k^{(m)}),\tag{1}$$

我们采用 CRP [Blei 和 Frazier， 2010]将 DPM 模型实例化为先验。具体而言，对于一系列表征为 $[v_1,v_2,...v_k,...]$ 的队友群组，其先验 $P(v_k^{(m)})$ 设置为：

$$P(v_k^{(m)})=\begin{cases}\frac{n^{(m)}}{k-1+\alpha}, & m\leq M\\ \frac{\alpha}{k-1+\alpha}, & m=M+1,\end{cases}\tag{2}$$

其中， $n^{(m)}$ 表示属于 m^{th} 簇的队友组数量， M 表示迄今为止创建的簇数量， $\sum_{m=1}^{M} n^{(m)} = k-1$，而 $\alpha > 0$ 是一个浓度超参数，用于控制创建新簇的概率。

为估计预测似然 $P(\tau_k^A|\tau_k^S; v_k^{(m)})$，我们使用基于RNN的解码器 D_{ω_2}，其以 τ_k^S， $v_k^{(m)}$ 作为输入并预测 τ_k^A。该解码器将每个样本表示为高斯分布 $\mathcal{N}(\mu(\tau_t^S, v)\ \sigma^2(\tau_t^S, ..., v))$，其中$\tau_t^S = (s_0, ..., s_t)$，使得

$$\begin{aligned} P(\tau_k^A|\tau_k^S; v_k^{(m)}) =& D_{\omega_2}(\tau_k^A|\tau_k^S; v_k^{(m)}) \\ =& \prod_{t=1}^{T} D_{\omega_2}(\boldsymbol{a}_t^k|\tau_{k,t}^S, v_k^{(m)}), \\ \text{where } v_k^{(m)} =& \begin{cases} \frac{n^{(m)}\bar{v}^m + v_k}{n^{(m)}+1} & m \le M \\ v_k & m = M+1. \end{cases} \end{aligned} \tag{3}$$

结合估计得到的先验公式（2）和预测似然公式（3），我们能够确定该 k^{th}队友组所属的簇，从而获得队友行为清晰可区分的边界。完成分配后， m^{th} 簇的均值也将随之更新。同时，为了促使学习到的表征 v捕捉每个队友组的行为信息，并更精确地估计预测似然，编码器 E_{ω_1}和解码器 D_{ω_2} 被优化为：

$$\mathcal{L}_{\text{model}}(\boldsymbol{\omega}) = -\log \mathbb{E}_{\tau \sim \cup_{k=1}^{K} \mathcal{D}_k}[D_{\omega_2}(\tau^A|\tau^S; E_{\omega_1}(\tau))], \tag{4}$$

其中， K 表示截至目前生成的队友组数量， $\boldsymbol{\omega} = (\omega_1, \omega_2)$。在生成队友组的同时，编码器和解码器会得到优化（详见附录 B.1）。

3.2 面向快速适应的集中式上下文学习

在获得生成的队友并将其划分为不同簇之后，本部分旨在训练一种鲁棒策略，以应对队友的突发变化，并通过根据其他队友的行为来调整可控智能体的策略，使其快速适应新队友。尽管未知队友的行为表现出多样性和复杂性，但前文形式化定义的中国餐馆过程（CRP）有助于根据队友的行为类型，从高层语义的角度获得清晰可辨的边界。受环境敏感上下文策略学习（ESCP） [Luo 等人 2022],的启发，该方法旨在引导上下文编码器快速识别并跟踪环境的突发变化；我们希望利用全局上下文编码器 g_θ 和局部上下文编码器 $\{f_{\phi_i}\}_{i=1}^{n}$ ，将历史交互嵌入紧凑且信息丰富的表示空间中。这些编码器应能快速识别新类型的队友，从而及时察觉突发变化；我们还可以提出一个目标函数来优化编码器，使其输出在早期迅速覆盖预言机，并在剩余步骤中保持一致。

在集中式训练阶段，我们设定 $z_t^m = g_\theta(\tau_t^m)$，其中 $\tau_t^m = (s_0^m, \boldsymbol{a}_0^m, ..., s_t^m)$ 是根据配对联合策略 $(\boldsymbol{\pi}, \bar{\boldsymbol{\pi}}^m)$ 与环境之间的交互生成的，而 $\bar{\boldsymbol{\pi}}^m$ 是属于 m^{th} 簇的不可控队友的联合策略。请注意，队友簇在每个回合开始时选定，并且在训练过程中不会改变；队友的突发变化只会发生在评估期间。我们可以得到 g_θ 的经验优化目标，如下：

$$\mathcal{L}_{\text{GCE}} = \sum_{m=1}^{M} \mathbb{E}[||z_t^m - \bar{z}^m||_2^2] - \log\det(R_{\{\bar{z}^m\}}), \tag{5}$$

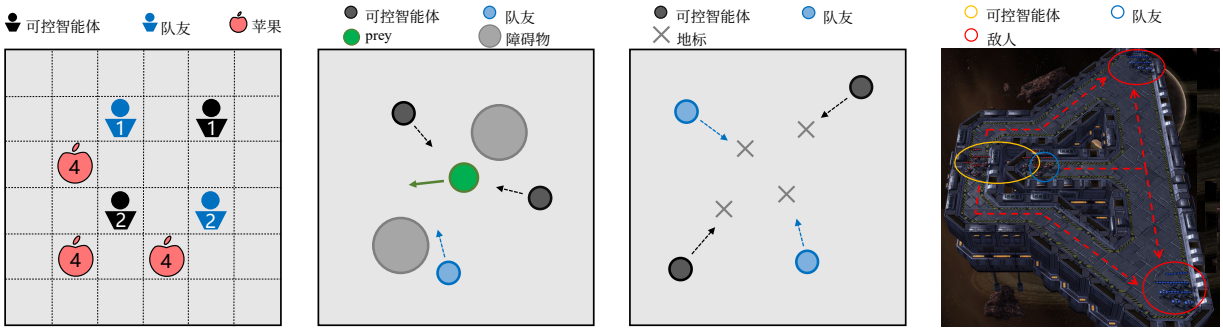
其中， $\bar{z}^m$ 是用于稳定训练过程的所有历史上下文向量的移动平均值， θ 是全局上下文编码器 g_θ 的参数，$\det(\cdot)$ 表示矩阵行列式，而 $R_{\{\bar{z}^m\}}$ 是一个关系矩阵。直观而言，该目标旨在帮助编码器的输出在早期快速覆盖，并在后续步骤中保持一致。具体来说，前一部分促使 z_t^m 在一个回合内快速、稳定地收敛；后一部分则促使 z_t^m 的期望趋向一组彼此可分但具有代表性的潜在向量。完整推导见附录 B.2。

在实践中，采用循环神经网络来实现 g_θ，该网络以$\tau_t^m = (s_0^m, \boldsymbol{a}_0^m, ... s_t^m)$作为输入，并输出多元高斯分布 $\mathcal{N}(\mu_\theta(\tau_t^m)\ \sigma_\theta^2\tau_t^m)$。因此，通过$z_t^m \sim g_\theta(\tau_t^m)$使用重参数化技巧，即可从高斯分布中获得队友上下文。由于我们可以将 Fastap 应用于任何基于价值的方法，因此也可以将全局嵌入z_t^m整合到集中式网络中。类似地，局部嵌入 $e^{m,\ t\ \ i}$和局部轨迹 $\tau^{m,\ t\ \ i}$也将被拼接起来，以计算局部 Q 值 $Q^i(\tau^{m,\ t\ \ i}, e^{m,\ t\ \ i}, \cdot)$，

其中，局部上下文编码器的优化将在下一部分中详细说明。因此，时序差分损失$\mathcal{L}_{\text{TD}} = [r_t^m + \gamma \max_{\boldsymbol{a}_{t+1}^m} Q_{\text{tot}}(s_{t+1}^m, \boldsymbol{e}_{t+1}^m, z_{t+1}^m, \boldsymbol{a}_{t+1}^m) - Q_{\text{tot}}(s_t^m, \boldsymbol{e}_t^m, z_t^m, \boldsymbol{a}_t^m)]$被用于加速集中式上下文学习，其中$\bar{Q}_{\text{tot}}$是周期性更新的目标Q网络，且$\boldsymbol{e}_t^m = (e^{m,1}_t)_{i=1}^n$。因此，$g_\theta$的总体优化目标可以推导为：

$$\mathcal{L}_{\text{ADAP}} = \mathcal{L}_{\text{TD}} + \alpha_{\text{GCE}} \mathcal{L}_{\text{GCE}}, \tag{6}$$

其中α_{GCE}是一个可调超参数，用于平衡两个优化目标。



(a) LBF (b) 捕食 (c) 协作 (d) 10m_对比_14m

Figu图 2：本文使用的实验环境

。

3.3 分散式团队情境识别与优化

尽管优化式（6）有助于获得紧凑且具有代表性的表征z_t^m，从而引导个体策略快速适应队友的突发变化，但多智能体强化学习的部分可观测性使得以分散方式执行的智能体无法获取由全局状态-动作轨迹编码得到的z_t^m。因此，我们为每个智能体i配备一个局部编码器f_{ϕ_i}，以识别团队情境。具体而言，f_{ϕ_i}的网络架构与g_θ类似，f_{ϕ_i}以局部轨迹$\tau^m_{t}{}^i = (o^m_0{}^i, a^m_0{}^i, \ldots, o^m_t{}^i)$作为输入，并输出$e^m_t{}^i \sim \mathcal{N}(\mu_{\phi_i}(\tau^m_t{}^i), \sigma^2_{\phi_i}(\tau^m_t{}^i))$。为了使$e^m_t{}^i$与$z_t^m$在信息上保持一致，我们引入互信息（MI）目标，通过最大化$e^m_t{}^i$与z_t^m之间的互信息$\mathcal{I}(e^m_t{}^i; z_t^m|\tau^m_t{}^i)$，其中该互信息以智能体$i$的局部轨迹$\tau^m_t{}^i$为条件。鉴于直接估计条件分布存在困难且可行性有限，我们使用变分分布$q_\xi(e^m_t{}^i|z_t^m, \tau^m_t{}^i)$来近似条件分布$p(e^m_t{}^i|z_t^m, \tau^m_t{}^i)$。受信息瓶颈 [（Alemi 等人， 2017],）的启发，我们将推导出互信息目标的一个可处理下界：

$$\begin{aligned}&\mathcal{I}(e_t^{m,i}; z_t^m|\tau_t^{m,i}) \geq \\ &\quad \mathbb{E}_{\mathcal{D}}[\log q_\xi(e_t^{m,i}|z_t^m, \tau_t^{m,i})] + \mathcal{H}(e_t^{m,i}|\tau_t^{m,i}),\end{aligned} \tag{7}$$

其中，$\mathcal{H}(\cdot)$表示熵，分布中的变量从经验回放缓冲区$\mathcal{D}$中采样。完整推导见附录 B.3。现在可以将互信息目标改写为：

$$\begin{aligned}&\mathcal{L}_{\text{MI}} = \\ &\sum_{m=1}^{M}\sum_{i=1}^{n} \mathbb{E}_{\mathcal{D}}[\log q_\xi(e_t^{m,i}|z_t^m, \tau_t^{m,i})] + \mathcal{H}(e_t^{m,i}|\tau_t^{m,i}),\end{aligned} \tag{8}$$

上述符号的定义与公式（5）类似。为促进学习过程，我们进一步设计了两个局部辅助优化目标。一方面，我们希望$e^m_t{}^i$能够识别团队情境，并适应突然变化的新队友，如z_t^m所示：

$$\mathcal{L}_{\text{LCE}} = \sum_{m=1}^{M}\sum_{n=1}^{n} \mathbb{E}[||e_t^{m,i} - \bar{e}^{m,i}||_2^2] - \log\det(R_{\{\bar{e}^{m,i}\}}). \tag{9}$$

另一方面，为了推导特定团队情境的描述性表示$e^m_t{}^i$我们希望$e^m_t{}^i$能够学习可控智能体与队友之间的关系。因此，我们期望$e^m_t{}^i$重构

队友采取的观测和动作：

$$\mathcal{L}_{\text{REC}} = \sum_{m=1}^{M} \sum_{n=1}^{n} \mathbb{E}_{\mathcal{D}}[-\log h_{\psi_i}(\bar{\boldsymbol{o}}_t^m, \bar{\boldsymbol{a}}_t^m | e_t^{m,i})], \tag{10}$$

其中， h 针对每个智能体 i由 ψ_i 参数化。由于 $e^m{}_{,t}{}^i$ 和 $\tau^m{}_{,t}{}^i$ 将拼接后输入个体Q网络 $Q^i(\tau^m{}_{,t}{}^i, e^m{}_{,t}{}^i, \cdot)$，因此也利用时序差分损失 $\mathcal{L}_{\text{TD}}$ 来促进局部上下文编码器的学习。由此，优化目标变为：

$$\mathcal{L}_{\text{DEC}} = \mathcal{L}_{\text{TD}} + \alpha_{\text{MI}}\mathcal{L}_{\text{MI}} + \alpha_{\text{LCE}}\mathcal{L}_{\text{LCE}} + \alpha_{\text{REC}}\mathcal{L}_{\text{REC}}, \tag{11}$$

其中， α_{MI}、 α_{LCE}、 α_{REC} 分别是三个目标函数对应的可调超参数。 s.

4 实验

本节中，我们设计了大量实验来探究以下问题：1）与不同场景下的其他基线方法相比，Fastap 在遇到队友行为突发变化时能否实现较强的适应性和泛化能力？各个组件又会如何影响其性能（第 4.2 节）？ 2）C RP 能否帮助学习到可区分的队友行为边界？Fastap 学习到了怎样的团队情境表示（第 4.3 节）？ 3）Fastap 展现出怎样的迁移能力？各个超参数又会如何影响其协作能力（第 4.4 节）？

4.1 环境与基线方法

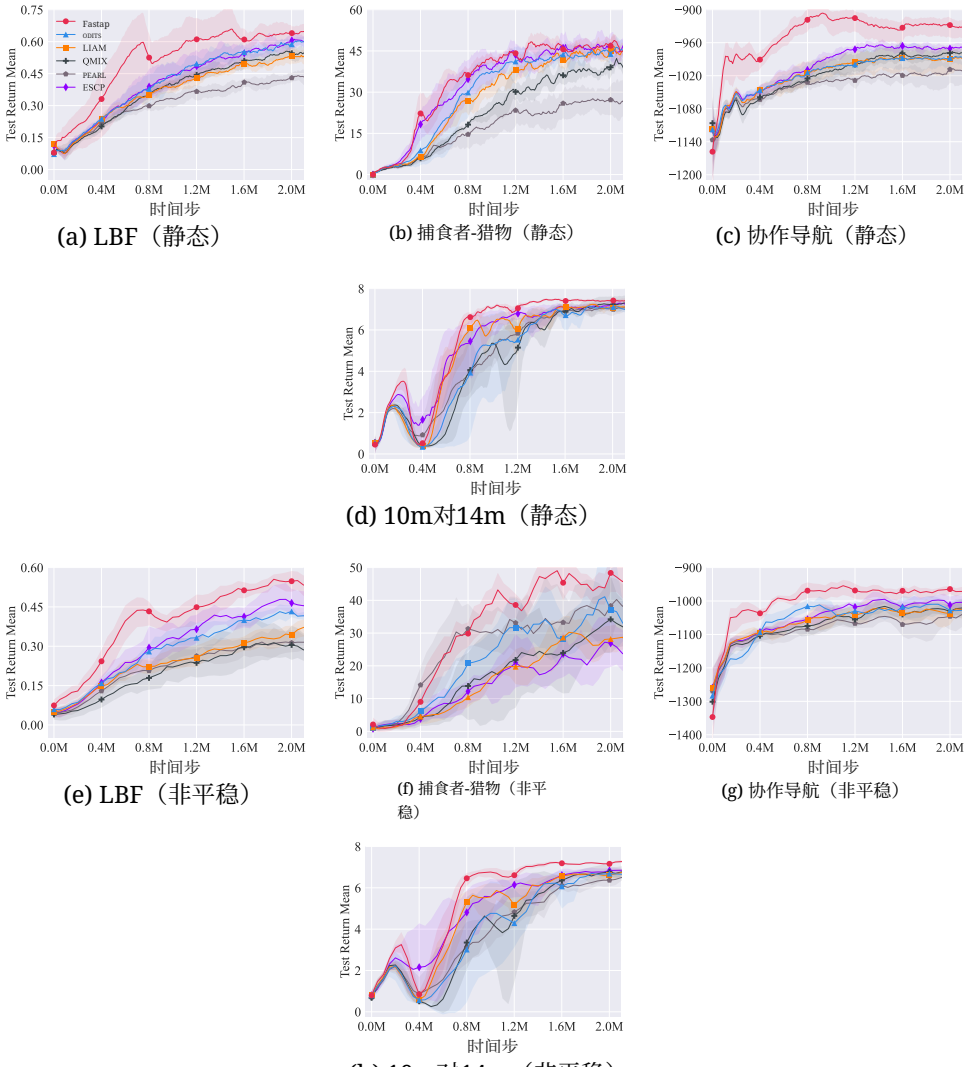
我们选择了四个多智能体任务作为环境，如图 2 所示。基于等级的觅食（LBF） [Papoudakis 等，2021b]是一种合作式网格世界游戏，其中智能体同时前往食物所在位置并将其收集后即可获得奖励。捕食者-猎物（PP）和协作导航（ CN）是源自 MPE环境 [Lowe 等人， 2017], 的两个场景：在 PP 中，多个智能体（捕食者）需要追逐并接触对抗智能体（猎物）以赢得游戏；在 CN 中，多个智能体经过训练，既要朝地标移动，又要避免彼此碰撞。我们还根据 SMAC [Samvelyan 等人， 2019], 创建了一张 10m对14m 地图，其中 10 个盟友在不同位置生成，并攻击 14 个敌人以获胜。

对于基线方法，我们考虑了多种方法，并将它们实现到一种流行的基于价值的方法 QMIX [Rashid 等人，2018] 以进行比较，包括：(1) 不含任何额外设计的原始 QMIX；(2) 元学习SARL方法：PEARL [Rakelly 等人， 2019] 使用近期收集的上下文来推断描述任务的概率变量；ESCP [Luo 等人， 2022] 通过学习上下文敏感策略来应对环境中的突发变化；(3) 基于上下文的多智能体强化学习方法：LIAM [Papoudakis 等，2021a] 基于局部观测历史预测队友当前的行为，以缓解训练阶段的非平稳性；ODITS [Gu 等人， 2022] 采用集中式“团队协作情境编码器”进行端到端学习，从而适应不同回合中的任意队友。关于环境和基线方法的更多详情，以及 Fastap 的详情，分别见附录 C 和附录 D。

4.2 竞争结果与消融实验

平稳与非平稳设置下的协作能力 首先，如图 3 所示，我们将 Fastap 与上述基线方法进行比较，以研究平稳和非平稳条件下的协作能力。我们发现，当队友处于非平稳状态时，所有算法的协作能力都会下降，这表明在非平稳环境中，需要特别考虑队友策略的突发变化。当仅使用局部信息获取上下文来捕捉队友的信息时，PERAL 和 LIAM 等方法在平稳和非平稳设置下的协作能力提升并不明显；PEARL 的表现甚至不如原始 QMIX，这表明，在单智能体强化学习中成功的元学习方法，未经修改便无法应用于多智能体强化学习设置。此外，当以全局和局部两种方式学习队友的行为上下文提取模型时，ODITS 在上述两种条件下均表现出色，体现了利用全局状态提高训练效率的必要性。此外，ESCP 也展现出相对较强的协作能力，证明了通过快速适应性优化上下文编码器的有效性。Fastap 在平稳和非平稳条件下的所有基准测试中均取得最佳表现，并且在大多数环境中，非平稳条件下测试时的性能下降幅度最小（见图 4），这表明所提出方法有效且高效。

消融实验 由于 Fastap 由多个组件构成，我们在基准测试 LBF 和 PP 上设计了消融实验，以研究这些组件如何影响 Fastap 在非平稳设置下的协调性能。



(a) LBF（静态） (b) 捕食者-猎物（静态） (c) 协作导航（静态）

(d) 10m对14m（静态）

(e) LBF（非平稳） (f) 捕食者-猎物（非平稳） (g) 协作导航（非平稳）

(h) 10m对14m（非平稳）

图3：在多个基准测试上与基线方法的性能比较。

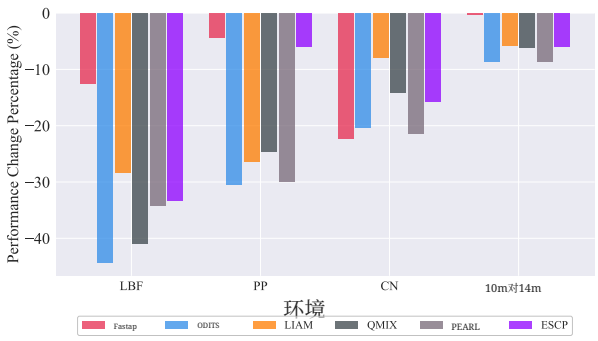
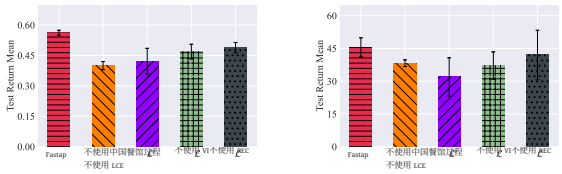


图4：平稳条件与非平稳条件下的性能差异。该值表示非平稳设置与平稳设置下的性能相对于最佳回报的差异。

$\mathcal{U}$	Fastap	Fastap wo 中国餐馆过程	ODITS	LIAM	QMIX	PEARL	ESCP
静态	$\mathbf{0.642 \pm 0.008}$	0.594 ± 0.015	0.637 ± 0.008	0.597 ± 0.029	0.569 ± 0.033	0.507 ± 0.021	0.618 ± 0.040
$U[5,8]$	$\mathbf{0.562 \pm 0.012}$	0.400 ± 0.020	0.352 ± 0.002	0.415 ± 0.026	0.306 ± 0.038	0.288 ± 0.019	0.404 ± 0.026
$U[6,7]$	$\mathbf{0.567 \pm 0.001}$	0.444 ± 0.314	0.487 ± 0.022	0.454 ± 0.157	0.444 ± 0.221	0.333 ± 0.000	0.556 ± 0.125
$U[2,9]$	0.484 ± 0.285	0.222 ± 0.133	0.416 ± 0.182	0.401 ± 0.078	0.443 ± 0.205	0.205 ± 0.114	$\mathbf{0.514 \pm 0.314}$
$U[3,6]$	$\mathbf{0.518 \pm 0.136}$	0.366 ± 0.217	0.444 ± 0.314	0.388 ± 0.283	0.353 ± 0.272	0.264 ± 0.066	0.502 ± 0.120
$U[3,3]$	$\mathbf{0.384 \pm 0.272}$	0.246 ± 0.141	0.342 ± 0.118	0.362 ± 0.208	0.222 ± 0.314	0.243 ± 0.172	0.271 ± 0.157

表 1：LBF 中的最终平均回报 $\pm$ 标准差，其中 $\mathcal{U}$ 是控制突变频率的开放式 Dec-POMDP 的突变概率分布，而 $U[m,n]$ 表示由 m 和 n 参数化的离散均匀分布。原始训练突变分布 $\mathcal{U} = U[5,8]$ 对应的行以灰色突出显示。



(a) LBF (b) PP

图 5：消融实验。

首先，对于动态队友生成的无限混合模型，我们通过移除 CRP过程，并将每个新生成的队友组视为一个新簇，得到 无 *CRP* 。接下来，为了探究对队友行为敏感的编码器是否有助于提高适应性，我们通过移除局部编码器的 $\mathcal{L}_{\text{LCE}}$ ，提出 无 *LCE* 。此外，我们提出 无 *MI* ，以研究最大化全局和局部上下文之间的互信息如何加快学习效率。最后，我们引入 无 *REC* ，以检验涉及智能体建模的辅助优化目标的影响。如图 5 所示，无 *CRP* 和无 *MI* 分别在 LBF 和 PP 中遭受最严重的性能下降，这表明引入 CRP 模型确有益处，且对队友行为敏感的编码器确实能帮助智能体快速适应队友的突变。此外，在移除$\mathcal{L}_{\text{MI}}$后，两个基准测试中无 *MI* 所表现出的性能差距证明，利用全局信息来促进局部上下文编码器的学习是必要的。最后，我们还发现，智能体建模有助于学习信息量更丰富的上下文，并略微提升协作表现。

（分布外）非平稳性设置下的比较。由于本研究考虑的是这样一种设置：不受控队友发生突发变化的频率遵循固定的概率分布$\mathcal{U}$设为均匀分布，我们通过在测试期间改变变化频率来评估泛化能力。LBF 上的实验在训练期间采用分布 $\mathcal{U} = U[5,8]$。如表 1 所示，我们通过改变分布 $\mathcal{U}$，比较了 LBF 中不同学习策略的最终回报。尽管不同方法在平稳条件下获得了相近的协调能力，但当队友策略的变化频率改变时，它们的性能都会大幅下降（例如，ODITS 在突发变化[3, 3]下的性能下降近一半）。另一方面，Fastap 和 ESCP 在分布内和 OOD 设置下大多表现出色的泛化能力。具体而言，在静态设置下，Fastap 比表现最佳的基线 ODITS 高出 0.005；而在原始非平稳设置下，差距扩大至 0.147。我们还发现，Fastap 在突发变化[2, 9], 设置下不如 ESCP；我们认为，这两种方法在 2 时间步的突发变化下都无法表现良好。

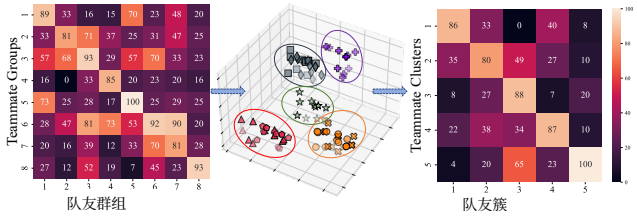
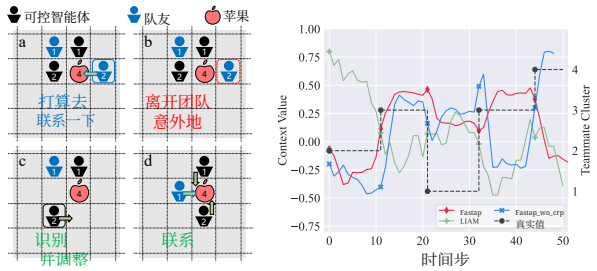


图 6：中国餐馆过程和队友行为嵌入前后的交叉对战表现 s.



(a) 快照 (b) 上下文曲线

图 7：队友适应可视化。

时间间隔，而 Fastap 在较大的时间步突发变化间隔下会牺牲部分性能，这种情况可能会在 $U[2,9]$ 中发生。未来将开发出能够适应多种条件的鲁棒策略。

4.3 队友适应性分析

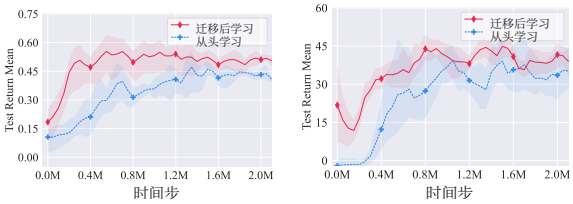
我们在此开展实验，以研究 CRP 模型和队友适应进程。首先，我们在 LBF 上分别于应用 CRP 前后进行交叉对战 [Hu 等人， 2020]实验，以验证 CRP 是否有助于获得可区分的队友行为边界。如图 6 左侧所示，对于 8 个队友群组的生成过程，我们发现从左上到右下的对角线上的数值相对较大。然而，其他若干点的表现也较高（例如队友群组 2 和 3），这表明生成的队友群组可能具有相似的行为。为减轻将行为相似的队友群组视为两种不同类型所造成的负面影响，我们采用 CRP 来学习行为类型，并将行为相似的队友分配到同一簇中。此外，我们对由 $E_{\omega_1}(\tau_k)$生成的潜变量进行采样，并通过主成分分析（PCA） [Wold 等人， 1987]降低其维度。我们发现，分配到同一簇的潜变量（椭圆区域）分布在相邻区域。我们还在应用 CRP 后的队友簇上进行了交叉对战实验；从图 6 右侧可以看出，属于不同簇的队友配对时表现较差，这表明 CRP 是有效的。

为研究队友行为敏感编码器如何帮助智能体快速适应队友的突发变化，我们还将测试期间某一回合的片段快照可视化，如图 7(a) 所示。当一名队友和两名受控智能体按预期尝试伸手摘取苹果并赢得分数时，该队友意外地离开了团队，导致他们暂时未能获得奖励。然而，由 Fastap 训练的受控智能体识别出了这一情况，并迅速切换策略，向下移动并与另一名队友协作以获得奖励。同时，我们记录了某一回合中不同时间步的潜在上下文向量。
Fastap 在 LBF 中将上下文编码为四维向量，我们通过主成分分析（PCA）将其降为一维标量。我们在图 7(b) 中绘制了这些点，并同时展示 LIAM 和消融版本 Fastap wo CRP 学到的上下文。结果表明，Fastap 学到的上下文能够敏锐地响应队友的突发变化；当队友保持稳定时，潜在上下文也稳定且平缓。尽管智能体建模有助于识别队友的行为，LIAM 的上下文曲线仍表现出滞后且不稳定的特征。同时，消融版本 Fastap wo CRP 也能快速适应新队友，但无法识别行为相似的队友，因此导致潜在上下文不稳定（例如，队友簇 3）。

4.4 迁移与敏感性研究

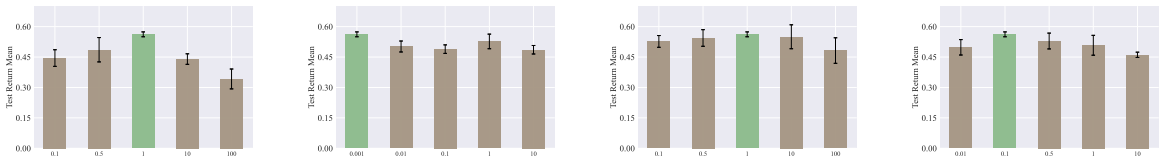
我们的 Fastap 学习队友识别模块，以应对在单个回合中可能突然变化的队友。控制变化频率的突发变化分布 $\mathcal{U}$ 是固定的；变化越频繁或等待间隔越长，训练就越困难。在此，我们通过比较微调后的表现与从头开始学习的表现，研究 Fastap 的策略迁移能力。具体而言，我们在突发变化分布 $\mathcal{U}_{\text{source}} = U[5,8]$ 下训练 Fastap 智能体 0.6M 个时间步，并在目标设置下使用 $\mathcal{U} = \mathcal{U}_{\text{target}} = U[3,6]$ 初始化训练好的网络。图 8 所示的学习曲线表明，在 $\mathcal{U}_{\text{source}}$ 下训练的智能体相比随机初始化具有先发优势；我们希望通过复用先前学到的知识，加速其在新环境中的学习。

由于 Fastap 包含多个超参数，我们在基准测试 LBF 上开展实验，以探究每个超参数对协作能力的影响。首先， α_{GCE} 用于平衡 TD 损失与全局上下文优化目标之间的权衡。如果它太小，智能体可能会过度适应静态环境中的协作，而忽视



(a) LBF (b) 捕食者-猎物

图 8：策略迁移能力。



(a) 敏感性 α_{GCE} (b) 敏感性 α_{MI} (c) α_{LCE}的敏感性 (d) α_{REC}的敏感性

图 9：LBF 敏感性研究。

对队友上下文信息的提取。另一方面，如果数值过大，智能体会过度关注队友识别，从而有过拟合特定队友类型的风险。因此，我们通过网格搜索确定每个超参数。如图 9(a) 所示，我们发现 $\alpha_{\text{GCE}} = 1$ 是该基准测试中的最佳选择。 α_{MI} 会影响局部编码器 f_{ϕ_i} 的优化，以识别团队情境。图 9(b) 显示， $\alpha_{\text{cont}_g} = 0.001$ 的表现最佳。我们可以用类似的方法发现， $\alpha_{\text{LCE}} = 1, \alpha_{\text{REC}} = 0.1$ 是相应的最佳选择。

5 结语

在这项工作中，我们研究了部分协调者遭遇突发策略变化时，队友的适应问题。我们首先将该问题形式化为开放式分散式部分可观测马尔可夫决策过程：团队中的部分协调者可能在一个回合内不可预测地持续发生策略变化，而我们训练多个受控智能体快速适应这种变化。为此，我们提出 Fastap，这是一种通过捕捉队友的策略变化信息来学习鲁棒多智能体协调策略的高效方法。来自不同基准测试的平稳与非平稳条件下的大量实验结果验证了 Fastap 的有效性，更多分析结果也从多个方面证实了这一点。我们的方法可视为开放环境设定下的一次初步尝试 [Zhou，2022] 在合作式多智能体强化学习领域；我们衷心希望它能成为将多智能体强化学习应用于实际场景的坚实基础。未来，研究多智能体强化学习系统中动作／观测空间的变化，或利用变换器等技术 [Vaswani 等人，2017] 从不同来源和不同程度的非平稳性中获得通用协调策略，都具有重要价值。

附录

6 相关工作

协作式多智能体强化学习 许多现实世界的问题由多个相互作用的智能体构成，通常可以建模为多智能体系统 [Dorri 等人， 2018]。在众多解决方案中，多智能体强化学习（多智能体强化学习） [Zhang 等人， 2021]凭借深度强化学习强大的问题解决能力取得了巨大成功 [Wang 等人， 2020]。此外，当智能体拥有共同目标时，这类问题称为合作式多智能体强化学习 [Oroojlooy 和 Hajinezhad， 2022], ，并已在多个领域取得显著进展，如寻路 [Sartoretti 等人， 2019], 主动电压控制 [Wang 等人， 2021], 以及动态算法配置 [Xue 等人， 2022], 等。人们提出了许多促进智能体间协作的方法，包括基于策略的方法（例如 MADDPG [Lowe 等人， 2017], MAPPO [Yu 等人， 2022]）、基于价值的方法系列，如 VDN [Sunehag 等人， 2018], QMIX [Rashid 等人， 2018], ，以及变换器 [Wen 等人， 2022]等其他技术和许多变体 [Gorsane 等人， 2022], 。这些方法在 SMAC [Samvelyan 等人， 2019], Hanabi [Yu 等人， 2022], GRF [Wen 等人， 2022]等各种任务中展现出卓越的协作能力。除上述方法及其相应变体外，人们还提出了许多其他方法，从不同方面研究合作式多智能体强化学习，

包括智能体间的因果推断 [Grimbly 等人， 2021], 面向现实应用的离线策略部署 [Yang 等人， 2021], 针对部分可观测性的通信 [Zhu 等人， 2022]通过模型学习提高样本效率 [Wang 等人， 2022],发生扰动时的策略鲁棒性 [Guo 等人，2022年，Yuan 等人，2023a], ，如 CTDE（集中式训练、分散式执行）等训练范式 [Lyu 等人，2021],用于持续协作验证的测试平台设计 [Nekoei 等人， 2021], 以及临时团队协作 [Mirsky 等人， 2022],多智能体强化学习中的离线学习 Guan 等人 [2023],Zhang 等人 [2023], 等。

非平稳性 是单智能体强化学习（单智能体强化学习）中的一个长期研究主题 [Padakandla 等人，2019年，Padakandla， 2020], 其中，学习系统的环境动态（例如转移函数和奖励函数）可能会随时间发生变化。对于单智能体强化学习，大多数现有工作关注回合间非平稳性，即决策过程在不同回合之间呈现非平稳性，包括多任务设置 [Varghese 和 Mahmoud， 2020], 持续强化学习 [Khetarpal 等人， 2022], 元强化学习 [Beck 等人，2023], 等。这些问题可以表述为上下文马尔可夫决策过程 [Hallak 等人， 2015], 并可通过任务嵌入学习等技术加以解决。其他工作还考虑了回合内非平稳性，即智能体可能在单个回合中遭遇动态漂移 [Kumar 等人，2021年，Ren 等人，2022年，Chen 等人，2022年，Luo 等人，2022年，Dastider 和 Lin，2022年，Feng 等人，2022]。具体而言，HDP-C-MDP [Ren 等人， 2022] 假设潜在上下文是有限且具有马尔可夫性的，并采用粘性层次狄利克雷过程（层次狄利克雷过程）先验进行模型学习；而 FANS-RL [Feng 等人， 2022]假设潜在上下文具有马尔可夫性，且环境可以建模为因子化马尔可夫决策过程；ESCP [Luo 等人，2022] 考虑了智能体可能遭遇的突发变化，并通过学习辅助上下文识别模型获得鲁棒策略。实验表明，在同时存在分布内和分布外参数变化的环境中，ESCP 不仅能更好地恢复环境编码，还能更快地适应变化后的环境；SeCBAD [Chen 等人， 2022] 进一步假设环境上下文通常会在一个随机时段内保持稳定，随后以突然而不可预测的方式发生变化。Linda Cao 等人 [2021] 学习分解局部信息，并为每个队友建立感知能力，从而提升在多个环境中的协作能力。

开放式多智能体系统研究的是智能体在进程进行期间可能加入或离开，从而导致系统组成和规模随时间变化的问题 Hendrickx 和 Martin [2017]。在以往的研究中，多智能体问题主要被建模为规划问题，因此形成了多种问题表述，例如开放式分散式部分可观测马尔可夫决策过程 Cohen 等人 [2017], 、团队部分可观测马尔可夫决策过程 Cohen 和 Mouaddib [2018, 2019], 、轻量级交互式部分可观测马尔可夫决策过程 Chandrasekaran 等人 [2016], 、Eck 等人 [2019], 提出的通信交互式部分可观测马尔可夫决策过程 Kakarlapudi 等人 [2022], ，以及其他问题。近年来，一些研究开始关注开放式多智能体强化学习问题。Rahman 等人提出的 GPL [2021] 将开放式临时团队协作建模为开放式随机博弈，并假设具备全局可观测性以提高效率，但这一条件在现实世界中可能难以实现。此外，该方法使用基于图神经网络的方法，只适用于单个可控智能体的情形，且扩展到多个可控智能体的情形时可扩展性不足。Yuan 等人提出的 ROMANCE [2023b] 将在不同环境中测试时出现策略扰动的问题建模为有限策略对抗者分散式部分可观测马尔可夫决策过程（LPA-Dec-POMDP），并提出**鲁棒多智**能体**协**同：通过**进**化生成辅助对抗攻击者（ROMANCE），使训练后的策略在训练期间能够应对多样且强大的辅助对抗攻击，从而在各种策略扰动下实现较高的鲁棒性。

与单智能体强化学习（SARL）的设置不同，非平稳性是多智能体强化学习（MARL）固有的挑战，因为其他智能体并行学习多种策略可能导致智能体的策略不稳定 [Papoudakis 等人， 2019]。以往的研究主要关注如何在训练阶段解决非平稳性，采用的技术包括智能体建模 [Albrecht 和 Stone， 2018], 元策略适应 [Kim 等人， 2021], 经验共享 [Christianos 等人， 2020]。其他研究则着眼于跨回合的非平稳性。以往的研究关注如何在训练阶段解决非平稳性，采用的技术包括多任务训练 [Qin 等人， 2022], 训练用于零样本协作的策略 [Hu 等人， 2020]。尽管这些方法取得了进展，但它们并未解决队友策略突发变化所导致的非平稳性，而这正是一个关键且迫切的需求。对于开放式多智能体强化学习，我们的工作采用了不同的视角，强调在多智能体强化学习的上下文中，已学习的可控智能体所具备的通用协调和快速适应能力。

7 推导详情

7.1 中国餐馆过程及簇分配推导详情

中国餐馆过程（CRP） [Blei 和 Frazier， 2010] 是一种离散时间随机过程，用于定义聚类结构上的先验分布，其描述如下：一位顾客走进一家中餐馆，以与浓度参数 α成正比的概率独自坐到一张新桌位，或以与已占用桌位上的顾客人数成正比的概率与其他顾客同桌。坐在同一桌位的顾客会被分配到同一个簇。具体来说，假设有 K 位顾客坐在

目前餐厅里。令 z_i 为一个指示变量，用来表示 i^{th} 坐在哪张桌位上；令 n_m 表示坐在第 m^{th} 张桌位上的顾客人数，令 M 表示非空桌位的总数。注意， $\sum_{m=1}^{M} n_m = K$。第 $K+1^{\text{th}}$ 位顾客坐到第 m^{th} 张桌位上的概率为:

$$P(z_{K+1} = m|\alpha) = \frac{n_m}{K+\alpha}, \quad m = 1, ..., M. \tag{12}$$

顾客决定坐到一张新桌位的概率不为零；如果新桌位的标签为 $M+1$，则:

$$P(z_{K+1} = M+1|\alpha) = \frac{\alpha}{K+\alpha}. \tag{13}$$

综合这两个方程即可刻画中国餐馆过程。

k^{th}生成的队友组 $P(v_k^{(m)}|\tau_k^S, \tau_k^A)$ 的簇分配可以分解为:

$$\begin{aligned} & P(v_k^{(m)}|\tau_k^S, \tau_k^A) \\ &= \frac{P(v_k^{(m}, \tau_k^S, \tau_k^A)}{P(\tau_k^S, \tau_k^A)} \\ &= \frac{P(\tau_k^S, \tau_k^A|v_k^{(m)})P(v_k^{(m)})}{P(\tau_k^S, \tau_k^A)} \\ &= \frac{P(\tau_k^A|\tau_k^S, v_k^{(m)})P(\tau_k^S|v_k^{(m)})P(v_k^{(m)})}{P(\tau_k^S, \tau_k^A)} \\ &\propto P(\tau_k^A|\tau_k^S, v_k^{(m)})P(\tau_k^S|v_k^{(m)})P(v_k^{(m)}). \end{aligned} \tag{14}$$

由于 τ_k^S 是一组状态，在忽略时间维度上的相关性的情况下，该状态组不由队友的行为类型决定。 $P(\tau_k^S|v_k^{(m)})$可以视为常数。因此，我们可以推导出$P(v_k^{(m)}|\tau_k^S, \tau_k^A) \propto P(v_k^{(m)})P(\tau_k^A|\tau_k^S;\ v_k^{(m)})$。

7.2 $\mathcal{L}_{\text{GCE}}$的完整推导

为了引导上下文编码器快速识别并跟踪突发变化，ESCP [Luo 等人 2022] 提出了以下优化目标:

$$\mathcal{L}_{\text{GCE}} = \sum_{m=1}^{M} \mathbb{E}[||z_t^m - \mathbb{E}[z_t^m]||_2^2] + ||\mathbb{E}[z_t^m] - u^m||_2^2, \tag{15}$$

其中， z_t^m是上下文编码器在 m^{th} 环境中嵌入的表征， u^m 是真实潜在上下文向量， M 是环境的数量。为了更好地理解，下面我们将结合本文的设定解释这些符号的含义。因此， z_t^m是与属于m^{th} 簇的队友配对时的潜在上下文向量，而 u^m 是真实行为类型。

由于我们无法获取真实的 u^m，因此需要一组具有较大多样性的替代变量，使其具备可分性和代表性。同时，u^m 是用于引导 $\mathbb{E}[z_t^m]$的中间变量，因此我们可以通过最大化关系矩阵$R_{\{\mathbb{E}[z_t^m]\}}$的行列式，直接最大化 $\{\mathbb{E}[z_t^m]\}_{m=1}^{M}$ 的多样性。关系矩阵的每个元素为:

$$R_{\{\mathbb{E}[z_t^m]\}}(i,j) = \exp(-\kappa||\mathbb{E}[z_t^i] - \mathbb{E}[z_t^j]||_2^2), \tag{16}$$

其中， κ 是径向基函数的半径超参数，该函数用于计算两个向量之间的距离。现在，目标函数可以写为:

$$\mathcal{L}_{\text{GCE}} = \sum_{m=1}^{M} \mathbb{E}[||z_t^m - \mathbb{E}[z_t^m]||_2^2] - \log\det(R_{\{\mathbb{E}[z_t^m]\}}). \tag{17}$$

为稳定训练过程，ESCP 将 $\mathbb{E}[z_t^m]$ 替换为 $\bar{z}^m$，即所有过去上下文向量的移动平均值。$\{\bar{z}^m\}$ 将在采样一个新的 z_t^m 批次后更新:

$$\bar{z}^m = \eta \text{sg}(\bar{z}^m) + (1-\eta)\mathbb{E}[z_t^m], \tag{18}$$

其中 sg(·) 表示停止梯度，而 η 是控制移动平均时间跨度的超参数 n.

7.3 队友上下文近似的变分界

为了使局部轨迹编码器 f_{ϕ_i} 生成的上下文向量 $e^{m,i}_t$ 与由 g_θ编码的全局上下文 z^m_t在信息上保持一致，我们提出最大化 $e^{m,i}_t$ 与 z^m_t之间的互信息，并以智能体 i的局部轨迹 $\tau^{m,i}_t$ 为条件。我们借鉴了变分推断的思路 [Alemi et al., 2017] ，并推导出该互信息项的一个下界。

定理 1. 令 $\mathcal{I}(e^{m,i}_t; z^m_t|\tau^{m,i}_t)$ 表示在智能体 i 的局部轨迹 $\tau^{m,i}_t$ 条件下，智能体 i 的局部上下文 $e^{m,i}_t$ 与全局上下文 z^m_t 之间的互信息。其下界为

$$\mathbb{E}_{\mathcal{D}}[\log q_\xi(e^{m,i}_t|z^m_t, \tau^{m,i}_t)] + \mathcal{H}(e^{m,i}_t|\tau^{m,i}_t). \tag{19}$$

此处 m 是在本回合中与受控智能体合作完成任务的队友所属的簇 ID。

证明。 通过一个由 ξ 参数化的变分分布 $q_\xi(e^{m,i}_t|z^m_t, \tau^{m,i}_t)$，我们有

$$\begin{aligned}
&\mathcal{I}(e^{m,i}_t; z^m_t|\tau^{m,i}_t) \\
=&\mathbb{E}_{\mathcal{D}}\left[\log \frac{p(e^{m,i}_t; z^m_t|\tau^{m,i}_t)}{p(e^{m,i}_t|\tau^{m,i}_t)p(z^m_t|\tau^{m,i}_t)}\right] \\
=&\mathbb{E}_{\mathcal{D}}\left[\log \frac{p(e^{m,i}_t|z^m_t; \tau^{m,i}_t)}{p(e^{m,i}_t|\tau^{m,i}_t)}\right] \\
=&\mathbb{E}_{\mathcal{D}}\left[\log \frac{q_\xi(e^{m,i}_t|z^m_t, \tau^{m,i}_t)}{p(e^{m,i}_t|\tau^{m,i}_t)}\right] + \\
&D_{\mathrm{KL}}(p(e^{m,i}_t|z^m_t, \tau^{m,i}_t)||q_\xi(e^{m,i}_t|z^m_t, \tau^{m,i}_t)) \\
\geq&\mathbb{E}_{\mathcal{D}}\left[\log \frac{q_\xi(e^{m,i}_t|z^m_t, \tau^{m,i}_t)}{p(e^{m,i}_t|\tau^{m,i}_t)}\right] \\
=&\mathbb{E}_{\mathcal{D}}[\log q_\xi(e^{m,i}_t|z^m_t, \tau^{m,i}_t)] + \mathcal{H}(e^{m,i}_t|\tau^{m,i}_t).
\end{aligned} \tag{20}$$

□

8 基线与基准的详细信息

8.1 使用的基线方法

QMIX [Rashid 等人， 2018]: 本文研究 Fastap 的整合能力，并在此介绍本文采用的基于价值的方法 QMIX [Rashid 等人， 2018]。我们提出的框架 Fastap 遵循基于价值的多智能体强化学习方法所采用的集中式训练与分散式执行（CTDE）范式，同时遵循个体-全局最大化（IGM） [Son 等人， 2019] 原则。该原则指出，联合价值函数 $Q_{\mathrm{tot}}(\boldsymbol{\tau}, \boldsymbol{a})$ 与个体价值函数 $\left[Q_i(\tau^i, a^i)\right]_{i=1}^n$所对应的联合贪心动作选择与局部贪心动作选择保持一致:

$$\begin{aligned}
&\forall \boldsymbol{\tau} \in \boldsymbol{\mathcal{T}}, \arg\max_{\boldsymbol{a}\in\boldsymbol{\mathcal{A}}} Q_{\mathrm{tot}}(\boldsymbol{\tau}, \boldsymbol{a}) = \\
&\left(\arg\max_{a^1\in\mathcal{A}} Q_1\left(\tau^1, a^1\right), \ldots, \arg\max_{a^n\in\mathcal{A}} Q_n\left(\tau^n, a^n\right)\right).
\end{aligned} \tag{21}$$

QMIX 将全局价值函数 $Q^{\mathrm{QMIX}}{}_{\mathrm{tot}}(\boldsymbol{\tau}, \boldsymbol{a})$ 分解为智能体局部价值函数 $\left[Q_i(\tau^i, a^i)\right]_{i=1}^n$的单调组合，从而扩展了 VDN:

$$\forall i \in \mathcal{N}, \frac{\partial Q^{\mathrm{QMIX}}_{\mathrm{tot}}(\boldsymbol{\tau}, \boldsymbol{a})}{\partial Q_i\left(\tau^i, a^i\right)} > 0. \tag{22}$$

我们主要在 QMIX 上实现 Fastap，因为多篇论文已证明其在各种任务中表现出色。QMIX 使用一个以全局状态为条件的超网络来生成局部 Q 值的权重与偏置，并通过绝对值运算确保权重为正，从而保证单调性。

PEARL [Rakelly 等人， 2019]： 这一基线源自单智能体和元学习领域，旨在根据某些隐表示来表征环境。具体而言，PEARL 利用转移数据作为上下文来推断环境的特征，并通过高斯分布乘积对其建模。将 PEARL 用于多智能体强化学习任务时，会采用 PEARL 模块，并针对每个可控智能体的局部上下文编码器进行优化。

ESCP [Luo 等人， 2022]： 作为一种单智能体强化学习算法，其旨在当环境发生突发变化时，能够快速识别并适应新环境，因此采用优化目标公式 15 来优化上下文编码器。为适应多智能体强化学习（MARL）的框架及具体任务，不对历史记录进行截断，并为每个可控智能体配备一个局部编码器。

LIAM [Papoudakis 等，2021a]： 该方法为每个智能体配备一个编码器-解码器结构，根据自身的局部观测历史 $\tau_t = \{o_{0:t}\}$，预测其他智能体在当前时间步的观测 $\boldsymbol{o}_t^{-1}$和动作 $\boldsymbol{a}_t^{-1}$。编码器和解码器经过优化，以最小化观测的均方误差与动作的交叉熵误差之和。为适用于我们工作中的多智能体强化学习设置，可控智能体的局部上下文编码器将根据其局部轨迹预测队友的观测和动作。其损失的均值用于优化编码器。

ODITS [Gu 等人， 2022]： 不同于前两种预测队友智能体实际行为的方法，ODITS 以端到端的方式提升零样本协作性能。我们采用两个变分编码器来提升协作能力。全局编码器将全局状态轨迹作为输入，并输出一个高斯分布。采样得到一个向量 z，并将其输入超网络；该超网络将临时智能体的局部效用Q_i 映射为全局效用 Q_{tot}，以逼近全局折扣回报。局部编码器具有类似的结构，采样得到的 e 被输入临时智能体的策略网络。编码器通过最大化回报以及在局部转移数据条件下两个上下文向量的互信息，以端到端的方式进行更新。由于 ODITS 仅考虑单个临时智能体，我们还为每个可控智能体配备局部轨迹编码器，并最大化互信息损失的均值，以适应我们的多智能体强化学习设置。

8.2 相关环境

基于等级的觅食 (LBF) [Papoudakis 等，2021b]： LBF 是一种混合了合作与竞争的部分可观测网格世界游戏，需要智能体高度协同才能完成收集食物的任务。每回合开始时，智能体和食物都会被随机分配等级和位置。每个智能体的动作空间包括向四个方向移动、收集身旁的食物以及“不操作”动作；但食物在整个回合中都不会移动。如果一组智能体的等级总和不低于食物的等级，就能收集该食物，并获得与食物等级相关的归一化奖励。智能体的主要目标是在有限时间内相互合作收集食物，从而最大化全局回报。

为了测试不同算法在此设定下的表现，我们考虑一个场景，其中最多有四个等级各异的智能体，以及三个最低等级分别为 $l \geq \sum_{i=1}^{3} sorted(levels)[i]$ 、位于 6×6 网格世界中的食物。智能体的视野有限，范围为 1（智能体周围的 3×3 个网格），回合时长上限为 25。在我们的开放式分散式部分可观测马尔可夫决策过程设定中，两个智能体可受控，并会在整个回合中留在环境内。队友数量可能为 1 或 2，策略网络也会相应改变。智能体获得的奖励为所收集食物的等级除以所有食物等级总和，具体如下：

$$r^i = \frac{\text{Food_with_Level_i}}{\sum_j \text{Food_with_Level_j}}. \tag{23}$$

捕食者-猎物（PP） [Lowe 等人， 2017]： 这是一个捕食者-猎物环境。优秀的智能体（猎物）速度更快，被对手（捕食者）击中时会受到负奖励（每次碰撞扣 10 分）。捕食者速度较慢，击中优秀智能体时会获得奖励（每次碰撞获得 +10 ）。障碍物会阻挡去路。默认情况下，环境中有 1 只猎物、3 只捕食者和 2 个障碍物。在我们的开放式分散式部分可观测马尔可夫决策过程设置中，两个捕食者是可控的，并会在整个回合中留在环境中。另一只捕食者是不可控的队友，其策略会突然变化。

协作导航（CN） [Lowe 等人， 2017]： 在这项任务中，四个智能体经过训练，要移动到四个地标处，同时避免彼此碰撞。所有智能体都会接收自身的速度、位置，以及相对于其他所有智能体和地标的位置。每个智能体的动作空间包含五种离散移动动作。智能体获得的奖励为：每个地标到任一智能体的最小距离的负值之和，此外还会增加一项惩罚智能体之间碰撞的奖励项。在我们的开放式分散式部分可观测马尔可夫决策过程设置中，两个智能体可受控，并将在

整个回合中留在环境内。队友数量可能为 1 或 2，策略网络也会相应改变。

星际争霸II微操基准（SMAC） [Samvelyan 等人， 2019]: SMAC 是一个星际争霸II单位微操任务的战斗场景。我们考虑部分观测设定，在该设定下，智能体只能看到以自身为中心、半径等于视野范围的圆形区域，视野范围设为 9。我们使用强化学习算法训练盟军单位，以击败由内置人工智能控制的敌方单位。每个回合开始时，盟军和敌军会在地图上的特定区域生成。每个智能体在每个时间步都会从离散动作空间中采取动作，包括以下动作：无操作、向 [方向], 移动、攻击 [敌方单位编号], 和停止。在这些动作的控制下，智能体可以在连续地图上移动和攻击。多智能体强化学习智能体在每个时间步都会获得全局奖励，其数值等于对敌方单位造成的总伤害。击杀每个敌方单位以及赢得战斗（击杀所有敌人）会分别带来 10 和 200 的额外奖励。这里我们创建了一张名为 10m 对 11m 的地图，其中 10 个盟军和 14 个敌军分别分成 2 组，并在不同地点生成，然后集合起来，对同一组敌人发动协同攻击以赢得任务。具体而言，我们控制 7 个盟军与另外 3 个队友协作完成任务，其中队友数量在一个回合内保持不变。

9 Fasta 的架构、基础设施和超参数选择 p

由于 Fastap 在主要实验中构建于 QMIX 之上，我们在本节详细介绍具体设置，包括网络架构、整体流程，以及不同环境选用的超参数。

9.1 网络架构

在本节中，我们将详细介绍以下网络：（1）CRP过程中的编码器 E_{ω_1} 和解码器 D_{ω_2}；（2）轨迹编码器 g_θ、f_{ϕ_i}和智能体网络；以及（3）变分分布 q_ξ 和队友建模解码器h_{ψ_i}。

8层Transformer编码器 E_{ω_1} 以全局轨迹 $\tau = (s_0, \boldsymbol{a}_0, \ldots, s_T)$作为输入，并输出16维行为嵌入 v。基于RNN的解码器 D_{ω_2}由一个隐藏维度为16的GRU单元组成，以 $\tau_t^X = (s_0, \ldots, s_t)$和 v作为输入，并重构动作 $\boldsymbol{a}_t$。

对于全局和局部轨迹编码器 g_θ 和 f_{ϕ_i}，我们将其分别设计为一个两层多层感知机和门控循环单元，隐藏维度为 64。然后通过一个线性层将嵌入转换为高斯分布的均值和标准差。上下文向量将从该分布中采样。全局上下文 z_t 和状态 s_t 将拼接后输入超网络。至于局部上下文 e_t^i，它将与局部轨迹 τ_t^i一起输入智能体 i 的个体Q网络。该网络包含一个维度为 64 的 GRU单元，用于编码历史信息，以及两个全连接层，用于计算局部Q值 $Q^i(\tau_t^i, e_t^i, \cdot)$。最后，局部Q值将输入混合网络，以计算时序差分损失。

为了在给定智能体 i的局部轨迹的条件下，最大化局部上下文向量与全局上下文向量之间的互信息，我们使用变分分布网络 q_ξ 来近似条件分布。具体而言， q_ξ 是一个隐藏维度为 64 的三层多层感知机，它输出一个高斯分布，并从中采样预测的局部上下文向量。智能体建模解码器 h_{ψ_i} 分为两个组成部分，即 $h_{\psi_i}^o$ 和 $h_{\psi_i}^a$，每个部分都是一个三层多层感知机。我们分别计算均方误差损失和最大似然损失，以优化目标函数。

9.2 Fastap 的整体流程

为了说明 Fastap 的整体流程，我们首先在算法 1 中展示基于CRP的无限混合过程。队友组可以通过任意多智能体强化学习算法生成，我们将少量轨迹存入经验回放缓冲区 $\mathcal{D}_k$（第 2~3 行）。训练编码器和解码器，使学习到的表征能够精准捕获行为信息，并精准估计预测似然（第 4 行）。随后，计算CRP先验和预测似然，以确定新生成的队友组的分配 m^* （第 5~7 行）。然后，我们根据分配结果更新现有簇，或实例化一个新簇（第 8~17 行）。

Fastap 的训练过程也如算法 2 所示。在轨迹采样阶段，我们首先从簇中采样一个队友组，并在本回合中固定该队友组。队友组与可控智能体配对，共同做出决策（第 3~12 行）。为了训练智能体策略网络和上下文编码器，移动

超参数 \ 环境	基于等级的觅食	捕食者-猎物	协作导航	10m对14m
浓度超参数 α	0.5	2.5	2.5	0.5
单次迭代中生成的队友组数量 L	4	1	1	2
半径超参数 κ	80	80	80	80
移动平均超参数 η	0.01	0.01	0.01	0.01
α_{GCE}	1	0.4	0.4	10
α_{LCE}	1	0.4	0.4	10
α_{MI}	0.001	0.001	0.001	0.001
α_{REC}	0.1	0.2	0.2	0.2
局部上下文向量维度 e	4	16	4	8
全局上下文向量维度 z	6	20	6	16

表 2：实验中的超参数。

平均上下文向量值会被更新，并计算优化目标（第 14~22 行）。此外，我们在算法 3 中展示了测试过程，其中队友可能会突然变化。突发变化分布 $\mathcal{U}$ 控制等待时间，从而决定变化频率（第 5~12 行）。

算法 1 Fastap：基于CRP的无限混合过程

输入：浓度参数 α，一次迭代中生成的队友群组数 L，迄今生成的队友群组数 K，迄今实例化的簇数 M，编码器 E_{ω_1}，解码器D

ω_2. 1：对于 $k = K+1, .., K+L$ 执行 2：生成 k^{th} 队友群组。 3：从 k^{th} 队友群组中采样一小批轨迹 τ_k，并将其存入 $\mathcal{D}_k$。 4：根据公式 4 中的 $\mathcal{L}_{\text{model}}$ 训练 E_{ω_1} 和 D_{ω_2}。 5：根据公式 2 计算中国餐馆过程先验 $P(v_k^{(m)})$ $m = 1,\ 2, ..., M+1$。 6：根据公式 3 计算预测似然 $P(\tau_k^Y|\tau_k^X;\ v_k^{(m)})$ $m = 1,\ 2, ..., M+1$。 7：$m^* = \arg\max_m P(v_k^{(m)})P(\tau_k^Y|\tau_k^X;\ v_k^{(m)})$。 8：如果 $m^* \leq M$，则 9：将 k^{th} 队友群组分配给 m^* 簇。 10：更新簇中心 $\bar{v}^{m^*} = \frac{n^{(m^*)}\bar{v}^{m^*} + v_k}{n^{(m^*)}+1}$。 11：更新簇的计数器 m：$n^{(m^*)} = n^{(m^*)} + 1$。 12：否则 13：使用 k^{th} 队友群组初始化 $M+1^{\text{th}}$ 簇。 14：初始化簇中心 $\bar{v}^{M+1} = v_k$。 15：初始化簇的计数器 $M+1$：$n^{(M+1)} = 1$。 16：更新 $M = M+1$。 17：结束 if 18：结束 for 19：更新 $K = K+L$。

我们的 Fastap 实现基于 EPyMARL[‡] [Papoudakis 等，2021b]代码库，使用 StarCraft 2.4.6.2.69223，并采用其默认超参数设置（例如，$\gamma = 0.99$）。不同环境所需的其他超参数选择列于表 2。

参考文献

Stefano V. Albrecht 和 Peter Stone。自主智能体对其他智能体建模：全面综述与开放问题。人工智能，258：66–95，2018。

Alexander A. Alemi、Ian Fischer、Joshua V. Dillon 和 Kevin Murphy。深度变分信息瓶颈。发表于*ICLR*，2017。

Jacob Beck、Risto Vuorio、Evan Zheran Liu、Zheng Xiong、Luisa Zintgraf、Chelsea Finn 和 Shimon Whiteson。《元强化学习综述》。预印本 *arXiv:2301.08028*，2023。

[‡]https://github.com/oxwhirl/epymarl

算法 2 Fastap：训练过程

输入：可控智能体策略网络 $\{\pi^i\}_{i=1}^n$、全局轨迹编码器 g_θ、局部轨迹编码器 $\{f_{\phi_i}\}_{i=1}^n$，
队友组簇 $\mathcal{C}$、迄今实例化的簇数 M、回合长度 T、采样回合数
$sample\text{-}num$、环境 env。
1：初始化移动平均值 $\bar{z}^m = \mathbf{0}$，$m = 1$，..... M。
2：初始化移动平均 $\bar{e}^{m,\,i} = \mathbf{0}$，$m = 1$，..... M；$i = 1$，..，n。
3: **对于** $l = 1, ..., sample\ num$ **执行 do**
4: 从属于 m^{th} 群集的 $\mathcal{C}$ 队友组样本。
5: $s_0^m = env.start()$.
6: **适用于** $t = 0$**的** $, ..., T$ **do**
7: $e_t^{m,i} = f_{\phi_i}(\tau^{m}_{,t}{}^{i})$，$i = 1$, ..., n。
8: $a_t^{m,i} = \pi^i(\tau_t^{m,i} e^{m}_{,t}{}^{i},)$ $i = 1$, ..., n。
9: $\boldsymbol{a}_t^m = (a^{m}_{,t}{}^{i})_{i=1}^n$。 // 可控智能体决策
10: $\bar{\boldsymbol{a}}_t^m = \bar{\boldsymbol{\pi}}^m(\bar{\boldsymbol{\tau}}_t^m)$。 // 不可控队友决策
11: $s_{t+1}^m, r_t^m = env.step(\langle \boldsymbol{a}_t^m, \bar{\boldsymbol{a}}_t^m \rangle)$.
12: **循环结束**
13: 将轨迹添加到经验回放缓冲区 $\mathcal{D}$
14: **适用于** $m = 1$**的** $, .., M$ **do**
15: 从 $\mathcal{D}$ 中采样 bs轨迹。
16: 计算估计的 Q 值和上下文向量 $z_t^m = g_\theta(\tau_t^m)$ $e^{m}_{,t}{}^{i} = f_{\phi_i}(\tau^{m}_{,t}{}^{i})$，$t = 0$, ..., T。
17: 更新 $\bar{z}^m = \eta$sg($\bar{z}^m$) +(1 − η)平均值(z_t^m)。
18: 更新 $\bar{e}^{m,i} = \eta$sg($\bar{e}^{m,i}$) +(1 − η)均值($e^{m}_{,t}{}^{i}$).
19: 根据 $\mathcal{L}_{\text{TD}}$优化智能体Q网络。
20: **结束循环**
21: 根据公式 6 中的 $\mathcal{L}_{\text{ADAP}}$ 优化 g_θ。
22: 根据 $\mathcal{L}_{\text{DEC}}$ 在公式 11 中优化 $\{f_{\phi_i}\}_{i=1}^n$。
23: **结束循环**

算法 3 Fastap：测试流程 输入：可控智能体策略网络 $\{\pi^i\}_{i=1}^n$，**局部轨迹编码器** $\{f_{\phi_i}\}_{i=1}^n$，**回合长度**T，
测试回合数 $test\ num$，**环境** env，**突发变化分布** $\mathcal{U}$，**队友集合**$\mathcal{N}$。**1: 对于** $l = 1$, ..., $test\ num$ **执行 2: 从** $\mathcal{N}$**中采样队友策略** $\bar{\boldsymbol{\pi}}$。**3:** $s_0 = env.start()$。**4: 对于** $t = 0$, ..., T**执行 5: 如果** $t = 0$**则 6: 采样等待时间** $u_0 \sim \mathcal{U}$。**7: 否则 8: 更新等待时间** $u_t = u_{t-1} - 1$。**9: 如果** $u_t \leq 0$**则 10: 重新采样** $u_t \sim \mathcal{U}$。**11: 从** $\mathcal{N}$**中重新采样队友策略** $\bar{\boldsymbol{\pi}}$。**12: 结束如果 13: 结束如果 14:** $e_t^i = f_{\phi_i}(\tau_t^i)$ $i = 1$, ..., n。**15:** $a_t^i = \pi^i(\tau_t^i, e_t^i)$ $i = 1$, ..., n。**16:** $\boldsymbol{a}_t = (a_t^i)_{i=1}^n$。// **可控智能体决策 17:** $\bar{\boldsymbol{a}}_t = \bar{\boldsymbol{\pi}}(\bar{\boldsymbol{\tau}}_t)$。// **不可控队友决策 18:** $s_{t+1}, r_t, done = env.step(\langle \boldsymbol{a}_t, \bar{\boldsymbol{a}}_t \rangle)$。**19: 结束循环 20: 结束循环**

David M. Blei 和 Peter I. Frazier。距离依赖的中国餐馆过程。载于*ICML*，第 87–94 页，2010。

Jiahan Cao、Lei Yuan、Jianhao Wang、Shaowei Zhang、Chongjie Zhang、Yang Yu 和 De-Chuan Zhan。Linda：面向队友感知的多智能体局部信息分解。*arXiv* 预印本：*2109.12508*，2021年。Muthukumaran Chandrasekaran、A. Eck、Prashant Doshi 和 Leen-Kiat Soh。开放式与类型化智能体系统中的个体规划。发表于*UAI*，2016年。Xiaoyu Chen、Xiangming Zhu、Yufeng Zheng、Pushi Zhang、Li Zhao、Wenxue Cheng、Peng CHENG、Yongqiang Xiong、Tao Qin、Jianyu Chen 和 Tie-Yan Liu。针对具有分段稳定上下文的非平稳环境的自适应深度强化学习方法。发表于*NeurIPS*，2022年。Kyunghyun Cho、Bart van Merri¨enboer、Caglar Gulcehre、Dzmitry Bahdanau、Fethi Bougares、Holger Schwenk 和 Yoshua Bengio。使用 RNN 编码器-解码器学习用于统计机器翻译的短语表征。发表于*EMNLP*，第 1724–1734 页，2014年。Filippos Christianos、Lukas Sch¨afer 和 Stefano V. Albrecht。用于多智能体强化学习的共享经验演员-评论家。发表于*NeurIPS*，2020年。Jonathan Cohen 和 Abdel-Illah Mouaddib。在不确定性下进行团队重组的蒙特卡洛规划：模型与性质。发表于*ICTAI*，第 458–465 页，2018年。Jonathan Cohen 和 Abdel-Illah Mouaddib。不确定性下团队重组规划的权力指数。发表于*AAMAS*，2019年。Jonathan Cohen、Jilles Steeve Dibangoye 和 Abdel-Illah Mouaddib。开放式分散式部分可观测马尔可夫决策过程。发表于*ICTAI*，第 977–984 页，2017年。Apan Dastider 和 Mingjie Lin。针对非平稳环境、带有策略性撤退的非参数随机策略梯度。发表于*CASE*，第 1377–1384 页。IEEE，2022年。Ali Dorri、Salil S. Kanhere 和 Raja Jurdak。多智能体系统：综述。*IEEE Access*，6:28573–28593，2018年。A. Eck、Maulik Shah、Prashant Doshi 和 Leen-Kiat Soh。开放式与类型化多智能体系统中的可扩展决策理论规划。发表于*AAAI*，2019年。Fan Feng、Biwei Huang、Kun Zhang 和 Sara Magliacane。面向非平稳强化学习的因子化适应。发表于*NeurIPS*，2022年。Rihab Gorsane、Omayma Mahjoub、Ruan John de Kock、Roland Dubb、Siddarth Singh 和 Arnu Pretorius。迈向合作式多智能体强化学习的标准化性能评估协议。发表于*NeurIPS*，2022年。St John Grimbly、Jonathan Shock 和 Arnu Pretorius。因果多智能体强化学习：综述与开放问题。*arXiv* 预印本：*2111.06721*，2021年。Pengjie Gu、Mengchen Zhao、Jianye Hao 和 Bo An。部分可观测条件下的在线临时团队协作。发表于*ICLR*，2022年。Cong Guan、Feng Chen、Lei Yuan、Zongzhang Zhang 和 Yang Yu。通过自监督信息聚合实现在线与离线多智能体强化学习中的高效通信。*arXiv* 预印本：*2302.09605*，2023年。Jun Guo、Yonghong Chen、Yihang Hao、Zixin Yin、Yin Yu 和 Simin Li。迈向对合作式多智能体强化学习鲁棒性的全面测试。*arXiv* 预印本：*2204.07932*，2022年。Assaf Hallak、Dotan Di Castro 和 Shie Mannor。上下文马尔可夫决策过程。*arXiv* 预印本：*1502.02259*，2015年。Julien M Hendrickx 和 Samuel Martin。开放式多智能体系统：具有随机到达与离开的信息传播。发表于*CDC*，第 763–768 页，2017年。Hengyuan Hu、Adam Lerer、Alex Peysakhovich 和 Jakob N. Foerster。用于零样本协作的“他人对弈”。发表于*ICML*，第 4399–4410 页，2020年。Anirudh Kakarlapudi、Gayathri Anil、Adam Eck、Prashant Doshi 和 Leen-Kiat Soh。开放式多智能体系统中带通信的决策理论规划。发表于*UAI*，2022年。Khimya Khetarpal、Matthew Riemer、Irina Rish 和 Doina Precup。迈向持续强化学习：综述与展望。人工智能研究杂志，75:1401–1476，2022年。Dong-Ki Kim、Miao Liu、Matthew Riemer、Chuangchuang Sun、Marwa Abdulhai、Golnaz Habibi、Sebastian Lopez-Cot、Gerald Tesauro 和 Jonathan P. How。用于在多智能体强化学习中学习如何学习的策略梯度算法。发表于*ICML*，第 5541–5550 页，2021年。Ashish Kumar、Zipeng Fu、Deepak Pathak 和 Jitendra Malik。RMA：腿式机器人的快速运动适应。载于 Dylan A. Shell、Marc Toussaint 和 M. Ani Hsieh 编，机器人学：科学与系统 *XVII*，2021年。

Soochan Lee、Junsoo Ha、Dongsu Zhang 和 Gunhee Kim。用于无任务持续学习的神经狄利克雷过程混合模型。发表于*ICLR*，2020年。Ryan Lowe、Yi Wu、Aviv Tamar、Jean Harb Pieter Abbeel 和 Igor Mordatch。面向混合合作竞争环境的多智能体演员-评论家。发表于*NIPS*，第 6379–6390 页，2017年。Fan-Ming Luo、Shengyi Jiang、Yang Yu、Zongzhang Zhang 和 Yi-Feng Zhang。通过学习上下文敏感策略适应环境突发变化。发表于*AAAI*，第 7637–7646 页，2022年。Xueguang Lyu、Yuchen Xiao、Brett Daley 和 Christopher Amato。比较多智能体强化学习中的集中式与分散式评论家。发表于*AAMAS*，第 844–852 页，2021年。Reuth Mirsky、Ignacio Carlucho、Arrasy Rahman、Elliot Fosong、William Macke、Mohan Sridharan、Peter Stone 和 Stefano V Albrecht。临时团队协作综述：定义、方法与开放问题。*arXiv* 预印本：*2202.10450*，2022年。Hadi Nekoei、Akilesh Badrinaaraayanan、Aaron C. Courville 和 Sarath Chandar。将持续协调作为终身学习的现实场景。发表于*ICML*，第 8016–8024 页，2021年。Frans A Oliehoek 和 Christopher Amato。分散式部分可观测马尔可夫决策过程简明导论。Springer，2016年。Afshin Oroojlooy 和 Davood Hajinezhad。合作式多智能体深度强化学习综述。应用智能，第 1–46 页，2022年。Sindhu Padakandla。动态变化环境中强化学习算法综述。*ACM* 计算综述（*CSUR*），54:1–25，2020年。Sindhu Padakandla、Prabuchandran K. J. 和 Shalabh Bhatnagar。面向非平稳环境的强化学习算法。应用智能，第 1–17 页，2019年。Georgios Papoudakis、Filippos Christianos、Arrasy Rahman 和 Stefano V Albrecht。应对多智能体深度强化学习中的非平稳性。*arXiv* 预印本：*1906.04737*，2019年。Georgios Papoudakis、Filippos Christianos 和 Stefano Albrecht。部分可观测条件下用于深度强化学习的智能体建模。发表于*NeurIPS*，第 19210–19222 页，2021a。Georgios Papoudakis、Filippos Christianos、Lukas Sch¨afer 和 Stefano V Albrecht。合作任务中的多智能体深度强化学习算法基准测试。发表于*NeurIPS*，2021b。Rongjun Qin、Feng Chen、Tonghan Wang、Lei Yuan、Xiaoran Wu、Zongzhang Zhang、Chongjie Zhang 和 Yang Yu。通过任务关系建模实现多智能体策略迁移。*arXiv* 预印本：*2203.04482*，2022年。Muhammad A Rahman、Niklas Hopner、Filippos Christianos 和 Stefano V Albrecht。通过基于图的策略学习实现开放式临时团队协作。发表于*ICML*，第 8776–8786 页，2021年。Kate Rakelly、Aurick Zhou、Chelsea Finn、Sergey Levine 和 Deirdre Quillen。通过概率上下文变量实现高效的离策略元强化学习。发表于*ICML*，第 5331–5340 页，2019年。Tabish Rashid、Mikayel Samvelyan、Christian Schroeder、Gregory Farquhar、Jakob Foerster 和 Shimon Whiteson。QMIX：用于深度多智能体强化学习的单调价值函数分解。发表于*ICML*，第 4295–4304 页，2018年。Hang Ren、Aivar Sootla、Taher Jafferjee、Junxiao Shen、Jun Wang 和 Haitham Bou-Ammar。离散马尔可夫上下文演化下的强化学习。发表于*ICLR*，2022年。Mikayel Samvelyan、Tabish Rashid、Christian Schr¨oder de Witt、Gregory Farquhar、Nantas Nardelli、Tim G. J. Rudner、Chia-Man Hung、Philip H. S. Torr、Jakob N. Foerster 和 Shimon Whiteson。星际争霸多智能体挑战。发表于*AAMAS*，第 2186–2188 页，2019年。Guillaume Sartoretti、Justin Kerr、Yunfei Shi、Glenn Wagner、TK Satish Kumar、Sven Koenig 和 Howie Choset。PRIMAL：通过强化学习与模仿式多智能体学习进行路径规划。*IEEE* 机器人与自动化快报，4(3):2378–2385，2019年。Kyunghwan Son、Daewoo Kim、Wan Ju Kang、David Earl Hostallero 和 Yung Yi。QTRAN：通过变换学习分解合作式多智能体强化学习中的价值函数。发表于 *ICML*，第 5887–5896 页，2019年。Peter Sunehag、Guy Lever、Audrunas Gruslys、Wojciech Marian Czarnecki、Vin´ıcius Flores 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。发表于*AAMAS*，第 2085–2087 页，2018年。Nelson Vithayathil Varghese 和 Qusay H. Mahmoud。多任务深度强化学习综述。电子学，2020年。

Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N. Gomez、Lukasz Kaiser 和 Illia Polosukhin。Attention is all you need。发表于*NIPS*，第 5998–6008 页，2017 年。Haonan Wang、Ning Liu、Yiyun Zhang、Dawei Feng、Feng Huang、Dongsheng Li 和 Yiming Zhang。深度强化学习：综述。*Frontiers ofInformation Technology & Electronic Engineering*，21：1726–1744，2020 年。Jianhong Wang、Wangkun Xu、Yunjie Gu、Wenbin Song 和 Tim C. Green。用于配电网络主动电压控制的多智能体强化学习。发表于*NeurIPS*，第 3271–3284 页，2021 年。Xihuai Wang、Zhicheng Zhang 和 Weinan Zhang。基于模型的多智能体强化学习：近期进展与展望。预印本 *arXiv:2203.10603*，2022年。Muning Wen、Jakub Grudzien Kuba、Runji Lin、Weinan Zhang、Ying Wen、Jun Wang 和 Yaodong Yang。多智能体强化学习是一个序列建模问题。发表于*NeurIPS*，2022年。Svante Wold、Kim Esbensen 和 Paul Geladi。主成分分析。*Chemometrics and Intelligent Laboratory Systems*，2(1-3)：37–52，1987 年。Ke Xue、Jiacheng Xu、Lei Yuan、Miqing Li、Chao Qian、Zongzhang Zhang 和 Yang Yu。多智能体动态算法配置。发表于*NeurIPS*，2022年。Yiqin Yang、Xiaoteng Ma、Chenghao Li、Zewu Zheng、Qiyuan Zhang、Gao Huang、Jun Yang 和 Qianchuan Zhao。相信所见：离线多智能体强化学习的隐式约束方法。发表于 *NeurIPS*，第 10299–10312 页，2021 年。Chao Yu、Akash Velu、Eugene Vinitsky、Jiaxuan Gao、Yu Wang、Alexandre Bayen 和 Yi Wu。近端策略优化在合作式多智能体博弈中的效果出人意料。发表于*NeurIPS*，2022年。Lei Yuan、Feng Chen、Zongzhang Zhang 和 Yang Yu。通过可适应辅助多智能体对手生成实现通信鲁棒多智能体学习。预印本 *arXiv:2305.05116*，2023a。Lei Yuan、Ziqian Zhang、Ke Xue、Hao Yin、Feng Chen、Cong Guan、Lihe Li、Chao Qian 和 Yang Yu。通过进化生成辅助对抗攻击者实现鲁棒多智能体协调。发表于*AAAI*，2023b。Fuxiang Zhang、Chengxing Jia、Yi-Chen Li、Lei Yuan、Yang Yu 和 Zongzhang Zhang。从多任务离线数据中发现可泛化的多智能体协调技能。发表于*ICLR*，2023年。Kaiqing Zhang、Zhuoran Yang 和 Tamer Bas¸ar。多智能体强化学习：理论与算法精选综述。*Handbook of Reinforcement Learning and Control*，第 321–384 页，2021 年。Zhi-Hua Zhou。开放环境机器学习。*National Science Review*，9(8)，2022年。Changxi Zhu、Mehdi Dastani 和 Shihan Wang。带通信的多智能体强化学习综述。预印本 *arXiv:2203.08975*，2022年。