异质目标下学习何时合作

Max Taylor-Davies[1],**Neil Bramley**[2],**Christopher G. Lucas**[1]

[1]爱丁堡大学信息学院 [2]爱丁堡大学心理学系

maxtaylordavies@gmail.com,

arXiv:2603.07253v1 [cs.MA] 7 Mar 2026

摘要

人类合作智能的一个重要组成部分，在于我们能够识别开展富有成效的协作的机会；反之，也能判断当前任务何时更适合独自完成。尽管这一元层次问题对于在异质、开放式环境中成功协作至关重要，机器灵活合作领域的研究却基本未对此展开探讨。本文将典型的临时团队协作（AHT）设定扩展为允许智能体具有异质目标：在任何特定情境下，这些目标可能重合，也可能并不重合。我们提出了一种在这一设定下学习策略的新方法，该方法基于模仿学习与强化学习的层次化结合；并通过两个合作环境的扩展版本证明，其表现优于基线方法。我们还探究了一个辅助模块的作用，该模块通过预测队友的行动来学习对其建模。研究发现，该模块对性能的影响与有关队友目标的可观测信息量呈负相关。

1 引言

人类的合作能力和协作解决问题的能力，是我们最卓越的适应能力之一；可以说，它们在促使我们发展出跨越时空的文化和制度方面发挥了关键作用（Tomasello 等人，2012；Henrich，2018）。因此，研究人工智能和机器人学的人员长期以来一直致力于打造具有人类般合作能力的自主系统，这不足为奇。最近，许多研究以“临时团队协作”（AHT）为主题，旨在开发能够与此前未曾谋面的队友“即时”协作的智能体。尽管取得了令人瞩目的进展，但通常所设想的临时团队协作情境忽略了现实世界协作中至关重要的某些方面。具体而言，临时团队协作往往假设每种场景都同样适合合作；也就是说，在任何给定的“回合”中，临时团队协作智能体总是始终 应该与环境中的其他智能体进行某种形式的协作。现实世界并非如此简单——有些场景为协作提供了有利机会，而在另一些场景中，独立行动、专注于我们独自能够取得的成果则更为合理。例如，如果你和朋友的目的地相同，拼车同行就很合理；但如果你们前往的方向相反，就不适合这样做。真正具有人类特质的协作者应当能够区分这些情形，并据此调整自己的行为。

尽管临时团队协作的定义明确允许智能体使用不同的奖励函数（Mirsky 等人，2022年），但以往研究对此鲜有探索。我们关注这样一种情形：智能体总体上拥有相同的根本目标（例如收集水果），但可能追求该目标的不同“变体”（例如收集苹果与收集橙子）。重要的是，智能体彼此之间事先可能并不知道对方的目标。我们的贡献有三点。首先，我们对这一情形进行了初步描述和形式化。其次，我们扩展了临时团队协作研究中使用的两种热门任务环境，使其支持异质目标。最后，我们提出了**GRILL**（通过强化学习与模仿进行低层控制的目标选择），这是一种新颖的分层方法，在这两种环境中的回报均高于基线方法，并且对不同等级的合作机会表现出更强的敏感性。

版权所有 © 2026，美国人工智能促进协会（www.aaai.org）。保留所有权利。

2 相关工作

临时团队协作

临时团队协作（AHT）领域关注这样一个问题：如何开发能够与此前从未见过的“队友”合作的智能体，且事先没有协调或沟通（Mirsky 等人，2022）。AHT 与多智能体强化学习（MARL）领域有一些基本共通之处；但 MARL 通常假定能够控制环境中的所有智能体，而在 AHT 中，我们只能控制单个智能体（通常称为“AHT 智能体”或“学习者”；本文将始终使用“自我智能体”这一称谓）。从本质上说，AHT 旨在适应潜在队友群体中的某种行为多样性。在实践中，这种多样性大多体现为队友采用不同风格的策略，但除此之外彼此同质，且由简单启发式方法或预训练（冻结）的强化学习策略控制。许多方法都涉及对队友策略类型进行显式推断和表示；传统上，这通常通过在离散队友空间上进行贝叶斯信念更新实现（Gmytrasiewicz 和 Doshi，2005；Barrett、Stone 和 Kraus，2011；Albrecht 和 Ramamoorthy，2013；Albrecht、Crandall 和 Ramamoorthy，2016），或者更

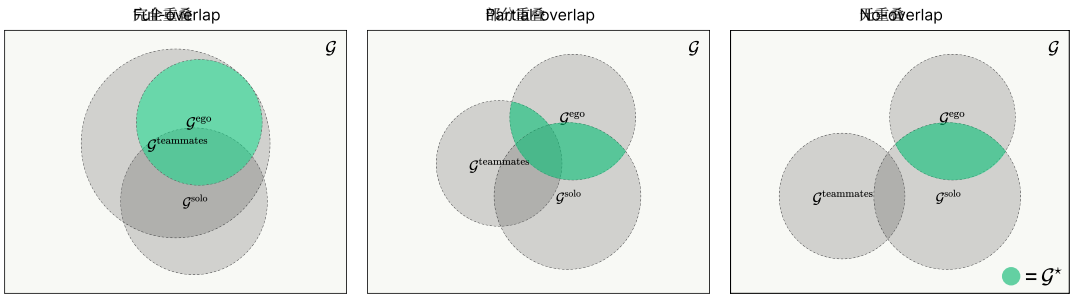

图1：三个不同场景下目标空间的示意图，其中值得追求（可获得奖励且原则上可实现）的目标集合以绿色突出显示。$\mathcal{G}^{\text{ego}}$ 表示能够为自我智能体带来奖励的目标集合；$\mathcal{G}^{\text{teammates}}$ 表示能够为至少一名队友带来奖励的目标集合；$\mathcal{G}^{\text{solo}}$ 表示单个智能体通过自身努力即可实现的目标集合。

近期则使用基于神经网络的编码器来学习潜在策略表示（Rabinowitz 等，2018；Papoudakis 和 Albrecht，2020；Rahman 等，2023a）。近期也有一些研究关注队友能力的异质性（例如，不同的机器人形态）（Liemhetcharat 和 Veloso，2014；Liu 等，2024）。然而，尽管 AHT 问题的定义允许智能体在更广泛的合作任务中拥有各自独立的奖励函数，绝大多数现有研究仍假定自我智能体及其所有队友共用单一奖励函数。虽然这一假设准确描述了某些现实世界中的合作任务，但它有所局限——人类环境通常存在追求各种不同目标的人，而这些目标之间的重合程度各不相同；为在这类环境中运行而设计的智能体，不应盲目假定自己遇到的每个人都是潜在的合作伙伴。在当前研究中，我们聚焦于这一尚未得到充分探索的多样性维度，考虑这样一种情境：在给定的高层任务下，智能体追求的目标可能各不相同。

分层强化学习

传统强化学习只考虑单一分辨率下的动作，而分层方法则在多个时间抽象等级上运行。通过以这种方式分解学习问题，分层强化学习算法原则上能够获得比非分层方法更低的样本复杂度和更好的泛化性能。尽管近期研究主要聚焦于更通用的“选项”框架（Sutton、Precup 和 Singh，1999；Bacon、Harb 和 Precup，2017；Barreto 等，2019；Klissarov 和 Precup，2021），但与我们的工作更相关的是早期的封建强化学习（FRL）（Dayan 和 Hinton，1992）。在封建强化学习中，智能体维护彼此独立的高层策略和低层策略：高层策略用于选择“子目标”，低层策略则用于根据给定的子目标选择基本动作。当这两个等级之间的分工明确时，封建强化学习效果良好；因此，对于当前场景而言，它是一个合理的起点，因为每个环境都由一组固定的目标来表征。与我们在此介绍的方法相比，主要区别在于，封建强化学习通过形式如下的奖励函数来优化低层策略 $R(s,\ a \mid s) = \mathbb{I}(s = g)$（见下一节），而我们使用带有学习所得目标标签的行为克隆（仅使用强化学习来学习高层目标选择）。

目标条件策略学习

另一类扩展基础强化学习范式的方法是目标条件强化学习（GCRL，有时也称为多目标强化学习），在这种方法中，智能体会学习一种以不同“目标” g为条件的动作策略；即 $\pi(a \mid o,\ g)$（Kaelbling，1993；Liu、Zhu 和 Zhang，2022年）。封建强化学习（Dayan 和 Hinton，1992）（见上一节）中的低层策略是 GCRL 的早期示例之一，但并非所有 GCRL 方法都包含分层结构。目标空间通常被具体化为状态空间的一个子集（智能体只有到达与当前目标对应的特定状态时才会获得奖励），但也可以用自然语言表示（Chan 等，2019；Colas 等，2020），或表示为期望回合回报的等级（Chen 等，2021；McCallum 等，2023）。虽然大多数 GCRL 方法都涉及预先指定目标，但近期也有一些研究探索了智能体学习生成自身目标的可能性（Colas 等，2022年）。

与我们的方法更直接相关的是将目标条件策略融入模仿学习的研究。例如，Lynch 等人（2019）结合自监督表征学习和监督行为学习，从非结构化玩耍数据中获得目标条件机器人策略。Ghosh 等人（2020）采用目标条件行为克隆进行自我模仿，利用智能体先前的轨迹作为监督学习中的成功示例，针对的是智能体实际到达的状态（无论这些状态是否是智能体“意图”到达的状态）。其他研究则探索了基于 GAIL 算法变体的目标条件模仿学习（Ding 等人，2020）、使用扩散策略的方法（Reuss 等人，2023），或采用 Transformer 架构的方法（Sundaresan 等人，2025）。

3 问题设定

我们考虑这样一个问题：自我智能体必须在一个由其他智能体组成的环境中采取行动，以实现目标

（为与 AHT文献保持一致，我们将使用队友这一术语），这些队友可能具有不同的目标。请注意，尽管我们呈现的实验结果仅使用了一个队友，但为了一般性起见，我们在此阐述任意 n名队友的情况。

形式上，我们在部分可观察随机博弈（POSG）的框架下定义问题。一个 POSG 由以下元组定义：$\langle \mathcal{N},\ \mathcal{S},\ \mathcal{T},\ \{\mathcal{A}^i\},\ \{\mathcal{O}^i\},\ \{\mathcal{Z}^i\},\ \{r^i\},\ \gamma\rangle$其中，$\mathcal{N}$ 是智能体集合（其中 $i \in \mathcal{N}$），$\mathcal{S}$ 是状态空间，而 $\mathcal{T}$: $\mathcal{S} \times \vec{\mathcal{A}} \mapsto \Delta(\mathcal{S})$ 是转移函数，其中 $\mathcal{A} = \mathcal{A}_1 \times ... \times A_N$ 表示联合动作空间。对于每个智能体 i，其动作空间为 $\mathcal{A}^i$，观测空间为 $\mathcal{O}^i$，观测函数为 $\mathcal{Z}^i: S \mapsto \mathcal{O}^i$，奖励函数为 $r^i: S \times \vec{\mathcal{A}} \mapsto \mathbb{R}$。$\gamma \in [0,1]$ 是折扣因子。

通常，我们考虑这样一种环境：它提供一组$\mathcal{G}$不同的可能目标；其中一些目标可由单个智能体独自行动实现，另一些则需要两个或更多智能体合作才能实现。形式上，目标 g可定义为一个元组 (Ω_g, r_g)，其中 $\Omega_g \subseteq \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ 是单步转移空间的一个特定子集，而 r_g 是一个标量奖励值。例如，在我们的觅食环境中（见第 5 节），目标“收集苹果”对应于 $\Omega_g = \{(s, a, s')$: 相邻的苹果$(s) \wedge a =$收集 $\wedge$ 苹果已收集$(s')\}$。随后，每个智能体的目标由 $\mathcal{G}$上的一个二元掩码刻画——也就是说，对于每个目标 $\bar{g} \in \mathcal{G}$，智能体 i 执行转移 $\in \Omega_g$时，将获得 r_g或 0 的奖励。因此，我们可以用不同智能体能获得非零奖励的目标子集 $\mathcal{G}^i \subseteq \mathcal{G}$ 来表示奖励上的所有个体差异。例如，在我们的觅食示例中，可能有 $\mathcal{G}$ = {收集苹果、收集橙子、收集李子}，以及 $\mathcal{G}^i$ = {收集苹果}。我们还假设，需要合作的目标总是比不需要合作的目标奖励更高，即 $r_g > r_{g'}\ \forall\, g \notin \mathcal{G}^{\text{solo}}, g' \in \mathcal{G}^{\text{solo}}$——这旨在反映现实中人们通常合作时比独自行动能取得更多成果。环境中的每个智能体都会发出一个可观察线索 ϕ^i，以带有噪声的方式传达其目标，即 $\phi^i \sim \mathcal{N}(\mathcal{G}^i, \sigma^2 I)$。除非另有说明，我们将在实验中始终使用 $\sigma^2 = .05$。

我们可以从对自我智能体而言值得追求的目标（即既有奖励、又有可能实现）这一视角，来思考该设定中的理性行为。此类目标的集合可以写为 $\mathcal{G}^{\text{ego}} \cap \left(\mathcal{G}^{\text{teammates}} \cup \mathcal{G}^{\text{solo}}\right)$ $\mathcal{G}^{\text{ego}}$ $\mathcal{G}^\star =$，其中是自我智能体的目标子集，$\mathcal{G}^{\text{teammates}}$ 是所有队友目标子集的并集，$\mathcal{G}^{\text{solo}}$ 是不需要合作的目标子集。图1展示了 $\mathcal{G}^\star$（以绿色高亮）如何在由这一一般设定产生的三种基本场景中发生变化：

1. **完全重叠**（自我智能体的所有目标都至少与一位队友共享）：$\mathcal{G}^{\text{ego}} \subseteq \mathcal{G}^{\text{teammates}}$2. **部分重叠**（自我智能体的目标中至少有一个、但并非全部，至少与一位队友共享）：$\mathcal{G}^{\text{ego}} \cap \mathcal{G}^{\text{teammates}} \neq \emptyset,\ \ \mathcal{G}^{\text{ego}} \not\subset \mathcal{G}^{\text{teammates}}$3. **无重叠**（自我智能体没有任何目标与任何队友共享）：$\mathcal{G}^{\text{ego}} \cap \mathcal{G}^{\text{teammates}} = \emptyset$

要成功完成整体任务，自我智能体应能够应对**所有**这些场景——判断何时追求协作目标，何时独自行动。

4 方法

我们提出两种方法：GRILL和GRILL-M，后者是一种变体，纳入了来自LIAM（Papoudakis、Christianos和Albrecht，2021）的辅助队友建模组件的一个版本。GRILL的核心思想是将以下问题分开：**(1)** 学习在给定特定环境状态下应追求哪些目标，以及**(2)** 学习为实现这些目标应采取哪些动作。也就是说，我们不是学习单一的端到端策略，而是分别学习用于目标选择的高层策略，以及用于目标条件动作选择的低层策略。这种层次结构并不新颖（见第2节）——相反，我们的洞见是，最优低层策略对群体中的所有智能体是通用的，而最优高层策略则取决于自我智能体及其当前队友的目标。因此，我们可以采用一个结合模仿与强化学习的两阶段过程（如图2所示）：

阶段 1：我们首先从随机采样的启发式智能体收集一个规模较小的离线数据集 $\mathcal{D}$ $=\{o_t, a_t\}$，其中包含观测和动作。随后将 $\mathcal{D}$ 划分为一组固定长度的轨迹 $\{\tau\}$，每条轨迹对应于追求单一目标的过程；然后利用这些轨迹优化一个编码器-解码器模型。对于每条轨迹，编码器都会生成一个离散目标标签。给定离散编码后，一个解码器会尝试根据此前的观测，预测轨迹中每个时刻队友的动作；另一个解码器则尝试根据第一个观测预测最终观测。整个系统仅通过重建 $\mathcal{D}$ 中的动作和观测来优化，不提供明确的目标信息：

$$\mathcal{L}_a = -\sum_{t=0}^{T-1} \log \text{dec1}_{\theta_{\text{dec1}}}\left(a_t \mid o_t, \hat{g} = \text{enc}_{\theta_{\text{enc}}}(\tau)\right)$$

$$\mathcal{L}_o = \left\| o_{T-1} - \text{dec2}_{\theta_{\text{dec2}}}\left(o_0, \hat{g} = \text{enc}_{\theta_{\text{enc}}}(\tau)\right)\right\|_2^2$$

$$\min_{\theta_{\text{enc}}, \theta_{\text{dec1}}, \theta_{\text{dec2}}} \mathbb{E}_{\tau \sim \mathcal{D}}\left[\lambda \mathcal{L}_a + (1-\lambda)\mathcal{L}_o\right] \quad (1)$$

其中，λ 是一个标量，用于权衡两个预测目标。训练完成后，我们舍弃编码器和观测解码器，同时保留动作解码器，将其作为低层目标条件策略 π_{action}。

阶段 2：我们使用近端策略优化来学习一个高层策略 π_{goal}，将当前观测映射为一个离散目标，并将其输出用于对上一阶段的行为克隆策略进行条件化。对于 GRILL-M，此阶段还包含一个辅助目标：通过具有以下目标函数的 LSTM 编码器-解码器，训练自我智能体根据自身的观测和动作预测队友的动作

$$\min_{\theta_{\text{enc}}, \theta_{\text{dec}}} -\log \text{dec}_{\theta_{\text{dec}}}\left(a_t^{-1} | z_t = \text{enc}_{\theta_{\text{enc}}}(o_t, a_{t-1})\right) \quad (2)$$

其中，π_{goal} 是在增广空间 $\mathcal{O} \times \mathcal{Z}$上学习得到的。这与 LIAM 的建模部分几乎完全相同

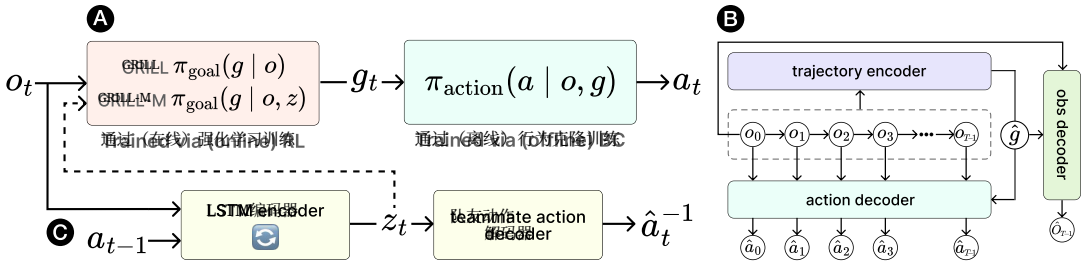

图2：（**A**）GRILL的分层架构（**B**）第1阶段离线优化的编码器-解码器架构，其中动作解码器成为第2阶段的低层策略 π_{action}（**C**）GRILL-M中使用（但GRILL中不使用）的辅助建模组件

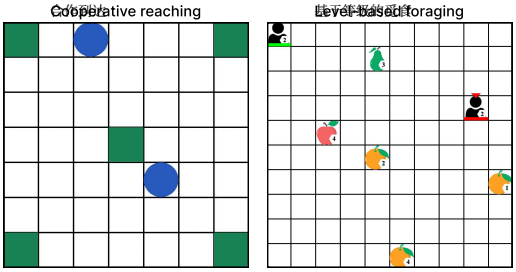

图3：我们扩展的两个临时团队协作环境中的示例帧。

（Papoudakis, Christianos 和 Albrecht 2021），唯一的区别是，由于我们的环境完全可观测，因此我们去掉了对队友观测的额外预测。

5 实验

环境

我们扩展了两种常用的、完全可观测的临时团队协作环境，以纳入目标异质性的概念。图 3 展示了这两种环境的状态示例。

合作到达：合作到达是一种简单的网格世界任务，两个智能体必须移动到四个产生奖励的角落格之一，并共同占据该格（Rahman et al. 2023b；Rahman、Cui 和 Stone 2024）。在原始版本中，不同角落对应不同数量的奖励，但到达某个角落所获的奖励在不同智能体之间没有差异。在我们的版本中，每个角落格对不同智能体的奖励要么为 1，要么为 0；我们还在网格中央增加了一个额外的目标格，单个智能体即可成功到达该格（即处于 $\mathcal{G}^{\text{solo}}$），但只能获得较低的 0.2 奖励。所有目标格都是吸收状态——智能体到达这五个格子中的任意一个后，就无法再执行任何动作。自我智能体接收的观测包含自身位置，以及队友的位置和可观察线索（ϕ^i）。动作空间仅由四个基本方向和一个无操作动作组成。

基于等级的觅食：基于等级的觅食（LBF）是另一种网格世界环境，智能体在其中协作收集物品。每个智能体和物品都有一个对应的“等级”，等级为 l 的物品只能由等级总和达到 $\geq l$ 的一组智能体协同努力才能收集（Albrecht 和 Ramamoorthy，2013；Yourdshahi 等，2018；Rahman 等，2023a）。我们的环境版本在此基础上进行了扩展：我们不再只设置一种物品类型，而是设置了三种不同的“水果”（它们的等级也可以不同）。因此，环境中可能出现的目标集合为 $\mathcal{G} = \{$（类型 l）$\forall$类型 $\in \{$苹果、橙子、李子$\}$、 $l \in [1, l_{\max}]\}$ 其中 $l_{\max}$ 是物品最高等级。自我智能体接收的观测包含环境中所有水果的位置、类型和等级，以及其自身的位置和等级，还有队友的位置、等级和可观察线索（ϕ^i）。动作空间包括四个基本方向、“收集”动作（尝试收集水果）和无操作动作。为简化起见，我们在实验中将自我智能体和队友的等级都固定为 1。

基线

我们将GRILL与三种基线进行比较：近端策略优化、LIAM和OMG。近端策略优化（Schulman等人，2017）是一种通用的强化学习算法；就我们的目的而言，它为完全未针对合作（乃至多智能体）设定做任何定制的方法所能达到的水平提供了下限。LIAM（Papoudakis、Christianos和Albrecht，2021）与OMG（Yu、Jiang和Lu，2024）是两种近期方法，它们利用某种形式的智能体建模，在与本文所用任务相似的AHT任务上取得了有竞争力的表现。LIAM和OMG都在一个“增强”观测空间上训练自我智能体的策略，该空间纳入了由辅助网络产生的潜在队友表示。在LIAM中，该表示由循环编码器-解码器产生，其训练目标是根据自我智能体自身的（观测，动作）对来预测队友当前的观测与动作。OMG则使用来自条件变分自编码器的潜在表示，该模型被优化用于建模队友的“子目标”（即某个未来状态的特征嵌入）。我们注意到，尽管LIAM的作者使用了优势演员-评论家，且OMG针对不同任务使用了不同的强化学习算法，但为了各方法之间保持一致，我们对二者的实现均以近端策略优化为骨干。对于每个环境，我们还将与一种“Oracle”策略进行比较，该策略采用由对所有智能体的目标及基础奖励的完全知晓所引导的最短路径启发式。除Oracle之外，所有方法均在合作到达上训练了 10^7 时间步，在基于等级的觅食上训练了 5×10^7 时间步。

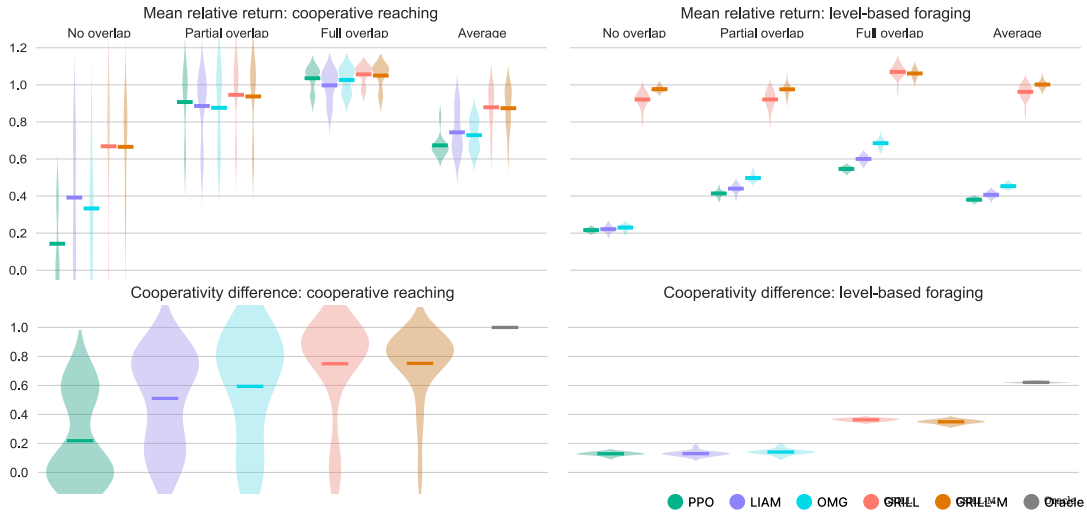

图 4：**上方：**相对于最优策略的评估回报，在 1000 个回合中测得 × 这 3 种场景下 × 20 次独立训练运行的结果。水平标记表示均值，下方的小提琴图显示核密度估计。智能体分别在 CR 和 LBF 上训练了 1e7 和 5e7 个时间步。**下方：**在同一组评估回合中，完整场景与无重叠场景下所追求的非单人目标比例之差。

评估

除了平均回合回报这一标准评估指标之外，我们还希望更深入地考察训练策略如何应对第3节所列举的三种场景。为此，我们考察自我智能体在评估回合中尝试到达哪些目标——即在CR中它会导航到哪些吸收格，以及在LBF中它会尝试收集哪些水果。我们考虑与目标选择相关的三种不同失败模式。第一种（也是最基本的）是追求 $\mathcal{G}^{\text{ego}}$之外的目标，即即使达成也不会产生任何奖励的目标。第二种是过度协作，即追求位于 $\mathcal{G}^{\text{ego}}$ 但无法实现的目标（$\notin \mathcal{G}^{\text{solo}} \cup \mathcal{G}^{\text{teammates}}$；请注意，该失败模式在完全重叠场景中并不存在）。第三种也是最微妙的是协作不足，即当存在可实现的合作目标时却不去追求这些目标（请注意，该失败模式在无重叠场景中并不存在）。成功的智能体应避免所有这三种失败模式。

我们测量自我智能体在四个不同子集中的目标选择分布，以便区分这三种失败模式（见图 1 和图 5）。作为额外的汇总指标，我们还报告“合作性差异”(Δ_{coop})，即完整场景与无重叠场景下所追求目标的比例差异 $\notin \mathcal{G}^{\text{solo}}$；该值越高，表示自我智能体的策略越灵活。

6 结果

GRILL 取得了比所有基线更高的回报

图 4（顶行）展示了训练后的平均回报（相对于每个环境的最优策略），涵盖第 3 节列举的三种场景。总体而言，我们发现，在两个环境的每一种场景中，GRILL 和 GRILL-M 的表现都优于所有基线。

仔细观察合作到达的结果（左图）可以看出，在部分重叠和完全重叠场景中，所有方法的表现都相近，所获得的回报接近甚至略高于最优策略。在无重叠场景中，所有方法（包括我们的方法）的表现都明显较差——有趣的是，这是 PPO 与另外两个基线之间差异最大的场景。在合作到达的所有场景中，我们没有发现 GRILL 和 GRILL-M 之间存在明显差异。在难度更高的 LBF 环境中（右图），我们观察到，在所有场景下，GRILL 的表现都明显优于三个基线。基线方法中，OMG 在 LBF 上始终优于 LIAM 和 PPO，不过优势并不算大。有趣的是，在这里我们确实看到了两种方法变体之间的差异：在无重叠和部分重叠场景中，GRILL-M 略占优势。

GRILL 追求更有价值的目标

图 5 展示了在两种环境中，训练后的自我智能体在评估期间追求的目标分布——我们将目标空间划分为四个子集，以揭示第 5 节概述的三种失败模式。在合作到达中，所有智能体都避免了第一种失败模式，即只追求实现后会带来奖励的目标。在无重叠场景中，所有智能体都在不同程度上陷入了第二种失败模式：尽管追求合作目标徒劳无益，它们仍然这样做。PPO 智能体的这一问题尤其明显，它选择正确格子的比例不到 30%。在部分重叠和完全重叠场景中，四种方法的目标分布非常相似，智能体相对于最优策略表现得不够合作。再看 LBF，我们发现 PPO、LIAM 和 OMG 在所有三种场景中都表现出第一种失败模式；此外，在无重叠和部分重叠场景中还表现出第二种失败模式，在第三种失败

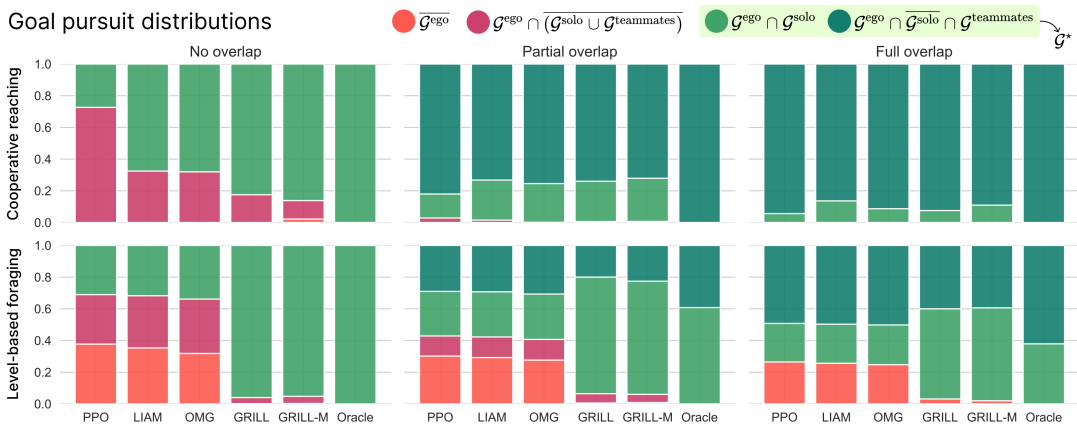

图 5：评估期间自我智能体尝试的目标分布。对于 CR，“尝试”是指智能体占据其中一个目标格；对于 LBF，则是指智能体在水果旁边时执行“收集”动作。

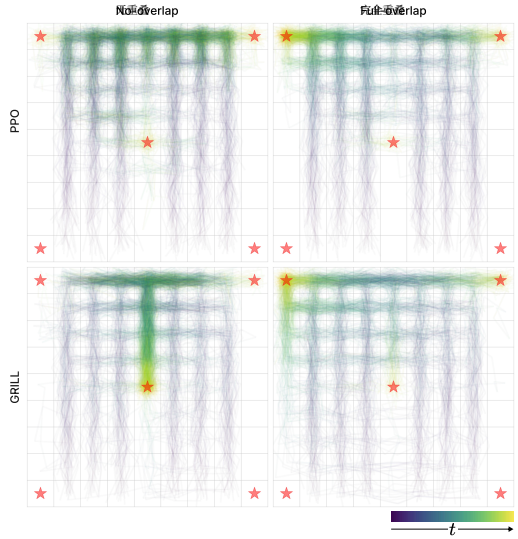

图 6：在合作抵达环境的无重叠和完全重叠场景中，从训练后的近端策略优化和 GRILL 智能体采样得到的轨迹可视化。轨迹从随机初始位置采样，并对每条路径施加少量噪声，以便更清晰地显示分布。

模式则出现在部分重叠和完全重叠场景中。相比之下，GRILL 完全避免了第一种失败模式，几乎完全避免了第二种失败模式，在所有三种场景中，超过 90% 的时间都追求有价值的目标。不过，在部分重叠和完全重叠设置下，它仍然不够合作；而且在这一特定指标上，表现实际上略逊于三个基线。总体而言，在两种环境中，使用我们的方法训练的智能体选择有价值目标的比例都高于任何基线方法。为直观展示这一结果，图 6 对比了 PPO 和 GRILL 在合作到达中的轨迹，凸显出后者的智能体在无重叠和完全重叠场景中的最终位置存在明显更大的差异。

图 4 最后一行展示了每种方法的“合作性差异” Δ_{coop}，即自我智能体在完全重叠场景与无重叠场景中追求合作目标的程度之差。与性能结果一致，我们发现，在两种环境中，GRILL 的 Δ_{coop} 值都高于基线；而在合作到达任务中，LIAM 和 OMG 的 Δ_{coop} 值仅高于近端策略优化。

当队友目标信息噪声更大时，GRILL-M优于 GRILL

最后，我们进行了一项小型补充实验，试图理解在我们方法的两种变体之间所观察到的结果模式。我们的假设是，在协作到达任务中，在观测中获取有关队友目标的信息对性能而言不太重要，因为队友的目标能更清楚地从其行为中体现出来。在LBF中，更大的可能行为空间使得目标与动作之间的关系不再那么简单，因此在观测中包含这些信息应当更有用。如果潜在表征能够捕获关于队友目标的信息，那么它们对性能的助益程度，应取决于该信息与“基础”观测不冗余的程度。

为了检验这一点，我们在 LBF环境中对 GRILL 和 GRILL-M 进行了训练和评估，并逐步提高队友可观测特征 ϕ的噪声水平，从 0.05（此前实验中使用的

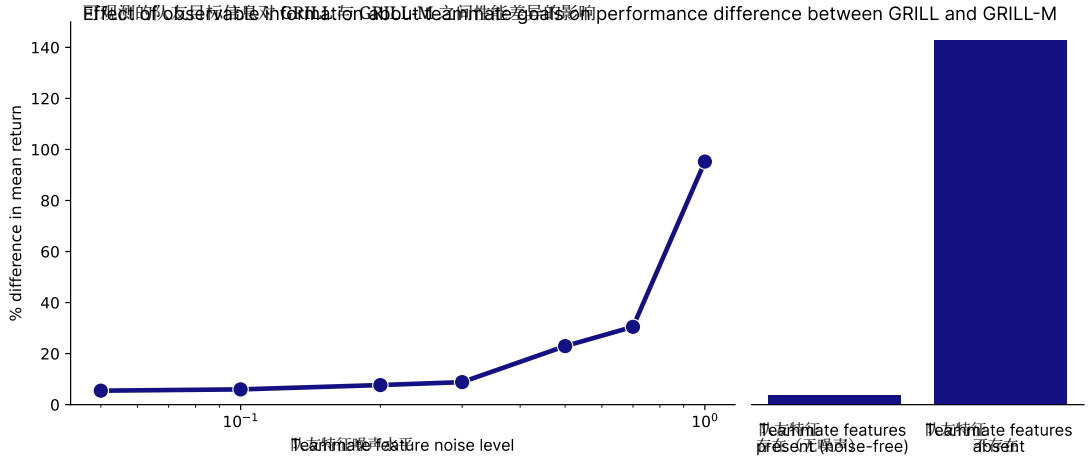

图7：**左：**在LBF环境中，GRILL-M相对于GRILL的平均评估回报百分比增幅，作为自我智能体对其队友目标向量观测中噪声量的函数。回报在1000个回合× 3个场景 × 20 独立训练运行上取平均。**右：**在无噪声与队友目标信息完全缺失这两种极端情况下，对同一增幅进行比较。

数值）提高到 1.0。我们还进行了一个更极端的对比实验，完全从观测中移除了 ϕ 。图 7 显示，随着噪声增加，性能差距单调扩大——当 σ^2 =1.0 时，GRILL-M 的平均回报高出 95.2%，而在 $\sigma^2 = 0.05$时仅高出 5.5%。当 ϕ 完全从观测空间中移除时，这一差距进一步扩大到 142.6%。这些结果支持了我们的假设：辅助潜在表示在某些情况下是有益的。

7 讨论

在本文中，我们提出了标准临时团队协作问题的一种变体：智能体可以拥有异质目标，这些目标在任一给定回合中可能重叠，也可能不重叠。我们研究这一设置的动机，是捕捉这样一种直觉：在许多现实场景中——并非所有智能体都拥有共同目标——人们通过理解协作何时可能带来成效，从而取得更好的结果。这与临时团队协作设置中的大多数工作形成对比；后者通常处理的问题是，如何适应那些在追求同一目标时行为可能有所不同的队友。

作为一项次要贡献，我们提出了 GRILL，这是一种新颖的分层方法，用于在异质目标场景中学习灵活策略。该方法通过在动作选择层进行模仿，并在目标选择层进行强化学习来实现这一点。借助对两种常见 AHT环境——合作到达和基于等级的觅食——的扩展版本，我们证明，GRILL 凭借更灵活的目标选择，优于基线方法。在一项小规模补充实验中，我们发现，辅助预测目标（类似 LIAM）的效果取决于观测空间中是否提供队友目标信息。

尽管我们在此着重于刻画灵活合作的某个特定维度，但我们认为，其核心理念应当广泛适用于任何多目标多智能体设置，无论其中是否涉及协作。例如，在某些明确具有竞争性的领域中，也可能存在类似的决策问题：智能体必须选择是追求竞争激烈但价值较高的目标，还是转而选择价值较低但竞争较小的目标。我们计划在未来的研究中检验这一想法，并探索其与人类行为学习的联系。后续研究还可以解决当前工作的局限性，例如，我们的实验仅限于完全可观测且只有两个智能体的环境，或者低层策略是在前期阶段离线学习的，而非与高层策略同步在线学习。

参考文献

Albrecht, S. V.；Crandall, J. W.；和 Ramamoorthy, S. 2016。假设行为中的信念与真相。人工智能，235：63–94。

Albrecht, S. V.；和 Ramamoorthy, S. 2013。多智能体系统中临时协调的博弈论模型与最佳响应学习方法。载于会议论文集 *o*

f

*2013*年自主智能体与多智能体系统国际会议，AAMAS'13，1155–1156。

Bacon, P.-L.；Harb, J.；以及 Precup, D. 2017。选项评论家架构。载于第三十一届*AAAI*人工智能会议论文集，AAAI'17，1726–1734。AAAI出版社。

Barreto, A.；Borsa, D.；Hou, S.；Comanici, G.；Ayg¨un, E.；Hamel, P.；Toyama, D.；Mourad, S.；Silver, D.；以及 Precup, D. 2019。选项键盘：结合强化学习中的技能。神经信息处理系统进展，32。

Barrett, S.；Stone, P.；以及 Kraus, S. 2011。追捕领域中临时团队协作的实证评估。载于第 *10* 届

自主智能体与多智能体系统国际会议——第 *2* 卷，AAMAS '11，567–574。自主智能体与多智能体系统国际基金会。

Chan, H.; Wu, Y.; Kiros, J.; Fidler, S.; and Ba, J. 2019. AC-TRCE：通过教师建议增强经验以实现多目标强化学习。arXiv:1902.04546。

Chen, L.; Lu, K.; Rajeswaran, A.; Lee, K.; Grover, A.; Laskin, M.; Abbeel, P.; Srinivas, A.; and Mordatch, I. 2021. 决策 Transformer：通过序列建模实现强化学习。In神经信息处理系统。

Colas, C.; Karch, T.; Lair, N.; Dussoux, J.-M.; Moulin-Frier, C.; Dominey, P.; and Oudeyer, P.-Y. 2020. 语言作为认知工具，在好奇心驱动的探索中构想目标。In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., 神经信息处理系统进展，第 33 卷，3761–3774。

Colas, C.; Karch, T.; Sigaud, O.; and Oudeyer, P.-Y. 2022年. 具有内在动机的目标条件强化学习自我目的智能体：简短综述。人工智能研究杂志，74。

Dayan, P.; and Hinton, G. E. 1992. 封建式强化学习。载于 Hanson, S.; Cowan, J.; and Giles, C. 编，神经信息处理系统进展，第5卷。Morgan-Kaufmann。

Ding, Y.; Florensa, C.; Phielipp, M.; and Abbeel,P. 2020. 目标条件模仿学习。ArXiv:1906.05838[cs]。

Ghosh, D.; Gupta, A.; Reddy, A.; Fu, J.; Devin, C.; Eysenbach, B.; and Levine, S. 2020. 通过迭代监督学习掌握达成目标。ArXiv:1912.06088 [cs]。

Gmytrasiewicz, P. J.；和 Doshi, P. 2005。《多智能体环境中的序贯规划框架》。人工智能研究杂志，24(1)：49–79。Henrich, J. 2018。《我们成功的秘密：文化如何推动人类进化、驯化我们的物种并让我们变得更聪明》。普林斯顿大学出版社。

Kaelbling, L. P. 1993。学习实现目标。载于国际人工智能联合会议。

Klissarov, M.；和 Precup, D. 2021。灵活的选项学习。载于《神经信息处理系统进展》，第34卷，4632–4646。Liemhetcharat, S.；和 Veloso, M. 2014。异构临时智能体的有效团队组建：加权协同图。人工智能，208：41–65。

Liu, M.；Zhu, M.；和 Zhang, W. 2022年。目标条件强化学习：问题与解决方案。载于 Raedt, L.D. 编，《第三十一届国际人工智能联合会议论文集，*IJCAI-22*》，5502–5511。国际人工智能联合会议组织。

Liu, X.; Li, P.; Yang, W.; Guo, D.; 和 Liu, H. 2024. 利用大语言模型进行异构临时团队合作。arXiv:2406.12224.

Lynch, C.; Khansari, M.; Xiao, T.; Kumar, V.; Tompson, J.; Levine, S.; 和 Sermanet, P. 2019. 从玩耍中学习潜在计划。ArXiv:1903.01973 [cs].

McCallum, S.; Taylor-Davies, M.; Albrecht, S.; and Suglia, A. 2023. 反馈就是你所需要的一切吗？利用自然语言反馈进行目标条件强化学习。发表于*NeurIPS 2023* 目标条件强化学习研讨会。

Mirsky, R.; Carlucho, I.; Rahman, A.; Fosong, E.; Macke, W.; Sridharan, M.; Stone, P.; and Albrecht, S. V. 2022年。《临时团队协作研究综述》。ArXiv:2202.10450[cs]。

Papoudakis, G.; and Albrecht, S. V. 2020. 用于多智能体系统对手建模的变分自编码器。*arXiv*预印本 *arXiv:2001.10829*。

Papoudakis, G.; Christianos, F.; and Albrecht, S. V. 2021. 部分可观测条件下的深度强化学习智能体建模。发表于第*35*届神经信息处理系统国际会议论文集，NIPS '21。Rabinowitz, N.; Perbet, F.; Song, F.; Zhang, C.; Eslami, S. M. A.; and Botvinick, M. 2018. 机器心智理论。发表于 Dy, J.; and Krause, A. 编，《第*35*届机器学习国际会议论文集》，机器学习研究论文集第80卷，4218–4227。PMLR。

˜Rahman, A.; Carlucho, I.; HA¶pner，N.; 和 Albrecht, S. V. 2023a. 使用基于图的策略学习实现开放式临时团队协作的通用学习框架。机器学习研究期刊,24(298): 1–74.

Rahman, A.; Fosong, E.; Carlucho, I.; 和 Albrecht, S. V. 2023b. 通过最佳响应多样性生成队友，以训练鲁棒的临时团队协作智能体。arXiv:2207.14138. Rahman, M.; Cui，J.; 和 Stone, P. 2024. 用于训练鲁棒临时团队协作智能体的最小覆盖集。载于第三十八届*AAAI*人工智能会议、第三十六届人工智能创新应用会议及第十四届人工智能教育进展研讨会论文集，AAAI'24/IAAI'24/EAAI'24。AAAI出版社。

Reuss，M.; Li，M.; Jia，X.; 与 Lioutikov，R. 2023. 利用基于分数的扩散策略进行目标条件模仿学习。arXiv:2304.02532 [cs]。

Schulman，J.；Wolski，F.；Dhariwal，P.；Radford，A.；以及 Klimov，O. 2017。《近端策略优化算法》。arXiv:1707.06347。Sundaresan，P.；Vuong，Q.；Gu，J.；Xu，P.；Xiao，T.；Kirmani，S.；Yu，T.；Stark，M.；Jain，A.；Hausman，K.；Sadigh，D.；Bohg，J.；以及 Schaal，S. 2025。《RT-Sketch：基于手绘草图的目标条件模仿学习》。载于《机器人学习》第八届会议论文集，70–96。PMLR。Sutton，R. S.；Precup，D.；以及 Singh，S. 1999。《MDP 与半马尔可夫决策过程之间：强化学习中的时间抽象框架》。人工智能，112(1)：181–211。

Tomasello，M.；Melis，A. P.；Tennie，C.；Wyman，E.；以及 Herrmann，E. 2012。《人类合作演化中的两个关键步骤：相互依存假说》。《当代人类学》，53(6)：673–692。Yourdshahi，E. S.；Pinder，T.；Dhawan，G.；Marcolino，L. S.；以及 Angelov，P. 2018。《迈向大规模临时团队合作》。载于*2018 IEEE* 国际智能体会议（*ICA*），44–49。Yu，X.；Jiang，J.；以及 Lu，Z. 2024。《基于子目标推断的对手建模》。载于 Globerson，A.；Mackey，L.；Belgrave，D.；Fan，A.；Paquet，U.；Tomczak，J.；以及 Zhang，C. 编，《神经信息处理系统进展》，第 37 卷，60531–60555。

A 架构细节

所有策略和价值网络（包括 GRILL 和 GRILL-M 的低层与高层网络）均使用大小为 (64, 64) 的隐藏层。其他维度如下：

方法	描述	值
LIAM	编码器隐藏层	(64, 64)（LSTM，全连接层）
LIAM	编码器输出	10
LIAM	解码器隐状态	(64, 64)
OMG	多层感知机隐层	128
OMG	多层感知机输出	32
OMG	循环神经网络隐藏层	(128, 128)
OMG	循环神经网络输出	96
OMG	变分自编码器隐藏层	(128, 128)
OMG	变分自编码器输出	16
OMG	条件变分自编码器隐藏层	(128, 128)
OMG	条件变分自编码器输出	16
GRILL-M	编码器隐藏层	(64, 64) (LSTM，全连接层)
GRILL-M	编码器输出	10
GRILL-M	解码器隐藏层	(64, 64)

所有策略和价值网络均使用“swish”激活函数；所有其他网络均使用 ReLU。

B 超参数

所有方法（近端策略优化、LIAM、OMG、GRILL、GRILL-M）共用的近端策略优化骨干网络采用了以下超参数：

超参数	值
批量大小	1024
小批量	8
训练轮数	2
最大梯度范数	0.5
学习率	1e-4
折扣因子 (γ)	0.99
GAE λ	0.95
裁剪参数 (ϵ)	0.2
熵系数	0.01
价值损失系数	0.5
并行环境数量	16

各方法特有的其他超参数如下：

方法	超参数	值
LIAM	编码器-解码器学习率	1e-3
OMG	循环神经网络、变分自编码器、条件变分自编码器学习率	1e-3
GRILL-M	编码器-解码器学习率	1e-3

C 训练细节

训练运行在单个英伟达 RT2080Ti 图形处理器上执行。智能体在合作到达任务上训练了 1e7 个时间步，在基于等级的觅食任务上训练了 5e7 个时间步。我们注意到，为了使“经历过的”时间步总数相等，我们让 GRILL 和 GRILL-M 分别在线训练 8e6 和 4.8e7 个时间步，因为这两种方法在离线模仿阶段都使用了包含 2e6 个时间步的数据集。每种方法/环境组合均使用 20 个独立种子（0–19）重复训练（总计 200 次训练运行）。对于三个重叠场景中的每一个场景，对每个训练得到的策略进行 1000 个回合的评估（使用不同的种子）。