arXiv:2611.32031v1 [cs.MA] 24 Nov 2026

# 统一合作伙伴与环境多样性，提升人机协作

**Patrick Q. Zhang**pqz317@uw.edu

**Yancheng Liang**yancheng@cs.washington.edu

**Adrienne F airhall**fairhall@uw.edu

**Natasha Jaques**
nj@cs.washington.edu

## 摘要

现实世界中现代人工智能系统的应用，要求其能够泛化到未见过的环境和新的人类用户。零样本协作（ZSC）领域主要关注如何部署人工智能，使其能够在其他方面熟悉的环境中与新伙伴协作。流行的基于群体的方法通过在固定环境中训练时，让智能体与多样化的模拟人类用户合作来实现这一目标，但并未考虑更现实的情况：用户和环境在训练期间都未曾出现。近期，研究者开始探索通过跨环境训练单个自博弈智能体，结果表明，这种方法在面对新的人类用户和未见过的环境时都能实现更好的泛化。在本研究中，我们探讨同时对伙伴多样性和环境多样性建模，能否在新环境中与新伙伴实现更强的协调。我们提出跨环境协同博弈（CECP），这是一种两阶段框架：首先，在程序生成环境的分布上训练一群自博弈智能体；随后，针对这一伙伴群体训练一个协作者策略。为进一步结合社会推理与环境推理，我们加入了一个下一状态预测模块，促使表征同时编码伙伴行为和任务结构。我们在神经科学启发的合作觅食任务以及广受欢迎的 ZSC 基准 Overcooked 上，分别在留出合作伙伴和留出环境的条件下评估 CECP。与所比较的基线方法相比，CECP取得了最优表现；按照我们针对人机协作的评估协议，其结果也达到了当前最先进水平。我们分析了由下一状态预测模块编码的模型潜在表征，结果表明，该模块不仅能提升协作型人工智能的表现，还能提供更易解释的潜在表征，例如可用于推断合作伙伴的技能水平。[1]

## 1 引言

人工智能持续取得极其迅速的进展，尤其是在数字领域。然而，当前人工智能的发展与我们许多人所设想的理想化未来大相径庭。我们或许已经解决了纳维–斯托克斯问题，但真正能够帮助处理家务、减轻体力劳动的家用机器人仍然
劳动仍然只是科幻小说中的内容；我们仍然得自己洗碗！要实现物理辅助人工智能这一愿景，我们需要一种能够在许多不同家庭中，针对广泛任务协助许多不同人们的人工智能智能体；这是人类可以毫不费力地做到的能力，但人工智能智能体仍难以掌握。我们将对环境和新接触的人类同时进行适应，定义为社会情境转变。

[1]请参阅我们的网站：https://pqz317.github.io/cecp-project-site/以及训练代码：https://github.com/pqz317/cecp。

第40届神经信息处理系统大会（NeurIPS 2026）。

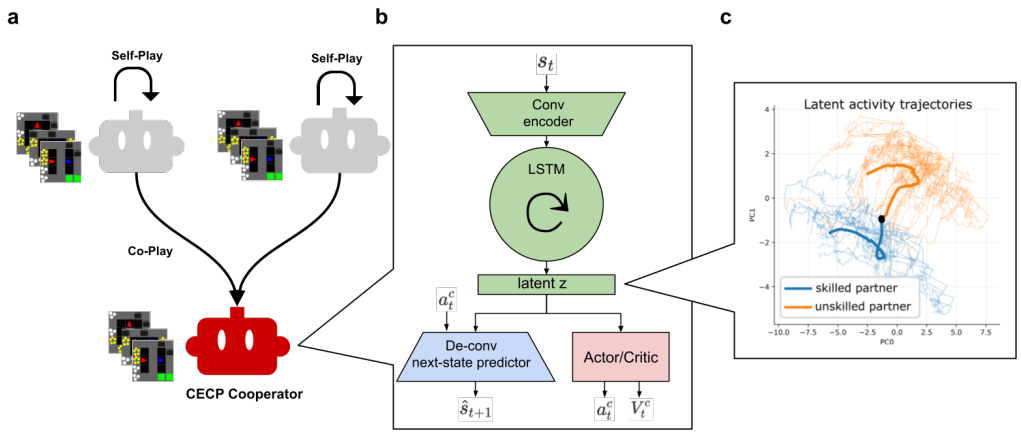


图 1：CECP 概览。**a**，训练范式。首先，在环境分布上通过自博弈训练合作伙伴智能体。随后，CECP 协作者通过与冻结的伙伴群体合作进行训练，同时也在环境分布上进行训练。**b,** 合作智能体的架构。CECP 配备了下一状态预测目标（蓝色）。**c,** 合作智能体中的潜在表示。CECP 潜变量形成对伙伴行为的表示，例如推断出的伙伴技能水平。

零样本协作（ZSC） [3]为研究人工智能在多样化人类用户之间的泛化提供了一个通用框架。在典型的ZSC基准中，例如合作烹饪游戏 Overcooked，智能体会与未见过的人类伙伴配对 [24, 30, 13, 4, 27]，并在用于训练的同一环境中进行协作。目前，已有700多篇论文使用 Overcooked 研究ZSC，其中几乎所有论文都只专注于覆盖可能的伙伴空间，却忽视了这样一个维度：能否在新环境中与新伙伴即时合作，而该环境可能还包含新的任务或目标。只有少数论文 [11, 20], 关注在ZSC背景下跨多个不同环境训练智能体这一理念。耐人寻味的是，Jha 等人 [11] 发现，仅在多样化环境中进行训练，就能实现对人类伙伴的有效泛化，即使采用的是自博弈。然而，迄今为止，还没有人成功探索将伙伴多样性与环境多样性结合起来。我们假设，能够应对这种高度任务复杂性和策略复杂性的算法，将是人工智能最有益应用所必需的，例如具身协作助手。

在本研究中，我们探讨同时发生社会情境转变下的协作泛化。我们提出跨环境协作（CECP），这是一种旨在统一这两个轴向的元强化学习方法。CECP 的训练分为两个阶段（图 1a）。训练环境和合作伙伴设置：通过在程序化生成的环境分布上进行自博弈，训练智能体群体，从而产生一组多样化的策略（合作伙伴池），这些策略建立在多样化的动态基础上。元强化学习训练阶段：针对从合作伙伴池中采样的冻结群体训练协作者策略，使其接触异质的合作伙伴行为，同时保留广泛的环境变化。评估：将在未见过的环境中，针对新伙伴（例如人类玩家）评估协作者。

同时沿着伙伴行为多样性和环境变化进行训练，由于高度非平稳性，使得鲁棒的表征学习尤其具有挑战性。为此，我们为合作智能体引入了下一状态预测辅助目标（图 1b，蓝色）。其直觉在于，在合作场景中，预测状态转移需要同时对环境动态和伙伴动作进行建模，从而促使智能体构建能够捕捉这两者的表征。使合作智能体能够隐式预测伙伴动作，还可能带来更快的推断速度，并更好地适应伙伴的协调行为。通过消融研究，我们表明，在联合多样性训练机制下，该辅助训练目标能够持续提升泛化性能。

我们在合作觅食任务和 Overcooked 领域中评估了 CECP，所采用的实验方案同时包含留出合作伙伴和留出环境。在这些设置下，CECP 始终优于强大的基线方法，展现出在社会和环境变化共同作用下更强的泛化能力。此外，我们开展了人类实验，结果表明，在与新颖人类玩家进行实时评估时，CECP 取得了最高的对人类参与者的协作性能；参与者也认为 CECP 智能体最有可能适应他们的动作。在可解释性分析中，我们发现，引入下一状态预测目标和伙伴多样性会促使模型形成关于伙伴动作和技能水平的表征（图 1c）。

我们的贡献如下：

- 我们提出了 CECP，这是一种两阶段训练范式，在统一设置中结合了伙伴多样性和环境多样性。
- 我们通过实验证明，在同时发生社会情境转变的情况下，该方法在合作觅食任务和 Overcooked（一项广受欢迎的 ZSC基准）上均取得了更好的性能。
- 我们发现，CECP 智能体能够形成可预测伙伴行为并推断合作伙伴技能水平的表征，为深入且可解释的分析提供了有用工具。

## 2 相关工作

**零样本协作（ZSC）**。**自博弈（SP）**描述了一种训练方法，在该方法中，智能体通过与自身进行重复对局来学习。尽管自博弈在 Starcraft 和围棋等竞争性任务中取得了巨大成功，但在协作场景下，SP 智能体已知会在合作任务中表现不佳，因为它们往往无法与新伙伴进行协调 [25, 23]，[24]。SP 智能体倾向于学习特定的协作模式，而这些模式只有在与自身对局时才有效。大量研究 [30, 13,14, 22, 4] 采用**基于群体的训练（PBT）**方法 [10, 3] 构建多样化的训练合作伙伴群体，旨在覆盖新颖人类玩家在测试时的行为。尽管上述方法通常都在同一个单一环境中进行训练和评估，但我们的工作进一步探索了环境多样性的引入：一方面是在创建伙伴群体的阶段，另一方面是在针对该群体训练协作者的阶段。

**元强化学习（Meta-RL）**旨在通过在任务或环境的分布上进行训练，而非局限于单一固定场景，从而学习能够快速适应新任务的策略 [2, 28, 6, 26, 31]。类似方法也已扩展至多智能体环境 [32, 19, 18, 1, 29]。近期，在零样本协调（ZSC）领域，Jha 等人 [11] 证明，在一系列程序生成环境中训练单个自博弈智能体，并利用元强化学习，能够提升该智能体与未见过的伙伴进行协调的能力。然而，当 Jha 等人 [11] 尝试结合伙伴多样性与环境多样性时，标准模型未能取得良好表现，这促使我们探究需要采用哪些技术，才能应对这种高度复杂的训练分布

**预测性辅助目标**近期研究表明，预测模块能够改善多智能体协调中的伙伴适应能力。具体而言，显式的伙伴行动建模已被证明有助于推断合作伙伴意图 [18]，并使自体行为能够依据预测的队友行动进行调整[32, 27, 17]。除了直接预测伙伴之外，辅助预测目标也改善了合作多智能体环境中的表征学习 [7, 9]。在元强化学习和部分可观测条件下，预测目标同样被证明能够改善贝叶斯自适应信念表征 [31, 12]。我们的工作受到 Ndousse 等人 [16], 的启发；他们证明，在多智能体情境中，下一状态预测辅助目标能够促成社会学习，其依据是：准确预测下一状态，必然需要隐式地捕捉伙伴行动和环境转移。

## 3 预备知识

双人合作马尔可夫博弈是一个元组 $\langle \mathcal{S},\ \mathcal{A},\ \mathcal{T},\ R,\ \rho_0,\ H \rangle$，其中 $\mathcal{S}$ 是状态空间，$\mathcal{A}$ 是共享动作空间，$\mathcal{T}\colon \mathcal{S} \times \mathcal{A} \times \mathcal{A} \to \Delta(\mathcal{S})$ 是转移函数，$R\colon \mathcal{S} \times \mathcal{A} \times \mathcal{A} \to \mathbb{R}$ 是共享奖励函数，$\rho_0 \in \Delta(\mathcal{S})$ 是初始状态分布，$H$ 是时域。对于

协作者策略 $\pi_c$ 和伙伴策略 $\pi_p$，我们将期望回报定义为

$$S(\pi_c, \pi_p) = \mathbb{E}_{\substack{s_0 \sim \rho_0, s_{t+1} \sim \mathcal{T}, \\ a_t^c \sim \pi_c,\, a_t^p \sim \pi_p}} \left[ \sum_{t=0}^{H} R(s_t, a_t^c, a_t^p) \right]. \tag{1}$$

**基于群体的训练（PBT）：** 我们假设伙伴策略的分布 $\pi_p \sim \mathcal{P}$，用于表示协作者在部署时可能遇到的策略、惯例和技能差异。在训练过程中，协作者的目标是：

$$J_{\text{PBT}}(\pi_c) = \mathbb{E}_{\pi_p \sim \mathcal{P}} \left[ S(\pi_c, \pi_p) \right]. \tag{2}$$

**元强化学习：** 我们考虑遵循先前工作 [11]的马尔可夫博弈分布 $\mathcal{M}$。每个$m \sim \mathcal{M}$ 共享相同的 $\mathcal{S}$、$\mathcal{A}$、$\mathcal{T}$、$R$、$H$，但具有不同的初始状态分布 $\rho_0^m$。将$S(\pi_c$、 $\pi_p$； $m)$记为博弈 $m$下的回报，则自博弈（CEC， [11]）的训练目标定义为：

$$J_{\text{CEC}}(\pi_c) = \mathbb{E}_{m \sim \mathcal{M}} \left[ S(\pi_c, \pi_c; m) \right]. \tag{3}$$

## 4 CECP（跨环境协作）

CECP训练能够在任务和伙伴两方面实现泛化的合作智能体。训练分为两个阶段。

**阶段1：跨环境的伙伴群体训练**。我们通过自博弈，在由程序化生成 [11]产生的环境分布下，独立训练 $N$ 伙伴智能体。我们采用Overcooked [21] 的JAX实现，以支持4096个并行环境模拟，并在NVIDIA A100 GPU上以约30,000帧/秒的速度进行训练。每个自博弈伙伴 $\pi_p^{(i)}$通过以下方式获得：

$$\pi_p^{(i)} = \arg\max_{\pi} \mathbb{E}_{m \sim \mathcal{M}} \left[ S(\pi, \pi; m) \right], \quad i = 1, \ldots, N \tag{4}$$

为进一步引入伙伴技能水平的多样性，我们遵循 FCP 协议 [24] 智能体，为每个伙伴智能体改变随机种子，并将一组中间训练检查点（早期、中期、晚期）加入伙伴群体 $\mathcal{P}_{\mathcal{M}}$。

**阶段 2：针对 社会情境转变进行协作者训练**。 随后对协作者 $\pi_c$ 进行训练，使其在任务分布和伙伴分布上的联合期望回报最大化：

$$J_{\text{CECP}}(\pi_c) = \mathbb{E}_{m \sim \mathcal{M},\, \pi_p \sim \mathcal{P}_{\mathcal{M}}} [S(\pi_c, \pi_p; m)] \tag{5}$$

**辅助下一状态预测**。 为支持社会情境转变下的表征学习，我们为协作者配备了辅助下一状态预测目标。协作者架构由三个组件构成：(1) RNN 编码器 $q_\phi(z \mid \tau_{0:t})$，将观测历史映射为潜在表征 $z$。（图 1b，绿色）。(2) 预测头 $p_\theta(s_{t+1} \mid z,\ a_t^c)$，根据潜在表征和协作者自身的动作解码下一状态（图 1b，蓝色）。(3) 策略头$\pi_\psi(a_t^c \mid z)$，以潜在表征 $z$ 为条件，并通过近端策略优化进行训练（图 1b，红色）。
辅助预测损失为：

$$\mathcal{L}_{\text{pred}}(\phi, \theta) = \mathbb{E}_{(\tau_{0:t},\, a_t^c,\, s_{t+1})} \left[ -\log p_\theta(s_{t+1} \mid q_\phi(\tau_{0:t}),\, a_t^c) \right]. \tag{6}$$

由于 $s_{t+1}$ 取决于伙伴行为 $a_t^p$ 和特定于任务的动态 $\mathcal{T}$，最小化 $\mathcal{L}_{\text{pred}}$要求编码器 $q_\phi$ 在 $z$ [16]中捕获有关伙伴行为和环境结构的信息。完整目标函数为：

$$J_{\text{CECP-}\lambda}(\pi_c) = J_{\text{CECP}}(\pi_c) - \lambda\, \mathcal{L}_{\text{pred}}(\theta, \phi). \tag{7}$$

## 5 实验设置

我们提出三个与方法相关的问题，以此作为实验的动机：

**问题1**：CECP 能否利用环境和伙伴多样性，泛化到新颖的环境/伙伴组合？
**问题2**：在与新的人类用户协调时，我们的方法是否比有竞争力的基线方法表现更好？
**问题3：** 下一状态预测和伙伴多样性如何影响表征学习和零样本协调（ZSC）性能？

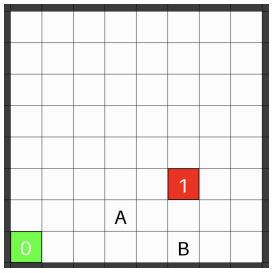

图2：合作觅食环境。两个智能体（0 和 1）必须抵达同一个目标（A 或 B），并停留在该目标上。

### 5.1 环境

**合作觅食：** 在这项任务中，我们受到合作任务的启发
用于神经科学研究 [5], 2智能体处于一个具有两个目标位置的网格世界中。为了共同获得奖励，两个智能体必须到达同一个目标位置并停留在那里。为了生成环境分布，我们均匀采样位置来放置两个目标，并留出一组目标位置用于评估（图2）。

**Overcooked：** 在广泛使用的ZSC基准Overcooked中 [3], 两名玩家需要相互合作，分担烹饪职责，并通过配送完成的菜肴共同获得奖励（图3）。传统上使用5个原始布局（图14，附录）来评估不同的协调挑战。我们在这5个原始布局上评估智能体；对于CEC和CECP方法，这些布局在训练期间均被留出；此外，我们还在另外100个程序生成布局上进行评估，这些布局同样在训练期间被留出。

### 5.2 基线方法

我们将方法CECP与文献中已有的基线方法进行比较，尤其是自博弈（SP）、虚拟协同博弈（FCP, [24]）和跨环境合作（CEC, [11]）。对于Overcooked评估，我们还纳入了：最大熵群体（MEP, [30]），一种通过群体熵奖励强制实现群体多样性的基于群体的方法；跨局优化、混合局强制多样性（CoMeDi, [22]），另一种通过混合博弈强制实现群体多样性的基于群体的方法；以及具有上下文感知探索的同伴适应方法（PACE, [15]），该方法通过添加同伴识别奖励来尝试改进协作者。

### 5.3 评估标准

**交叉对战：** 我们评估来自同一训练范式和架构、但使用不同初始随机种子的独立训练智能体之间的协调能力。在所有方法中，使用不同初始随机种子的智能体在训练期间从未作为伙伴合作过，因此交叉协作表现衡量的是智能体在人工智能体间协调情境下与新伙伴协调的能力。

**人机协作：** 在 Overcooked 中，我们通过评估智能体与行为克隆的人类代理智能体及真实人类伙伴的合作表现，来评估人机零样本协调。我们从热门众包工作平台 Prolific 招募了 73 名参与者。参与者的平均报酬为每小时 19.46 美元。所有实验流程均经机构审查委员会（IRB）批准。招募的人类参与者分别与每个智能体进行 200 个时间步的对局，并记录其表现以及关于协调感受的问卷回答。

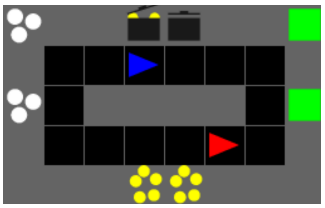
图 3：Counter Circuit 布局的可视化，这是 Overcooked 原有的 5 种布局之一。两位厨师（蓝色和红色）需要合作，高效地制作并递送汤。

# 6 结果

## 6.1 CECP 在合作觅食中的新环境和新伙伴设置下实现泛化

为研究**问题1**，我们在合作觅食任务中评估 CECP，并将其与现有基线方法进行比较。为测试其对新环境和新伙伴的泛化能力，我们留出一组目标位置和一组伙伴智能体作为评估测试集。图4展示了这些结果。虽然所有方法在熟悉的伙伴和熟悉的环境中表现良好，但当目标位置不同于训练环境时，SP（紫色）和 FCP（蓝色）都无法实现泛化。相比之下，CEC（绿色）能够很好地泛化到留出的目标位置，但与新伙伴配对时性能会下降。在同时发生社会情境转变的情况下（图4，右下角），CEC 的表现约为 CECP（红色）的一半。我们的方法是唯一能够同时泛化到在新环境中与新伙伴进行合作的方法。

CEC 从熟悉伙伴到新伙伴时性能下降，这表明在该设置下，仅凭环境多样性并不能阻止自博弈智能体形成惯例，因此促使我们利用 CEC 智能体群体中形成的多样化惯例来训练合作智能体（CECP），使其能够泛化到未见过的伙伴。

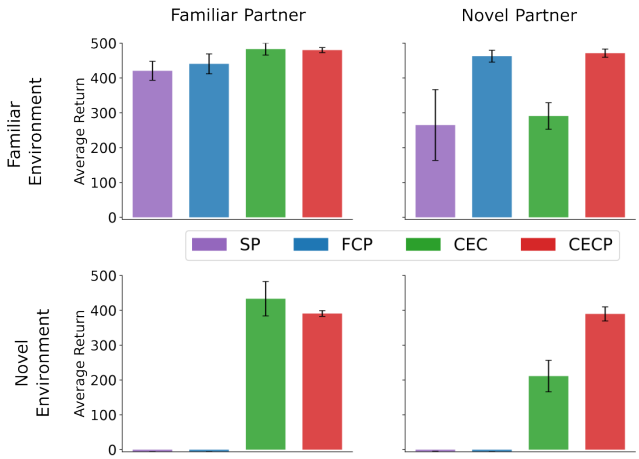


图4：在合作觅食任务中，CECP 与基线方法在熟悉/新环境和伙伴设置下的性能。除 CEC 外，所有先前方法都无法应对留出环境；CEC 同样采用跨环境训练。CECP 在<条目 id='1'>社会情境转变</条目>下表现出额外的鲁棒性。

## 6.2 CECP 在 Overcooke 中的零样本人机协作方面优于基线方法 d




图5：在 Counter Circuit 布局的特定场景中，CEC 智能体群体的行为差异示例。**a,** 一个智能体正拿着洋葱，面向左侧，位于右侧的锅下方。**b,** 当 CEC 智能体群体面对<样式 id='5'>a</样式>中的场景时的平均动作分布。

用于训练 CECP 的自博弈智能体群体中也会形成不同的惯例吗？我们收集了 CECP 伙伴群体中的自博弈智能体在一些 Overcooked 布局中的轨迹，

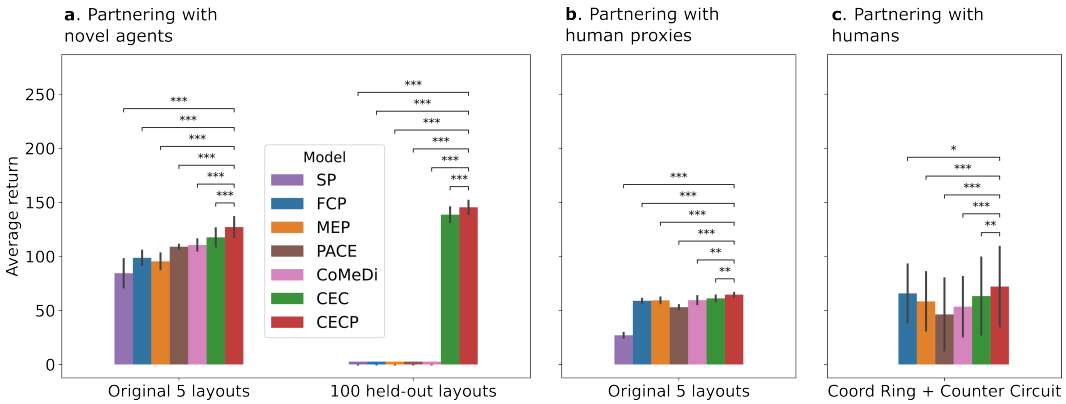


图6：在 Overcooked 布局上的性能。**a**、CECP 与基线方法在原始的 5 个 Overcooked 布局上的交叉对战平均回报。**b**、各方法与通过行为克隆（BC）训练的人类代理配对时的平均回报。**c**、各方法与人类配对时的平均回报。采用双侧 Welch t检验评估显著性，***：$p < 0.001$，**：$p < 0.01$，*：$p < 0.05$。

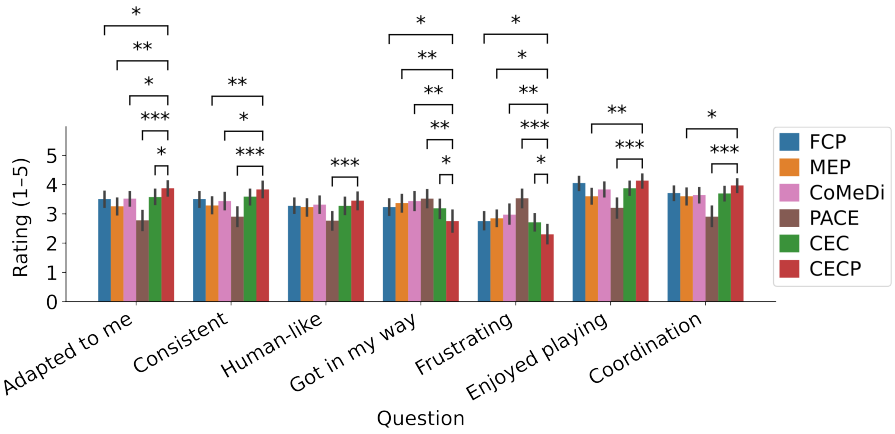


图7：针对调查问题的各方法参与者平均评分（李克特量表，1-5），采用配对Wilcoxon符号秩检验进行显著性检验，*：$p < 0.05$，**：$p < 0.01$，***：$p < 0.001$

并评估不同智能体在相似情境下是否会表现出不同的行为。图5展示了 Counter Circuit 布局中的一个例子：当智能体手持洋葱、面朝左侧并位于右侧锅的正下方时，群体分成了两种策略，一种倾向于向左移动，另一种倾向于向上移动并与锅交互。伙伴群体中的这些不同惯例使合作型CECP智能体能够在训练期间利用群体多样性。

为了评估我们的方法在 Overcooked 中的表现，我们首先在 5 种原始布局下，通过交叉对战将其与现有基线方法进行了比较。为确保公平比较，对于未在环境分布上进行训练的先前方法（SP、FCP、MEP、PACE、CoMeDi），我们分别在这 5 种布局上进行训练；而 CEC 和 CECP 则在程序化生成布局的环境分布上进行训练，并将 5 种原始布局留作测试集。平均来看，在这 5 种原始布局上，尽管 CECP 从未直接在这些布局上训练过，其表现仍显著优于现有基线方法（图 6a，左）。在另外 100 种程序化生成的留出布局上（图 6a，右），单环境方法（SP、FCP、MEP、PACE、CoMeDi）无法取得良好表现，而 CECP 仍优于 CEC。当将现有基线方法和我们的方法与人类代理进行配对，在

最初的 5 种 Overcooked 布局中，我们再次观察到类似结果：我们的方法 CECP 优于现有基线方法（图 6b）。

为直接回答**Q2**，我们接下来测试了各方法在 Coordina- tion Ring 和 Counter Circuit 布局中与人类协调的能力。$N = 73$ 参与者需要分别与每种方法（FCP、MEP、CoMeDi、PACE、CEC、CECP）进行一场 200 时间步的游戏，游玩顺序随机，并被询问对各方法的体验。在这两种布局中，CECP 在平均回报方面再次显著优于所有基线方法（图 6c）。在问卷反馈中（图 7），用户对 CECP 适应性的评分高于基线方法，而对其挫败感和干扰的评分低于基线方法。

### 6.3 下一状态预测和伙伴多样性塑造 CECP 表征

CECP 如何在社会情境转变下实现泛化？为探究这一问题（**Q3**），我们试图了解哪些组成部分推动了 CECP 人类零样本协调表现的提升。实证结果表明，移除下一状态预测模块会导致训练期间的交叉对战表现明显下降（图 8），这凸显了预测的重要性。

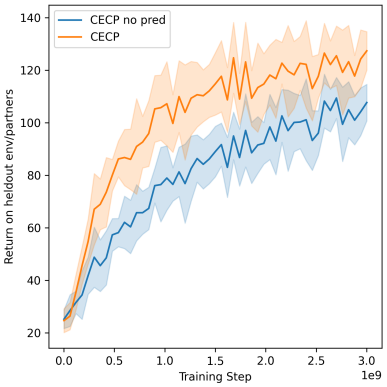


图 8：比较训练期间 CECP 的表现：使用下一状态预测模块（橙色）与不使用该模块（蓝色）。阴影区域表示各智能体之间的标准差。

为更好地理解其中原因，我们借鉴 Ndousse 等人 [16]，的观点：他们认为，在合作任务中，成功的下一状态预测不仅取决于理解环境如何演变，还取决于预判伙伴将采取何种行动（见附录第 8.7 节，图 17）。

更广泛地说，在面对众多伙伴和环境时，训练良好的 CECP 智能体在整个回合中预测 AI 合作伙伴动作（上、下、左、右或停留）的准确率始终显著高于随机水平（图 9a，“最佳 AI 伙伴”），这表明下一状态预测促使智能体学习伙伴行为的隐式模型。这种动作预测同样能够泛化至人类：CECP 对人类受试者动作的预测准确率也高于随机水平（图 9a，“人类受试者”）。此外，我们发现，CECP 准确预测其人类伙伴动作的能力与人类的表现相关，这表明，技能更高的伙伴所产生的动作更容易预测。

我们假设，CECP 的泛化和适应能力源于其能够更好地推断伙伴，因为潜在变量 $z$ 是通过对具有不同偏好和技能水平的伙伴进行表征学习而训练得到的。为探究这一点，我们分析潜在变量 $z$ 预测伙伴训练阶段（早期/中期/后期）的能力（使用线性分类器）。在一个 CECP 模型示例中（图10a，红色），预测准确率始终高于随机水平，并随着回合进行而提高。使用 CEC 智能体的隐藏层进行预测时，则明显没有这一现象（图10a，绿色）。在原则主成分（PC）空间中可视化 CECP 模型每回合的潜在$z$ 活动轨迹后，我们进一步发现，潜在活动会根据伙伴的检查点阶段逐渐分离（图

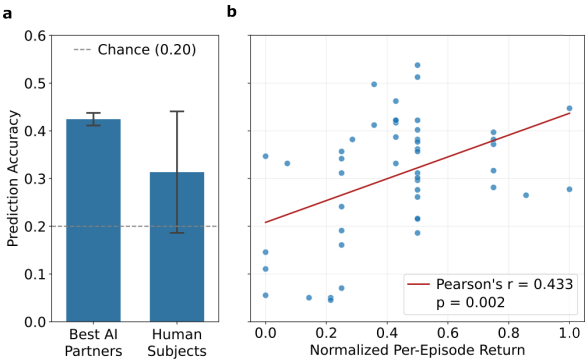


图 9：基于下一状态预测的 CECP 伙伴动作预测准确率。**a,** 人工智能训练合作伙伴和人类参与者在各回合/时间步中的预测准确率。**b,** CECP 对人类动作的预测准确率与回合表现之间的相关性。

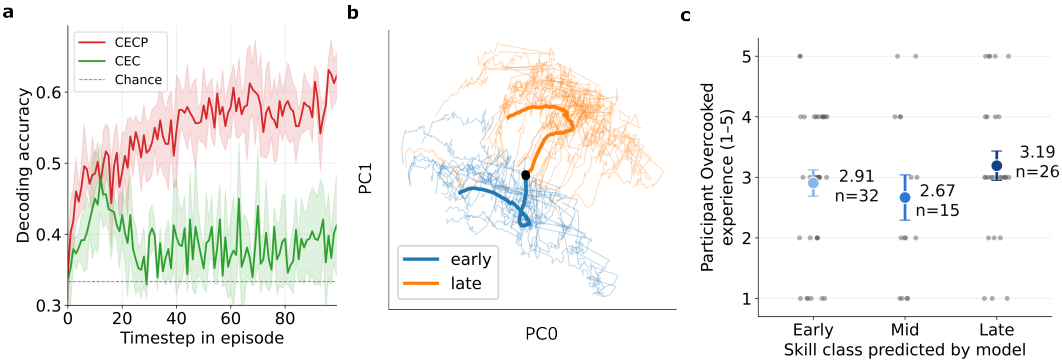


图10：CECP 潜变量中的技能水平表征。**a,** 在回合的各个时间步预测伙伴训练阶段的准确率，显示 CECP 的潜变量 $z$ 比 CEC 的隐藏层向量具有更好的表征学习能力。**b,** 在主成分空间中可视化每回合的潜在活动轨迹，并按伙伴检查点阶段（早期与后期）着色。粗线表示各检查点阶段的平均活动轨迹，浅色线表示20条子采样轨迹。黑点表示回合开始时的活动状态。**c,** 使用在**a**中训练的相同解码器，根据 CECP 伙伴在交互期间推断出的技能水平，对参与者自我报告的 Overcooked 经验进行分组。

10b），这进一步表明，在交互过程中，CECP 的潜在空间会围绕伙伴能力动态组织。当人类参与者与 CECP 智能体交互时，我们将相同的伙伴训练阶段分类器（早期/中期/后期）应用于 CECP 智能体的隐藏状态，从而读出 CECP 智能体对其人类伙伴技能水平的估计。我们发现，CECP 智能体对伙伴技能的估计大致与参与者自我报告的 Overcooked 经验相符（图10c）：被分类为后期伙伴的参与者报告的平均经验最高。

推断伙伴的技能水平是否也会带来自适应行为？为研究这一点，我们收集了 CECP 智能体在不同训练阶段与伙伴 CEC 智能体交互的轨迹，找出各训练阶段中与伙伴一起频繁到访的特定状态，并考察 CECP 智能体在各伙伴训练阶段下的平均动作分布。图 11 展示了协调环中的一个示例状态，并突出显示了 CECP 智能体会根据伙伴的技能水平调整行为：如果伙伴训练充分，CECP 智能体会优先把洋葱放进右侧的锅，因为它预期伙伴也会拿起洋葱并将其放入锅中。然而，如果伙伴训练不足，CECP 智能体则倾向于向上移动，把洋葱放进上方的锅，以便至少让一口锅开始烹饪（见图 11 及附录第 8.9 节）。

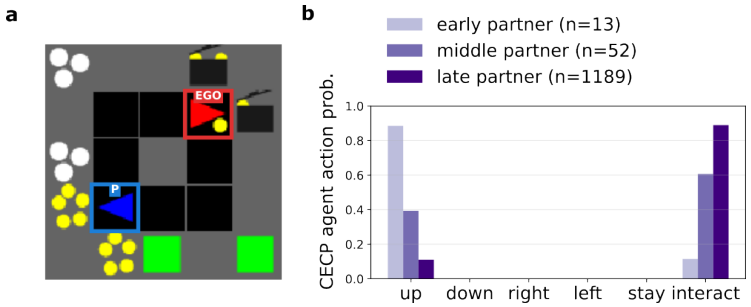


图11：在 Coordination Ring 布局中的特定状态下，CECP 智能体与处于训练早期、中期或后期阶段的伙伴进行交互时的平均动作分布。**a,** 状态示意图。处于不同训练阶段的 CECP（自体）智能体以红色表示，伙伴智能体以蓝色表示。**b,** CECP（自体）智能体在该状态下针对各伙伴训练阶段的平均动作分布。CECP 智能体会根据伙伴的技能水平调整动作选择：如果伙伴训练程度较低，它倾向于向“上”转向，并将最后一个洋葱放入上方的锅中，以至少完成一锅汤。面对更专业的伙伴时，CECP 智能体则会采取更高效的动作，通过“交互”将自己的洋葱放入最近的可用锅中。

## 7 结论

我们提出 CECP，这是一种将伙伴多样性与环境多样性相结合、用于零样本协作的两阶段训练框架。CECP 先通过在程序生成环境中进行自博弈来训练伙伴群体，再针对该群体训练协作者，从而生成能够适应伙伴与环境同时变化的智能体。我们还表明，下一状态预测目标和伙伴多样性能够提升性能，并促使协作者形成可隐式追踪伙伴行为和技能水平的表征。

**局限性与未来工作：** CECP 的性能受限于其训练所用的伙伴群体和环境分布中的多样性。未来若能提高伙伴群体的多样性或扩大环境范围，或可进一步提升 CECP 的泛化性能。

我们发现，下一状态预测能够改善协调，但其作用机制尚未得到充分理解：潜在表征捕捉的是广泛的技能水平，而非细粒度的伙伴特定行动意图。将潜在表征 $z$ 中的伙伴特定信息与环境特定信息分离（例如通过结构化瓶颈），或可使表征更易解释且更具样本效率。

最后，尽管 Overcooked 任务一直是测试人机协作的标准，更新的基准测试 [8] 若能纳入更高的复杂性和部分可观测性，或可进一步展现 CECP 的有效性。

## 致谢与经费披露

本研究获得了 Schmidt Sciences 可信人工智能科学项目、UW-Amazon Science Gift Hub、UW-Tsukuba Amazon NVIDIA 跨太平洋人工智能倡议（XPAI）、Tinker 研究资助、Open Philanthropy、Coefficient Giving、丰田研究院 Kaizen 倡议、Schmidt AI2050 研究员项目、NSF CISE RI 项目（奖项编号：#2550849）以及 NSF GRFP 2023356971 的支持。本研究还部分获得了超威半导体公司通过 AMD 大学项目人工智能与高性能计算集群提供的支持。本材料基于美国国防高级研究计划局和美国空军研究实验室支持的工作，合同编号为：FA8650-23-C-7316。本材料中表达的任何观点、研究结果、结论或建议均属于作者，并不一定反映 AFRL 或 DARPA 的观点。

## 参考文献


[1] Maruan Al-Shedivat、Trapit Bansal、Yuri Burda、Ilya Sutskever、Igor Mordatch 和 Pieter Abbeel。《通过元学习在非平稳和竞争性环境中持续适应》，2018年2月。网址：http://arxiv.org/abs/1710.03641。arXiv:1710.03641[cs]。[2] Jacob Beck、Risto Vuorio、Evan Zheran Liu、Zheng Xiong、Luisa Zintgraf、Chelsea Finn 和 Shimon Whiteson。《元强化学习综述》，2023年1月。网址：http://arxiv.org/abs/2301.08028。arXiv:2301.08028 [cs]。[3] Micah Carroll、Rohin Shah、Mark K. Ho、Thomas L. Griffiths、Sanjit A. Seshia、Pieter Abbeel 和 Anca Dragan。《学习人类对人机协作的效用》，2020年1月。网址：http://arxiv.org/abs/1910.05789。arXiv:1910.05789 [cs]。[4] Paresh Chaudhary、Yancheng Liang、Daphne Chen、Simon S. Du 和 Natasha Jaques。《通过在线对抗训练和生成模型改善人机协作》，2025年10月。网址：http://arxiv.org/abs/2504.15457。arXiv:2504.15457 [cs]。[5] Yuan Cheng、Yusi Chen、Myungji Kwak、Ross P. Kempner、Rudramani Singha、Jared Winslow、Runqi Liu、Umais Khan、Tessa Spangler、Alvi Khan、Talmo Pereira、Matthew Whiteway、Evan S. Schaffer、Nuttida Rungratsameetaweemana、Nan Yang 和 Herbert Zheng Wu。《合作行为中前额叶皮层的非对称社会表征》。*bioRxiv*，页面 2025.08.27.672249，2025年8月。ISSN 2692-8205。doi：10.1101/2025.08.27.672249。网址：https://pmc.ncbi.nlm.nih.gov/articles/PMC12407837/。[6] Yan Duan、John Schulman、Xi Chen、Peter L. Bartlett、Ilya Sutskever 和 Pieter Abbeel。RL$^2$：《通过慢速强化学习实现快速强化学习》，2016年11月。网址：http://arxiv.org/abs/1611.02779。arXiv:1611.02779 [cs, stat]。[7] Mingxiao Feng、Wengang Zhou、Yaodong Yang 和 Houqiang Li。《多智能体强化学习的联合预测表征》，2022年9月。网址：https: // openreview.net/forum?id=S80ioOGLpD9&utm_source=chatgpt.com。[8] Tobias Gessler、Tin Dizdarevic、Ani Calinescu、Benjamin Ellis、Andrei Lupu 和 Jakob Nicolaus Foerster。《OvercookedV2：重新思考 Overcooked 以实现零样本协作》，2025年3月。网址：http://arxiv.org/abs/2503.17821。arXiv:2503.17821 [cs]。[9] Qinwei Huang、Stefan Wang、Simon Khan、Garrett Katz 和 Qinru Qiu。《基于信念的多智能体系统的预测辅助学习》，2025年11月。网址：http://arxiv.org/ abs/2511.01078。arXiv:2511.01078 [cs]。

[10] Max Jaderberg、Valentin Dalibard、Simon Osindero、Wojciech M. Czarnecki、Jeff Donahue、Ali Razavi、Oriol Vinyals、Tim Green、Iain Dunning、Karen Simonyan、Chrisantha Fernando 和 Koray Kavukcuoglu。《基于种群的神经网络训练》，2017年11月。网址：http://arxiv.org/abs/1711.09846。arXiv：1711.09846 [cs]。

[11] Kunal Jha、Wilka Carvalho、Yancheng Liang、Simon S. Du、Max Kleiman-Weiner 和 Natasha Jaques。《跨环境合作实现零样本多智能体协调》，2025年4月。网址：http://arxiv.org/abs/2504.12714。arXiv：2504.12714 [cs]。

[12] Po-Chen Kuo、Han Hou、Will Dabney 和 Edgar Y. Walker。《预测编码增强元强化学习，在部分可观测性下实现可解释的贝叶斯最优信念表示》，2025年10月。网址：http://arxiv.org/abs/2510.22039。arXiv：2510.22039 [cs]。

[13] Yancheng Liang、Daphne Chen、Abhishek Gupta、Simon S. Du 和 Natasha Jaques。《使用生成式智能体学习与人类合作》，2024年11月。网址：http://arxiv. org/abs/2411.13934。arXiv：2411.13934 [cs]。

[14] Andrei Lupu、Brandon Cui、Hengyuan Hu 和 Jakob Foerster。《零样本协作的轨迹多样性》。载于第*38*届国际机器学习会议论文集，第7204–7213页。PMLR，2021年7月。网址：https://proceedings.mlr.press/ v139/lupu21a.html。

[15] Long Ma、Yuanfei Wang、Fangwei Zhong、Song-Chun Zhu 和 Yizhou Wang。《具情境感知探索的快速同伴适应》，2024年8月。网址：http://arxiv.org/abs/ 2402.02468。arXiv:2402.02468 [cs.AI].[16] Kamal Ndousse、Douglas Eck、Sergey Levine 和 Natasha Jaques。《通过多智能体强化学习实现的涌现式社会学习》，2021年6月。网址：http://arxiv.org/abs/2010. 00581。arXiv:2010.00581 [cs].[17] Andrew Ni、Simon Stepputtis、Stefanos Nikolaidis、Michael Lewis、Katia P. Sycara 和 Woojun Kim。《心智理论引导的零样本协同策略适应》，2026年2月。网址：http://arxiv.org/abs/2602.12458。arXiv:2602.12458 [cs] 版本：1。[18] Neil C. Rabinowitz、Frank Perbet、H. Francis Song、Chiyuan Zhang、S. M. Ali Eslami 和 Matthew Botvinick。《机器心智理论》，2018年3月。网址：http://arxiv.org/abs/ 1802.07740。arXiv:1802.07740 [cs].[19] Roberta Raileanu、Emily Denton、Arthur Szlam 和 Rob Fergus。《在多智能体强化学习中以自我建模他者》，2018年3月。网址：http://arxiv.org/abs/1802.09640。arXiv:1802.09640 [cs].[20] Constantin Ruhdorfer、Matteo Bortoletto、Anna Penzkofer 和 Andreas Bulling。《〈Overcooked〉泛化挑战：使用无监督环境设计，在未知环境中评估与新伙伴的合作》，2025年9月。网址：http://arxiv.org/abs/2406.17949。arXiv:2406.17949 [cs].[21] Alexander Rutherford、Benjamin Ellis、Matteo Gallici、Jonathan Cook、Andrei Lupu、Gardar Ingvarsson、Timon Willi、Ravi Hammond、Akbir Khan、Christian Schroeder de Witt、Alexandra Souly、Saptarashmi Bandyopadhyay、Mikayel Samvelyan、Minqi Jiang、Robert Tjarko Lange、Shimon Whiteson、Bruno Lacerda、Nick Hawes、Tim Rocktaschel、Chris Lu 和 Jakob Nicolaus Foerster。《JaxMARL：基于 JAX 的多智能体强化学习环境与算法》，2024年11月。网址：http://arxiv.org/abs/2311.10090。arXiv:2311.10090 [cs].[22] Bidipta Sarkar、Andy Shih 和 Dorsa Sadigh。《人类与人工智能协作的多样化惯例》，2023年10月。网址：http://arxiv.org/abs/2310.15414。arXiv:2310.15414 [cs].[23] David Silver、Julian Schrittwieser、Karen Simonyan、Ioannis Antonoglou、Aja Huang、Arthur Guez、Thomas Hubert、Lucas Baker、Matthew Lai、Adrian Bolton、Yutian Chen、Timothy Lillicrap、Fan Hui、Laurent Sifre、George van den Driessche、Thore Graepel 和 Demis Hassabis。《在没有人类知识的情况下精通围棋》。《自然》，550(7676)：354–359，2017年10月。ISSN 1476-4687。doi：10.1038/nature24270。网址：https://www.nature.com/articles/nature24270。[24] D. J. Strouse、Kevin R. McKee、Matt Botvinick、Edward Hughes 和 Richard Everett。《在没有人类数据的情况下与人类合作》，2022年1月。网址：http://arxiv.org/abs/ 2110.08176。arXiv:2110.08176 [cs].[25] Oriol Vinyals、Igor Babuschkin、Wojciech M. Czarnecki、Michaël Mathieu、Andrew Dudzik、Jun- young Chung、David H. Choi、Richard Powell、Timo Ewalds、Petko Georgiev、Junhyuk Oh、Dan Horgan、Manuel Kroiss、Ivo Danihelka、Aja Huang、Laurent Sifre、Trevor Cai、John P. Agapiou、Max Jaderberg、Alexander S. Vezhnevets、Rémi Leblond、Tobias Pohlen、Valentin Dalibard、David Budden、Yury Sulsky、James Molloy、Tom L. Paine、Caglar Gulcehre、Ziyu Wang、Tobias Pfaff、Yuhuai Wu、Roman Ring、Dani Yogatama、Dario Wünsch、Katrina McKinney、Oliver Smith、Tom Schaul、Timothy Lillicrap、Koray Kavukcuoglu、Demis Hassabis、Chris Apps 和 David Silver。《通过多智能体强化学习达到《星际争霸II》宗师级水平》。《自然》，575(7782)：350–354，2019年11月。ISSN 1476-4687。doi：10.1038/s41586-019-1724-z。网址：https://www.nature.com/articles/s41586-019-1724-z。[26] Jane X. Wang、Zeb Kurth-Nelson、Dhruva Tirumala、Hubert Soyer、Joel Z. Leibo、Remi Munos、Charles Blundell、Dharshan Kumaran 和 Matt Botvinick。《学习强化学习》，2017年1月。网址：http://arxiv.org/abs/1611.05763。arXiv:1611.05763 [cs, stat].[27] Xue Yan、Jiaxian Guo、Xingzhou Lou、Jun Wang、Haifeng Zhang 和 Yali Du。《用于零样本人机协同的高效端到端训练方法》，2023年11月。网址：https://openreview.net/forum?id=6ePsuwXUwf。

[28] Tianhe Yu、Deirdre Quillen、Zhanpeng He、Ryan Julian、Avnish Narayan、Hayden Shively、Adithya Bellathur、Karol Hausman、Chelsea Finn 和 Sergey Levine。《Meta-World：多任务与元强化学习的基准和评估》，2021年6月。URL http://arxiv.org/abs/1910.10897。arXiv:1910.10897 [cs].[29] Shenao Zhang、Li Shen、Lei Han 和 Li Shen。《多智能体强化学习中智能体的元表示学习》，2023年6月。URL http://arxiv.org/abs/2108.12988。arXiv:2108.12988 [cs].[30] Rui Zhao、Jinming Song、Yufeng Yuan、Hu Haifeng、Yang Gao、Yi Wu、Zhongqian Sun 和 Yang Wei。《用于零样本人机协作的最大熵群体基于训练》，2022年6月。URL http://arxiv.org/abs/2112.11701。arXiv:2112.11701 [cs].[31] Luisa Zintgraf、Sebastian Schulze、Cong Lu、Leo Feng、Maximilian Igl、Kyriacos Shiarlis、Yarin Gal、Katja Hofmann 和 Shimon Whiteson。《VariBAD：通过元学习实现的变分贝叶斯自适应深度强化学习》。2021年。[32] Luisa Zintgraf、Sam Devlin、Kamil Ciosek、Shimon Whiteson 和 Katja Hofmann。《通过元学习实现深度交互式贝叶斯强化学习》，2022年4月。URL http: //arxiv.org/abs/2101.03864。arXiv:2101.03864 [cs].

# 8 附录

## 8.1 CECP 训练

预测器模块详情。(1) 一个 RNN 编码器 $q_\phi(z \mid \tau_{0:t})$，通过 CNN 特征提取器输入观测，将观测历史映射为潜在表征 $z$。（图 1b，绿色）。(2) 一个预测头 $p_\theta(s_{t+1} \mid z,\ a_t^c)$，根据潜在表征和协作者自身的动作解码下一状态（图 1b，蓝色）。潜在表征 $z$ 与独热动作编码进行拼接，投影为平面向量，再重塑为网格，随后通过反卷积层重构下一次观测。(3) 一个策略头 $\pi_\psi(a_t^c \mid z)$，以潜在表征 $z$ 为条件，并通过近端策略优化进行训练（图 1b，红色）。

程序化生成详情。为了生成环境分布，我们采用了 [11], 中提出的程序化环境生成方法，通过改变洋葱、锅和交付口的位置以及布局方向来生成环境；该方法是 JaxMARL 项目的扩展 [21]。

CECP 训练首先包括在环境分布中训练一组自博弈智能体。为此，我们遵循 CEC[11] 中提出的程序化环境生成与训练方法。

为了训练 CECP 协作者智能体，我们使用 6 个经过训练的伙伴智能体及其训练早期、中期和后期的检查点来构建伙伴群体，并留出 2 个伙伴智能体用于测试。协作者智能体和伙伴智能体在由 NVIDIA A100 GPU 组成的集群上进行训练。每个步骤约需 1 天训练时间。训练 CECP 协作者智能体时使用了以下超参数：

| 超参数 | 值 |
|---|---|
| CNN 卷积核 | [2, 2], [2, 2] |
| CNN 通道 | [64, 32] |
| 编码器全连接层尺寸 | [512, 512] |
| 读出层尺寸 | 256 |
| 循环层类型 | LSTM |
| 循环层大小 | 256 |
| 循环层 | 1 |
| 演员头大小 | [256, 192, 128, 64] |
| 评论家头大小 | [512, 256, 192, 128] |
| 解码器投影大小 | $9 \times 9 \times 32$ |
| 反卷积通道 | [64, 26] |
| 反卷积核 | [2, 2], [2, 2] |
| 反卷积步幅 | [1, 1] |
| 环境步数 | 3B（每个智能体） |
| 并行环境 | 4096（每个智能体） |
| 回合长度 | 256 |
| PPO 批次大小 | 16384 |
| PPO 训练轮数 | 4 |
| PPO 学习率 | 0.0003 |
| 学习率调度 | 线性 |
| 广义优势估计器（GAE）$\lambda$ | 0.95 |
| 折扣因子 $\gamma$ | 0.99 |
| 熵系数 | 0.005 |
| 比率裁剪 | 0.2 |
| 梯度裁剪范数 | 0.5 |
| 预测损失系数 | 1000 |

表 1：CECP 合作者的超参数

### 8.2 CEC 与 CECP 智能体在合作觅食中的交叉对战比较

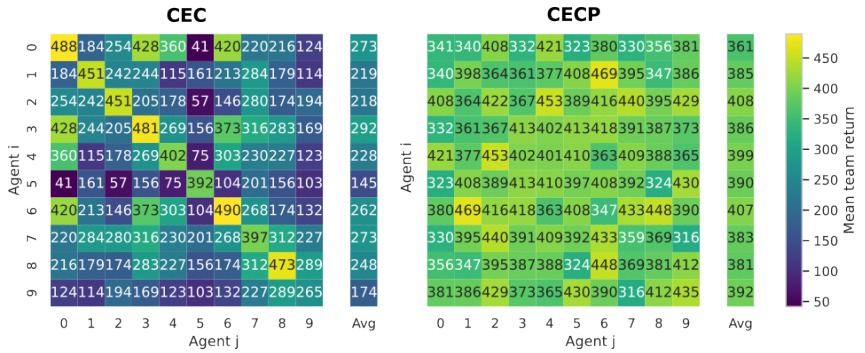


图12：CEC 智能体（左）和 CECP 智能体（右）在合作觅食任务中，针对留出的目标位置进行逐个智能体交叉对战的回报比较。每个热图中的智能体均采用相同方法训练，但使用不同的初始随机种子。对角线表示自博弈结果。对角线与非对角线项目之间的差距缩小，表明与 CEC 相比，CECP 提高了对新伙伴的鲁棒性。

### 8.3 评估伙伴群体的行为多样性

我们的方法 CECP 与纳入比较的许多基线方法（FCP、MEP、CoMeDi）一样，是一种基于群体的训练方法，因此我们试图量化各方法伙伴群体内部行为多样性的相对程度。

对于原始的 5 种 Overcooked 布局中的每一种，我们针对由 $N = 8$ 个智能体 $\{\pi_1, \ldots, \pi_N\}$组成的群体，首先通过汇集每个智能体在固定数量回合中的自博弈轨迹，收集一组轨迹 $\mathcal{D}$。对于每个智能体 $\pi_j$，我们重放来自 $\mathcal{D}$的轨迹，并计算该智能体在每个唯一状态 $s$下、对 $|\mathcal{A}| = 6$ 个动作的平均动作分布：$\pi_j(\cdot \mid s)$。随后，我们计算群体的詹森-香农散度（JSD），并对各状态取平均：

$$\mathrm{JSD}_{\mathrm{pop}} = \mathbb{E}_{s\sim\mathcal{D}}\Big[H\big(\bar{\pi}(\cdot \mid s)\big) - \frac{1}{N}\sum_{j=1}^{N} H\big(\pi_j(\cdot \mid s)\big)\Big], \tag{8}$$

其中，$\bar{\pi}(\cdot \mid s) = \frac{1}{N}\sum_{j=1}^{N} \pi_j(\cdot \mid s)$ 是群体在 $s$ 的动作分布的平均值，而 $H$ 是香农熵。

尽管 $\mathrm{JSD}_{\mathrm{pop}}$ 对群体多样性进行了总体评估，但我们还关注群体中一对智能体在最坏情况下可能有多相似。对于群体中的所有智能体对（$i$，$j$），我们计算状态平均成对 Jensen-Shannon 散度的最小值：$\min \mathrm{JSD}_{\mathrm{pair}} = \min_{i\neq j} \mathrm{JSD}_{ij}$。表2报告了各方法伙伴群体在原始的 5 种 Overcooked 布局中的评估结果，其中 $\mathrm{JSD}_{\mathrm{pop}}$及括号内的 $\min \mathrm{JSD}_{\mathrm{pair}}$ 分别表示相应指标。

表2：Overcooked 中各方法的伙伴群体多样性。每个单元格显示 $\mathrm{JSD}_{\mathrm{pop}}$（最小 $\mathrm{JSD}_{\mathrm{pair}}$）。每种方法的群体包含 8 个智能体。

| 布局 | CECP | FCP | MEP | CoMeDi |
|---|---|---|---|---|
| 非对称优势 | 0.387 (0.151) | 0.242 (0.067) | 0.370 (0.069) | 0.389 (0.079) |
| 协调环 | 0.245 (0.089) | 0.384 (0.152) | 0.409 (0.124) | 0.456 (0.169) |
| Counter Circuit | 0.334 (0.131) | 0.377 (0.137) | 0.366 (0.136) | 0.493 (0.179) |
| 狭窄房间 | 0.222 (0.093) | 0.261 (0.104) | 0.336 (0.120) | 0.467 (0.155) |
| 强制协调 | 0.308 (0.127) | 0.281 (0.112) | 0.200 (0.005) | 0.342 (0.088) |
| 所有布局 | 0.299 (0.089) | 0.309 (0.067) | 0.336 (0.005) | 0.429 (0.079) |

总体而言，我们观察到 CECP 伙伴群体的行为多样性与 FCP 的伙伴群体相近，不过在所有布局中，CECP 的 minJSD$_{\text{pair}}$ 都略高于其他方法。尽管考虑多样性的算法（MEP、CoMeDi）提高了 JSD$_{\text{pop}}$，但在某些布局中也会产生较低的 min JSD$_{\text{pair}}$，这表明其策略最终会收敛到少数特定惯例。

为了进一步确认 CECP 的伙伴群体中的智能体确实形成了不同的惯例，我们接下来考察了在协调环和 Counter Circuit 布局中的特定场景下，每个伙伴的平均动作分布。在协调环中，当一名智能体手持洋葱、面朝上方并位于左下角时，群体分成两类策略：一类偏好向上移动并发起顺时针移动，另一类偏好向右移动并发起逆时针移动（图13a）。类似地，在 Counter Circuit 中，当一名智能体手持洋葱、面朝左方并位于右侧锅的正下方时，群体分成两类策略：一类偏好向左移动，另一类偏好向上移动/与锅交互（图13b）。

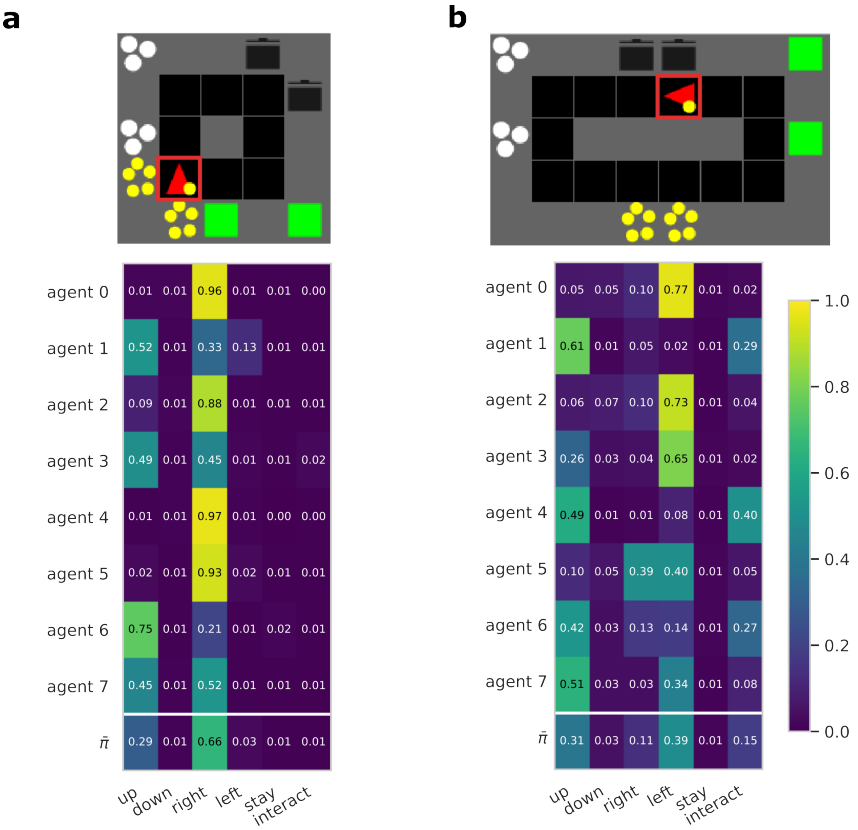


图13：CECP 的伙伴群体在协调环（a）和 Counter Circuit（b）布局中呈现不同惯例的两个示例场景。上图为相关智能体的位置、朝向和物品状态。下图为伙伴群体中每个智能体的平均动作分布。

### 8.4 训练人类代理智能体

**人类数据集** 我们使用 Carroll 等人发布的人类数据集 [3]，涵盖五种 Overcooked 布局（非对称优势、协调环、Counter Circuit、狭窄房间和强制协调），并采用与 Liang 等人相同的数据处理和行为克隆（BC）实现 [13]。

人类数据集按 70% / 30% 的比例划分为训练集划分和测试集划分。平均回合长度为 $T \approx 1200$。训练集划分/测试集划分的统计数据如下表所示。

| | 试验 | |
|---|---|---|
| 布局 | 训练 | Test |
| 非对称优势 | 14 | 5 |
| 协调环 | 13 | 5 |
| Counter Circuit | 13 | 5 |
| 狭窄房间 | 14 | 5 |
| 强制协调 | 9 | 3 |

对于每种布局，我们分别在训练集划分和测试集划分上训练两个行为克隆模型。在测试集划分上训练的留出行为克隆模型仅用于评估，不会用于训练。训练所用的超参数如下表所示。

| 超参数 | 值 |
|---|---|
| 隐藏层大小 | [64, 64] |
| MLP层数量 | 2 |
| 学习率 | 0.001 |
| 批量大小 | 64 |
| 训练轮数 | 100 |

表 3：BC人类代理模型的超参数

### 8.5 人机协作评估

**参与者：**我们通过 Prolific 招募了 $N = 73$ 名参与者，招募方案经本校机构审查委员会批准。参与者年龄为 19–72 岁（中位数为 = 28.5 岁），其中 43.1% 自我认同为男性，55.6% 自我认同为女性，1.4% 未透露。实验开始前，参与者使用 5 点李克特量表报告了自己对电子游戏 Overcooked 的熟悉程度（1 = 没有经验， 5 = 非常熟悉）。实验时间的中位数为 34 分钟，每位参与者均因付出的时间和精力获得 11 美元（平均每小时 19.46 美元）的报酬。

**流程：** 完成一轮教程后，每位参与者会被随机分配到两种布局之一：Counter Circuit 或协调环。随后，参与者与每种受测人工智能方法各进行一个回合，每位参与者总共进行 6 个回合。对于每种方法，均从由 8 个经过训练的智能体组成的池中随机抽取一个智能体（每个智能体使用不同的随机种子初始化），以控制种子层面的方差。每个回合持续 200 个时间步。

**回合后问卷：** 每个回合结束后，参与者使用 5 点李克特量表（1 = 非常不同意， 5 = 非常同意）对以下陈述评价自己的体验 [11]：

- 该智能体在做决策时会适应我。
- 该智能体的行为始终如一。
- 该智能体的行为像人类。
- 该智能体经常妨碍我。
- 智能体的行为令人沮丧。
- 总的来说，我很喜欢和智能体一起玩。
- 总的来说，我觉得智能体与我协调配合的能力.....

### 8.6 Overcooked 按布局分类的评估结果

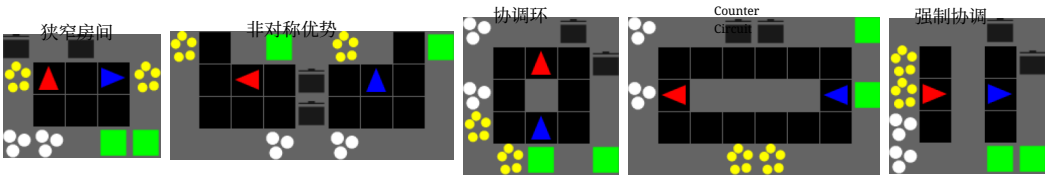


图 14：原始 5 个 Overcooked 环境的可视化。在 Overcooked 中，两名厨师需要相互合作，高效地制作并运送汤。

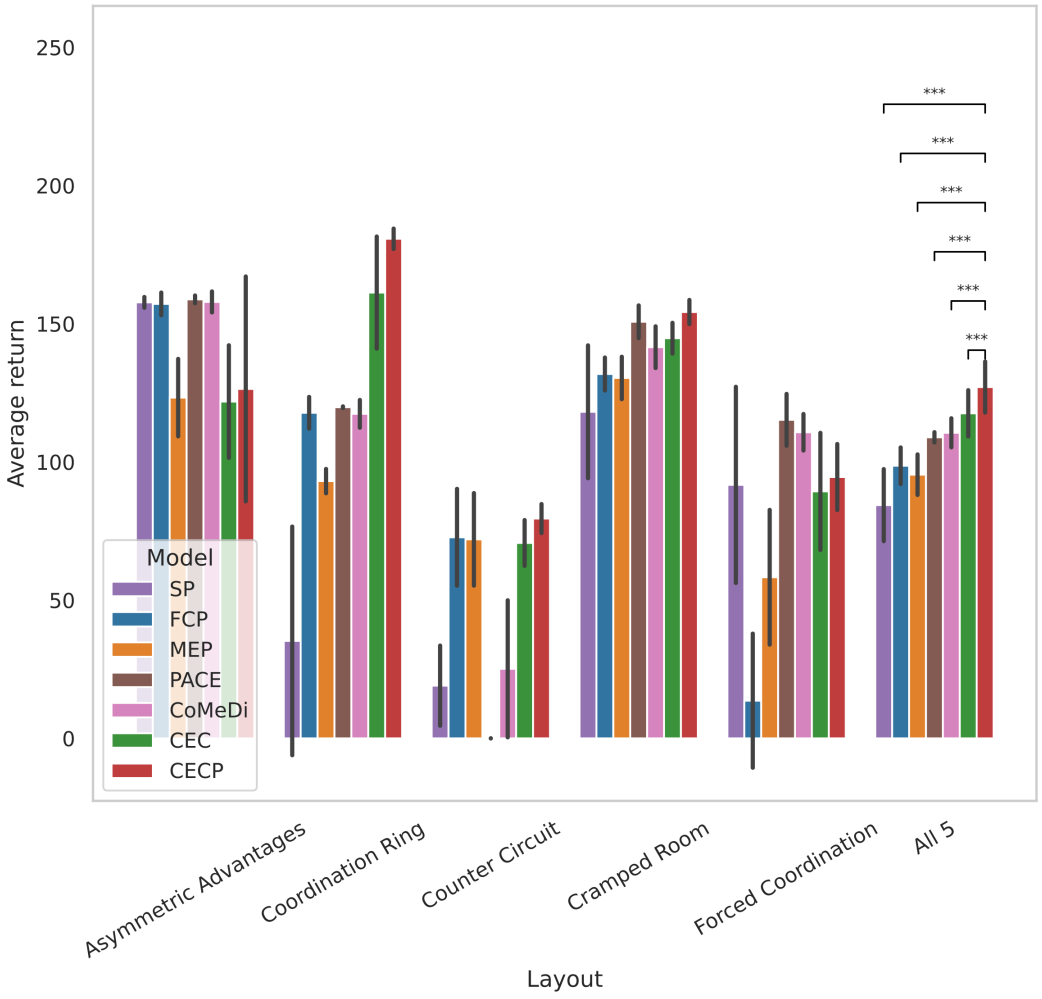


图 15：按原始 5 个 Overcooked 布局划分的交叉对战平均回报

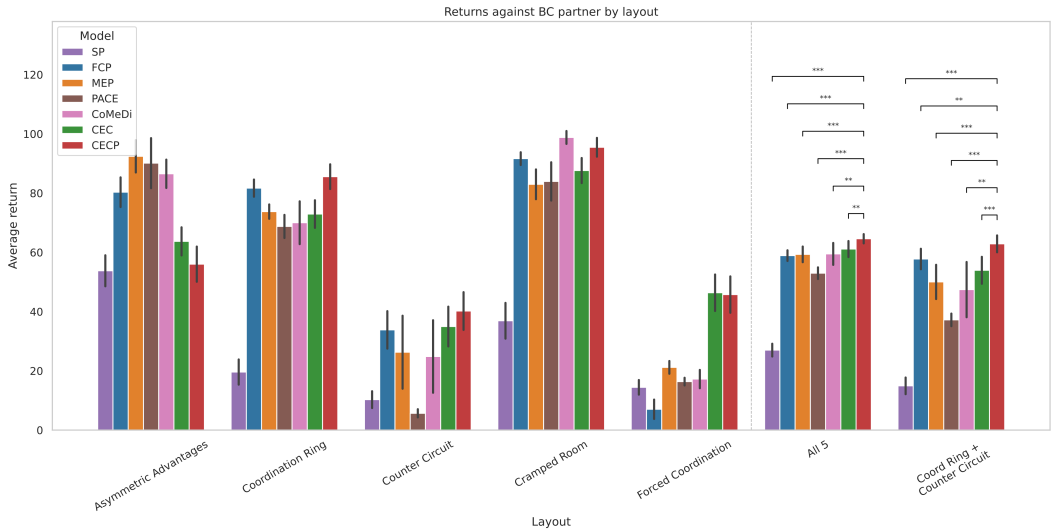


图 16：与人类代理合作时的平均回报，按原始 5 个 Overcooked 布局划分

### 8.7 伙伴动作预测

为评估智能体是否隐式地对其伙伴的行为进行建模，我们将伙伴动作预测准确率作为所学习伙伴表征的代理指标。在每个时间步，我们向 CECP 合作智能体查询其在以自身采样动作为条件时预测的下一状态。随后，通过考察预测的下一状态与当前状态之间伙伴位置的差异，推断伙伴的预测动作。

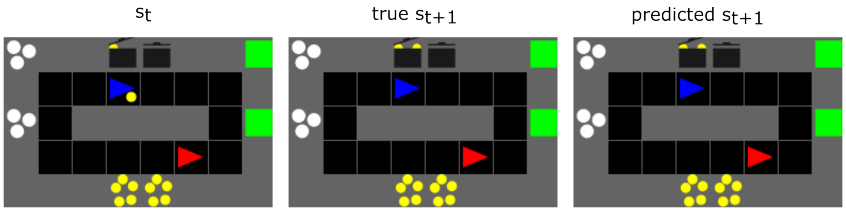


图 17：CECP 智能体（红色）与伙伴（蓝色）协作时的下一状态预测示例。在状态 $s_t$（左）下，为了完整预测真实的下一状态 $s_{t+1}$（中），CECP 的预测（右）必须同时捕捉蓝色伙伴将洋葱放入锅中的动作，以及锅中洋葱数量增加 1 的环境状态转移。

### 8.8 潜在表征分析

**伙伴检查点解码：** 我们评估一个经过训练的 CECP 智能体，该智能体分别与处于早期/中期/晚期检查点阶段的伙伴配对。对于每个伙伴，我们运行 20 个回合，并收集 CECP 智能体潜在 $z$ 层中的逐时间步激活值。在每个时间步，使用所有伙伴和回合的数据，在完整潜在向量上拟合一个多类线性判别分析 LDA 分类器，以对检查点阶段进行分类。我们报告使用 5 折交叉验证得到的平均测试分类器准确率。

**PCA 分析：** 将从 CECP 智能体的潜在 z 活动中收集的矩阵重塑为 (n_units，n_时间步 $\times$ n_回合)，按单元进行均值中心化，并通过完整奇异值分解进行分解。将回合轨迹投影到前两个主成分构成的空间中，并绘制为按伙伴标签（早期与晚期）着色的时间索引曲线，同时绘制平均轨迹。

### 8.9 CECP 智能体会根据处于不同训练阶段的伙伴调整策略吗?

在图10中，我们观察到可以从 CECP 智能体的潜在层解码出伙伴训练阶段。此处，我们还评估 CECP 智能体是否会针对处于不同训练阶段的伙伴调整其策略。令 $A$ 表示 CECP 智能体的动作， $Z \in \{$早期，中期，晚期$\}$ 表示伙伴检查点的训练阶段， $S$ 表示环境状态。我们收集了 8000 条 CECP 智能体与来自各训练阶段的伙伴协作时的回合轨迹 $z$，并选取 CECP 智能体与每个阶段的伙伴协作时访问次数超过 10 次的状态 $s$。为进一步确保 CECP 智能体有足够时间推断伙伴训练阶段，我们只考虑回合开始后前 10 个时间步之后访问的状态。对于每个此类状态和伙伴阶段，我们计算 CECP 智能体的平均动作分布 $\pi(a \mid s, z)$，并据此计算智能体动作与伙伴阶段之间的条件互信息，其中将 $Z$ 视为在三个阶段上均匀分布：

$$I(A; Z \mid S = s) = \frac{1}{3} \sum_{z} \mathrm{KL}\left(\pi(a \mid s, z), |, \bar{\pi}(a \mid s)\right), \qquad \bar{\pi}(a \mid s) = \frac{1}{3} \sum_{z} \pi(a \mid s, z). \quad (9)$$

对于原始的 5 种 Overcooked 布局，在选定状态上均匀取平均得到的互信息显著高于打乱的零分布。该零分布通过在轨迹间随机置换伙伴训练阶段标签并重新计算平均互信息生成，重复 1000 次（图18）。

当我们考察互信息较高的特定状态时，发现 CECP 智能体会根据伙伴所处的训练阶段调整其行为策略。例如，图19a展示了狭窄房间中的一个 CECP 智能体（红色），当伙伴（蓝色）处于训练早期阶段时，它往往会避开伙伴（更多地向左移动，因为伙伴在右侧）。此外，图19b展示了协调环中的一个特定场景：上方的锅里已经有两个洋葱，右侧的锅里有一个。如果伙伴训练得较充分，CECP 智能体会优先把自己的洋葱放进右侧的锅里，因为它预计伙伴也会取来一个洋葱并放进锅里。然而，如果伙伴训练得较少，CECP 智能体则更倾向于向上移动，把自己的洋葱放进上方的锅里，以开始烹饪。

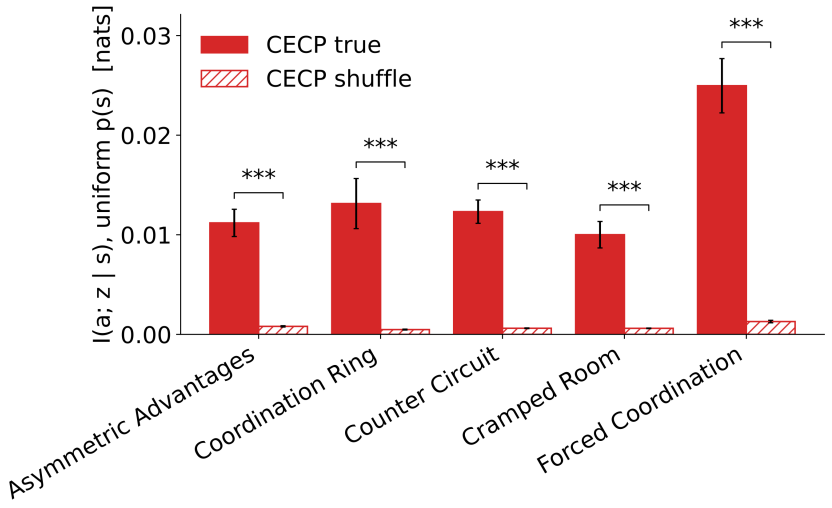


图18：对于五种原始 Overcooked 布局中的每一种，在选定状态上均匀取平均后，CECP 智能体的动作与伙伴训练阶段之间的平均互信息。实心柱：真实的伙伴阶段标签。斜线柱：打乱的零分布。柱形显示群体中 8 个 CECP 智能体的均值；误差线表示均值标准误。 ***： $p < 0.001$ ，通过置换检验；适用于每个布局中的每个智能体

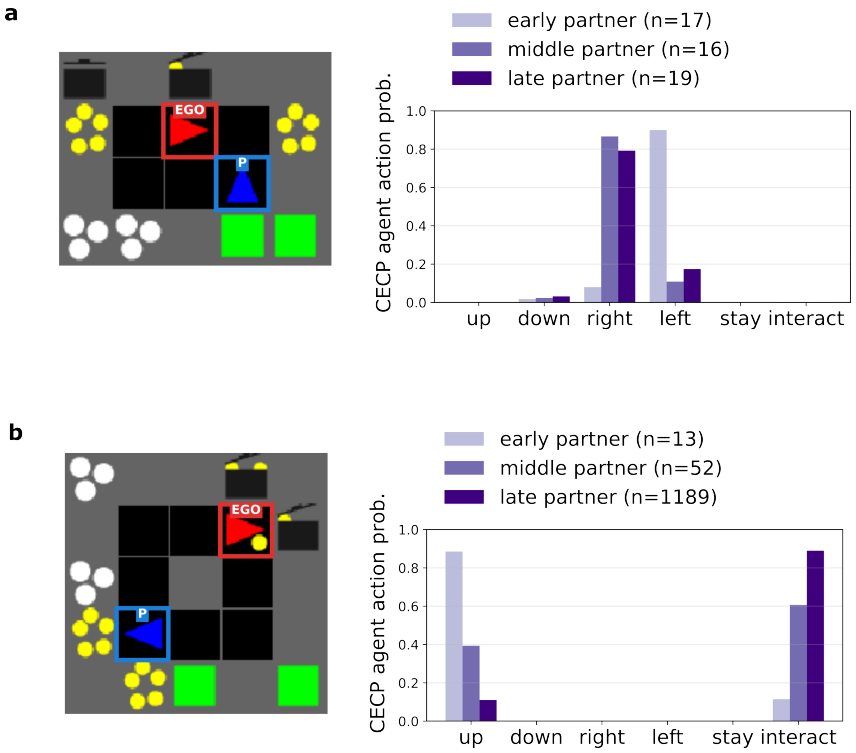


图19：伙伴处于训练早期/中期/后期阶段时，CECP 智能体在特定状态下的平均动作分布。狭窄房间 (a) 和协调环 (b) 的示例状态。左侧为状态示意图；处于不同训练阶段的 CECP（自体）智能体以红色表示，伙伴智能体以蓝色表示。右侧为 CECP（自体）智能体在每个伙伴训练阶段下，于此状态中的平均动作分布。

#### 8.10 大语言模型使用情况

我们借助 Claude Code 协助实现所提出的方法和分析，并帮助撰写和编辑这篇稿件。