frontiers

用于高效多智能体合作的人格混合增强脉冲行动者网络

Xiyun Li[1,2], **Ziyi Ni**[1,3], **Jingqing Ruan**[1,2],**LinghuiMeng**[1,3],**JingShi**[1,3], **TielinZhang**[1，3，*] **和 Bo Xu**[1，2，3，4，*]

[1]中国科学院自动化研究所，中国北京[2]中国科学院大学未来技术学院，中国北京[3]中国科学院大学人工智能学院，中国北京[4]中国科学院脑科学与智能技术卓越创新中心，中国上海

通讯作者*：Tielin Zhang 和 Bo Xu
tielin.zhang@ia.ac.cn，xubo@ia.ac.cn

arXiv:2305.05898v1 [cs.AI] 10 May 2023

摘要

在人类与智能体、智能体与智能体的合作中实现自适应，正日益成为多智能体强化学习（MARL）研究领域的关键问题；借助深度神经网络，该领域已取得显著进展。然而，许多现有算法只能在学习范式中表现良好，与其他未见过的合作伙伴协作时却缺乏泛化能力。认知心理学中的人格理论指出，人类能够通过先预测他人的人格，再推断其复杂动作，从而妥善应对上述合作挑战。受这一两步心理学理论的启发，我们提出一种具有生物学合理性的人格混合（MoP）增强脉冲演员网络（SAN）。该方法使用行列式点过程模拟 MoP 中不同类型人格的复杂形成与整合，并将动态脉冲神经元融入 SAN，以实现高效的强化学习。我们选取对协同烹饪要求很高的基准任务 Overcooked，来测试所提出的 MoP-SAN。实验结果表明，MoP-SAN 不仅在学习范式中表现出色，在泛化测试范式（即与其他未见过的智能体合作）中也能取得优异表现，而大多数对照深度演员网络在该范式中均表现不佳。我们还进行了必要的消融实验和可视化分析，以解释 MoP 和 SAN 在多智能体强化学习场景中为何有效，以及深度神经网络（DNN）在泛化测试中为何表现不佳。

关键词：多智能体合作、人格理论、脉冲演员网络、多智能体强化学习、心智理论。

1 引言

随着近年来深度强化学习（强化学习）的迅速发展和巨大进步（Vinyals 等人，2019年；Silver 等人，2017；Vaswani 等人，2017；Yu 等人，2021；Meng 等人，2023），越来越多的研究人员对多智能体合作或人在回路中的合作表现出浓厚兴趣。

（Shih 等人，2021；Strouse 等人，2021；Carroll 等人，2019；Zhao 等人，2021；Shih 等人，2022；Ruan 等人，2022；Lou 等人，2023）。然而，与未见过的合作伙伴合作通常需要持续收集专家数据，这既昂贵又耗时（Carroll 等人，2019；Shih 等人，2022）。其他方法尝试通过构建群体池来模拟多样化的候选合作伙伴，从而在不使用专家数据的情况下实现更好的泛化。然而，这些研究试图通过依赖于与大量精心设计的合作伙伴进行训练来提高泛化合作得分，却忽视了对智能体真实思维能力和共情能力的培养。

对伙伴智能体心理特征的考虑不足，可能是这些人工智能体不如其对应的生物智能体的关键原因。在日常生活中，人类可以与从未见过的人良好合作（Boyd and Richerson，2009；Rand and Nowak，2013）。这一现象很有趣，但并不难以理解。我们能够迅速推断他人的个性，并借助推测出的人格妥善处理接下来的合作行为。人格理论属于心智理论（ToM）的框架（Gallagher and Frith，2003；Frith and Frith，2005；Aru 等，2023；Roth 等，2022）。心智理论指的是我们推测他人意图、行为和目标的能力，它从认知视角解释了人类为何能够与未见过的合作伙伴协作。事实上，与其将未见过的人归为某种特定人格，不如将其视为几种“人格”的组合。因此，尽可能少地找出具有代表性的人格，并使它们彼此正交，对实现更高效的组合大有裨益。认知心理学中的人格理论（Ryckman，2012；Schultz and Schultz，2016；McCrae and Costa Jr，2008）为更清晰、具体地建模合作伙伴提供了契机，其中包括大五人格（De Raad，2000）和十六种人格因素（16PF）（Cattell and Mead，2008）。这些理论有助于描述独特而多样的人群（Anglim and Horwood，2021），并能指导许多认知任务，例如人格特质测试（O’Connor and Paunonen，2007），以分析人们适合的职业。

不同于认知科学中常用于离散分类的人格理论，我们提出了类似于人格空间中基向量的基准人格，可用于推断个性。为进一步确保多个基础人格之间的差异，我们采用行列式点过程（DPP）约束作为内在奖励。基于包含这些基础人格的人格模型，智能体能够自然地预测并理解任何未曾见过的合作伙伴，从而更好地做出响应并实现合作。

因此，受上述人格理论启发，我们提出了人格混合（MoP），并结合我们此前提出的脉冲智能体网络（SAN）；该网络已被证明在单智能体强化学习中效率较高（Zhang 等，2022）。SAN 在生物学上具有合理性，包含更多动态神经元；在动态强化学习任务中，这些神经元展现出优势，能耗更低且泛化能力更强。在此，我们进一步将 SAN 应用于多智能体强化学习协作场景。我们的主要贡献可概括如下：

1. 我们首次提出了 MoP 概念，该概念受到心理学人格理论的启发，描述了一种两步预测：首先，人格估计器（PE）接收上下文，并在 DPP约束下估计合作伙伴的个性；然后由多人格网络进行行为预测。

2. 我们引入高效的 SAN 和 MoP 模型，以实现多尺度生物合理性，其中具有神经动力学的脉冲神经元已被证实在类强化学习任务中具有较高的效率（Zhang 等，2022），

此外，我们进一步尝试将神经元尺度的动态与合作伙伴尺度的合作相结合，以提升智能体在多智能体协作中的泛化能力。

3. 随后，我们在 Overcooked 基准环境中测试了所提出的 MoP-SAN。实验结果表明，与其他深度神经网络基线相比，该方法的泛化能力明显更强，尤其是在与其他未见过的合作伙伴协作时。这意味着，我们提出的算法能够在零样本协作测试中成功推断未见过的合作伙伴的个性。我们还进行了分析实验，以探究 SAN 方法的泛化效果为何优于深度神经网络基线。

2 相关工作

强化学习是机器学习中的一种重要范式，也适用于许多序贯决策任务。近年来，强化学习方法在许多任务中取得了良好成果（Vinyals 等人，2019年；Silver 等人，2017、2018）。现有的传统强化学习方法可分为基于价值的方法（Mnih，2013年）和基于策略的方法（Schulman 等人，2015）。演员-评论家方法的提出在强化学习领域具有里程碑意义，它结合了基于价值的方法和基于策略的方法的优点。近端策略优化（PPO）（Schulman 等人，2017）是这一框架中最经典的方法之一，在许多任务中均取得了令人瞩目的表现，例如控制任务（Schulman 等人，2017）和星际争霸（Yu 等人，2021）。

多智能体强化学习描述了多智能体从零开始学习策略的过程，旨在通过与环境依次或同时交互，最大化全局奖励。例如，在双人合作任务 Overcooked 中，自我智能体和伙伴智能体需要协同合作，以最大化从 Overcooked 环境中获得的团队奖励。在多智能体强化学习中，合作式多智能体强化学习任务是一个极具挑战性的方向。尽管已有一些研究探索如何解决合作式多智能体强化学习任务中的挑战性问题，例如信用分配（Rashid 等人，2020年；Sunehag 等人，2017），但如何设计能够泛化到未见过的合作伙伴的模型仍然具有挑战性。针对多智能体合作，近期一些研究（Shih 等人，2021；Carroll 等人，2019年；Zhao 等人，2021；Strouse 等人，2021；Shih 等人，2022；Lou 等人，2023）重点研究未见过的合作伙伴的泛化问题。尽管传统的自我博弈方法（Silver 等人，2018）取得了显著优势，并且往往能够在竞争性博弈中收敛至最优均衡策略，但在合作任务中，它们往往会对特定合作伙伴过拟合。一些研究尝试通过模仿学习解决过拟合问题（Carroll 等人，2019年；Shih 等人，2022），尽管据报道，在许多真实场景中收集专家数据颇具挑战。为提升人机协作的泛化能力，研究者提出了模块化方法，明确分离依赖约定的表征与依赖规则的表征（Shih 等人，2021）。其他研究（Strouse 等人，2021；Zhao 等人，2021）则尝试通过设计包含众多精心设计的标准和智能体的各种群体池，解决与未见过的合作伙伴共同完成合作任务的问题。

由于类脑脉冲神经网络在许多方面都具有优势，许多研究开始使用脉冲神经网络解决强化学习问题（Florian，2007年；Tang 等人，2020年；Zhang 等，2022；Bellec 等人，2020年；Patel 等人，2019年；Fr´emaux 等人，2013年）。我们此前的工作提出了一种多尺度动态编码改进脉冲演员网络（MDC-SAN），用于单智能体场景，以实现高效决策（Zhang 等，2022）。与大多数在单智能体强化学习任务中探索脉冲神经网络方法的研究不同，本文旨在将脉冲神经网络方法应用于多智能体协作任务。在这项工作中，我们需要在合作任务中与不同风格的合作伙伴协作，因此构建用于伙伴建模的模型至关重要。

心智理论（Gallagher and Frith，2003；Frithand Frith，2005；Aru 等，2023；Roth 等，2022）是认知心理学中的一个基本概念，它使个体能够预测并解释他人的行为、有效沟通并更好地参与合作交互，这也是我们希望人工智能智能体实现的目标。一些研究设计了心智理论模型（Yuan 等，2022；Wang 等，2021；Tabrez 等，2020），用于解决强化学习任务。借助心智理论模型，智能体可以更高效地与其他合作伙伴沟通，并学习合作伙伴的一些惯例。在一些研究中（Rabinowitz 等，2018；Roth 等，2022），心智理论模型旨在理解其他智能体的行为，这对许多强化学习任务至关重要。虽然心智理论涵盖许多方面，包括心理模拟、动作预测和推理，但在此上下文中，我们将聚焦于一个特定方面——人格特质，以增强智能体模型。

3 方法

3.1 双人合作的问题设定

我们可以将这个双人马尔可夫博弈定义为一个元组 $\left(\mathcal{O}, \mathcal{A}, \mathbb{P}, \gamma, \pi, \rho^i, r, m\right)$，其中 $\mathcal{O}$ 表示观测空间，$\mathcal{A}$ 表示自我智能体和合作伙伴共享的动作空间。我们可以将 o $=(o^1, o^2)$ 定义为包含自我智能体观测和合作伙伴观测的内容。我们可以将标签 a $=(a^1, a^2)$ 表示为所有玩家的联合动作，包括自我智能体的动作和合作伙伴的动作。$\mathbb{P}: \mathcal{O} \times \mathcal{A} \rightarrow \mathcal{O}$表示环境转移概率函数，$\gamma \in [0,1)$ 是折扣因子。π 是联合策略，而自我智能体 ρ^1 的策略是我们 MoP-SAN 中 SAN 智能体的脉冲策略，ρ^2 表示合作伙伴的策略。所有智能体共享同一个团队奖励函数 r(o, a) : o $\times$ a $\rightarrow R$。$\tau =$ $(\mathrm{o}_0, \mathrm{a}_0, \mathrm{o1}, ...)$ 表示由联合策略 π生成的轨迹，而 $\tau^2 = (o_0^2, a_0^2, o_1^2, ...)$ 是合作伙伴的轨迹。MoP模型 m 可以基于合作伙伴的历史轨迹信息对其进行建模，并为 SAN 智能体提供可执行指导。在每个时间步，SAN 智能体感知观测 $o_t^1 \in \mathcal{O}$，并从 MoP模型 m接收引导动作 $\hat{a}_t^2$，然后从脉冲策略 $\rho^1: \mathcal{O} \times \mathcal{A} \rightarrow [0,1]$中抽取动作 $a_t^1 \in \mathcal{A}$，记作 $a_t^1 = \rho^1(\cdot|o_t^1, \hat{a}_t^2)$。合作伙伴的策略可以表示为 $a_t^2 = \rho^2(\cdot|o_t^2)$。SAN 智能体和合作伙伴以概率 $\mathbb{P}(\mathrm{o}_{t+1} \mid \mathrm{o}_t, \mathrm{a}_t)$ 进入下一个状态 o_{t+1}，并从环境中获得数值奖励 r_{t+1}。所有智能体共同协调，以最大化累积折扣回报 $\mathbb{E}_{\boldsymbol{\tau} \sim \pi}\left[\sum_{t=0}^{\infty} \gamma^t r(\mathrm{o}_t, \mathrm{a}_t)\right]$。

我们假设，在完全合作博弈中，至少存在一个联合策略，使所有智能体都能获得最大累积奖励。问题、目标陈述和我们的方法将在以下各节中形式化。

3.2 MoP-SAN 的算法架构与流程

上一节定义了合作式多智能体强化学习问题。本节将介绍我们针对学习阶段和泛化阶段提出的算法架构与流程。本文提出了一个用于多智能体协作的鲁棒框架。图1左侧展示了我们实验中的两个阶段，我们将在下一节中讨论这两个阶段。图1右侧展示了 MoP-SAN 在零样本协作中的流程，图2则说明了 MoP-SAN 的详细结构。

如图1和图2所示，考虑到生物学合理性和能源效率，我们提出的框架包含一个作为自我智能体的 MoP模型和 SAN模型。MoP作为伙伴心智模型，能够理解合作伙伴的行为并对其进行建模，首先估计合作伙伴的个性，然后指导SAN智能体的动作。SAN智能体能够更好地提升泛化

能力，以应对合作伙伴异质性（与多样化的未见过的合作伙伴进行零样本协作），并在MoP模型的帮助下与未见过的合作伙伴协作 m。

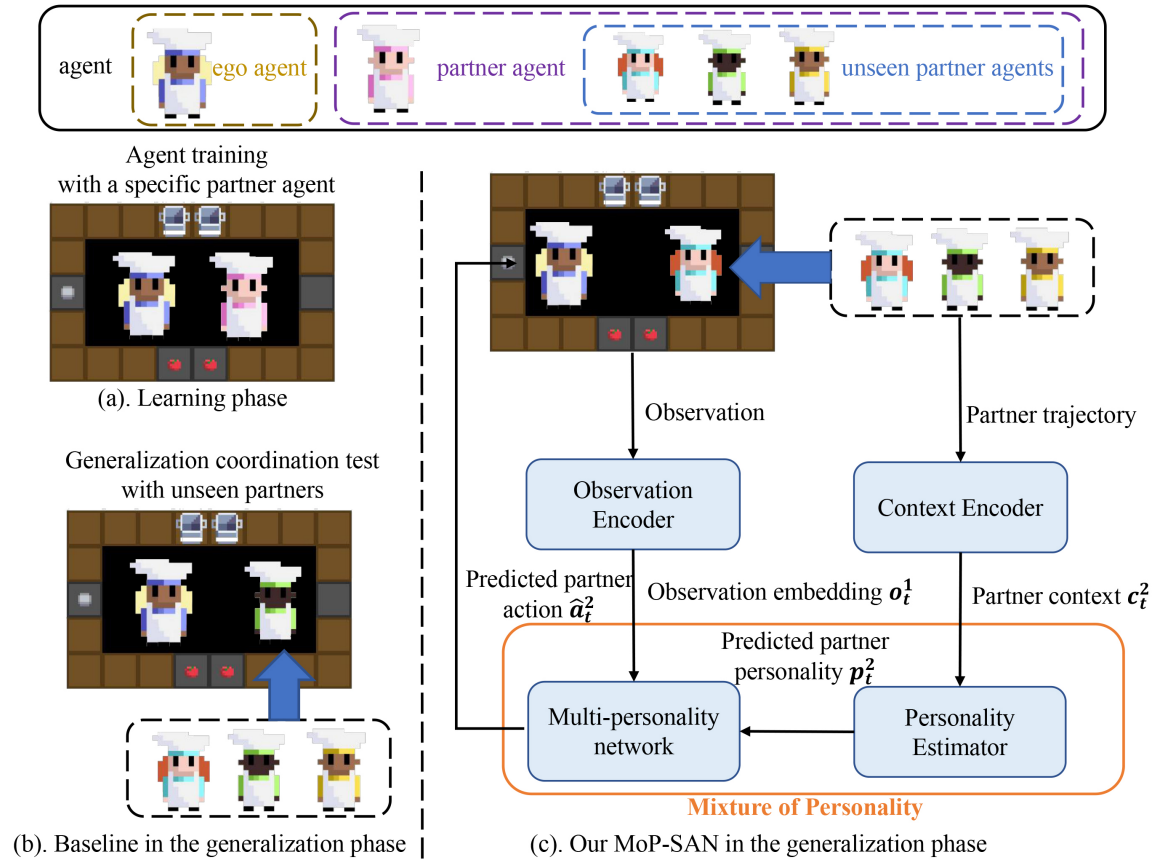


图1。 我们提出的 MoP-SAN 的学习阶段和泛化阶段。（a）在学习阶段，自我智能体与特定合作伙伴智能体协作完成 Overcooked任务。（b）在泛化阶段，自我智能体需要与一些未见过的合作伙伴智能体协作，以测试泛化能力（零样本协作）。（c）我们的 MoP-SAN 也需要与未见过的合作伙伴进行协作。通过构建MoP模型，我们可以首先估计合作伙伴的个性，并根据合作伙伴的个性预测其动作。三幅图中位于同一厨房内的两个智能体代表这两个智能体之间为完成该烹饪任务而建立的协作关系。

如图1所示，我们可以将整个过程分为学习阶段和泛化阶段，也称为训练和测试过程。我们提出了一个在学习阶段不需要额外专家监督数据的通用框架。在当前模型中，为了简化起见，我们假设观测编码器是恒等映射，并将来自环境的观测作为 MoP 的输入。为了在不使用额外专家数据的情况下对 MoP模型进行自监督训练，出于简化考虑，我们直接在学习过程中将 MoP 训练为一个合作伙伴。

一方面，MoP模型可以作为由众多不同智能体组成的池，促进SAN智能体的学习。另一方面，MoP模型还可以学习各种人格。在泛化阶段，我们希望进行更好的推断，并适应具有特定人格的未见合作伙伴，因此需要在学习过程中尽可能多地发掘人格空间中的基础人格。

在泛化阶段，我们框架中的参数是固定的。如图1所示，当 SAN 智能体需要与一个未见过的合作伙伴协作时，人格估计器（PE）首先根据该合作伙伴的历史上下文信息确定其个性，然后多人格网络推断合作伙伴当前的意图和动作。我们的目标是基于未见过的合作伙伴的历史信息，最大化总奖励和熵。在以下各节中，我们以泛化阶段为例描述我们的方法及其公式。MoP模型的输出是 SAN 的脉冲策略的输入，ρ_θ^1 和 θ^1 是 SAN 中策略网络的参数。φ 和 η 是 MoP模型的参数，联合策略可写为:

$$\pi\left(\mathbf{a}_t \mid \mathbf{o}_t\right)=\rho_\theta^1\left(a_t^1 \mid o_t^1, \hat{a}_t^2\right) \rho_\theta^2\left(a_t^2 \mid o_t^2\right), \tag{1}$$

其中，o_t^i 是第 i个玩家的观测，$\hat{a}_t^2$ 表示我们的 MoP 模型预测的动作分布。

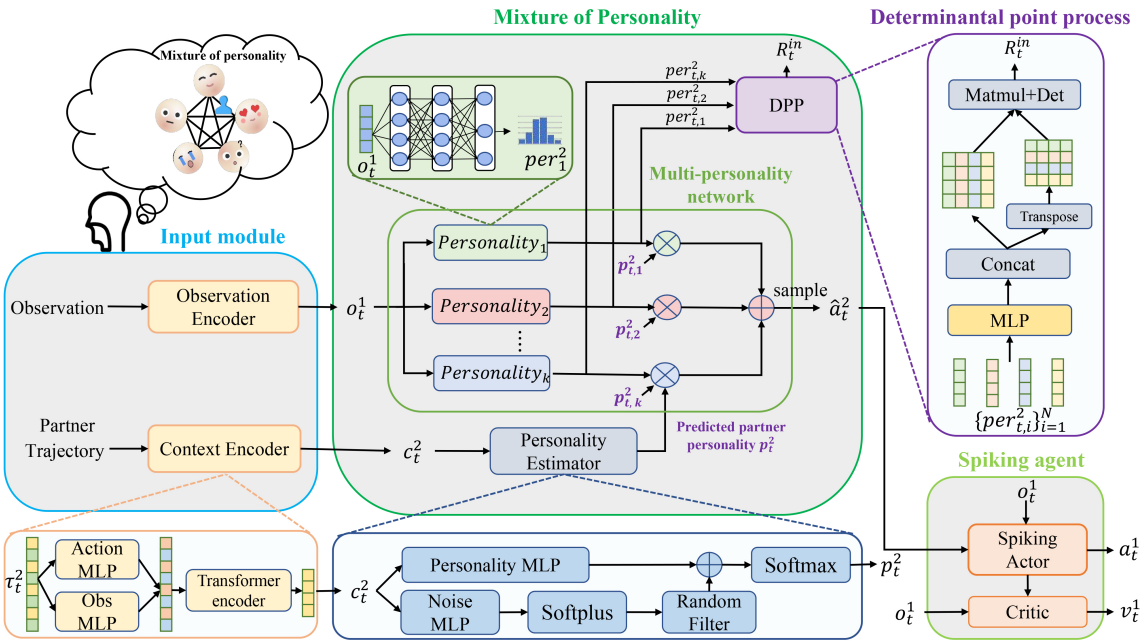


图2. MoP-SAN的详细结构。MoP-SAN由一个SAN智能体、一个MoP模型和一个输入模块组成，输入模块包括上下文编码器和观测编码器。SAN PPO用于通过MoP模拟自我智能体。MoP模型用于模拟自我智能体的心智理论过程，以对未见过的合作伙伴的个性进行建模。我们的MoP模型包含人格估计器（PE）模块、多人格网络和DPP模块。

3.3 SAN模型和上下文编码器

我们的 MoP-SAN 中的 SAN 模型指的是一个 SAN PPO 智能体，它根据 MoP 模型的引导动作来执行动作，以最大化合作奖励和熵。输出动作a_t^1 从 SAN 模型中脉冲策略在动作空间上的概率分布中采样得到 $\rho_\theta^1\left(a_t^1 \mid o_t^1,\ \hat{a}_t^2\right)$。SAN PPO 智能体包括一个脉冲演员和一个评论家。SAN 模型由漏积分发放（LIF）神经元组成，这是对霍奇金-赫胥黎模型的抽象。LIF 神经元的非微分膜电位和不应期是符合生物学实际的特征，

这些特征可以模拟神经元动力学。我们将 LIF 神经元定义如下：

$$\tau \frac{dV_t}{dt} = -V_t + I_t, \tag{2}$$

其中， V_t 表示时间 t 时膜电位的动态变量， dt 是最小模拟时间片。 I_t 表示积分后的突触后电位， τ 是积分时间段。输入 I_t ，持续时间为 τ。当 V_t 大于发放阈值 V_{th}时，神经元将被激活并产生一个脉冲，膜电位 V_t 将重置为重置电位 V_{reset}。当 V_t 小于发放阈值时，神经元大多处于漏电状态。SAN 的详细配置见我们之前的工作（Zhang 等，2022）。

上下文编码器是我们具备良好泛化和适应能力的关键。我们使用 Transformer 模型作为上下文编码器，其输入是特定上下文大小下合作伙伴的历史轨迹，作为上下文信息。对于上下文信息，历史动作和观测具有不同的维度。因此，我们引入动作 MLP 网络和观测 MLP 网络，将历史动作和观测转换为相同维度，并按照轨迹 τ 中的时间 t 顺序交替拼接，类似于（Chen 等，2021；Meng 等，2023）。

3.4 MoP模型

我们的MoP-SAN的心智理论能力由MoP模型 m实现，该模型由多人格网络、PE模块和DPP模块组成。

多人格网络包含 k 个不同的人格网络，每个人格网络由三层MLP组成，分别以不同的策略表示一类基准人格。多人格网络的输入是SAN智能体的观测，而 i个人格网络的输出$per^2_{t,\ i}$则是在相同环境观测下与相应基准人格对应的动作分布。

PE模块的输入是合作伙伴的上下文信息 c^2_t，即通过上下文编码器从合作伙伴的历史轨迹中获得的上下文嵌入。与完全理性的AI智能体不同，未见过的合作伙伴会受到一些影响其决策的非理性因素的影响。因此，我们的PE模块由人格多层感知器（MLP）和噪声MLP组成，前者由可训练的权重矩阵 W_p 表示，后者由 W_{noise}表示。噪声MLP的输出经过softplus函数和随机滤波器处理后，加到人格MLP的输出上。所得和再经过softmax函数，以获得 p^2_t 未见过的合作伙伴的估计人格画像。 e 表示PE函数， R 表示随机滤波器函数：

$$e(c^2_t) = \text{Softmax}\left(c^2_t \cdot W_p + R(c^2_t \cdot W_{noise})\right), \tag{3}$$

MoP模型 $\hat{}a^2_t$ 从动作空间上的概率分布中采样得到，$\eta\left(\hat{}a^2_t \mid o^1_t,\ c^2_t\right)$。PE模块 p^2_t的输出对应于预测的合作伙伴个性，而第 i 个人格网络 $per^2_{t,\ i}$ 的输出表示当前观测下第 i 个基准个性的动作空间上的概率分布。 η 是 MoP 中 DPP 的参数， φ 是

MoP模型的参数。我们的 MoP 策略可以定义如下 :

$$m_{\varphi,\eta}\left(\hat{a}_t^2 \mid o_t^1, c_t^2\right) = \sum_{i=1}^{n} p_{t,i}^2 \cdot per_{t,i}^2, \tag{4}$$

与针对不同任务选择不同分支的稀疏激活模型不同，我们的 MoP 方法整合了所有基础人格的输出，而不是每次都选择一个基准人格。因此，PE 模块的输出，即对合作伙伴人格的预测，并非离散的独热向量，而是一个总和为 1 的浮点向量。

我们的 MoP 可以对合作伙伴进行建模，并推断其他合作伙伴的人格，从而帮助任意强化学习智能体提升其泛化能力和适应性，使智能体能够应用于许多零样本协作场景。

3.5 MoP中的DPP模块

在本节中，我们首先介绍 DPP，然后介绍我们提出的 MoP-SAN 中的 DPP。DPP（Kulesza等，2012年）是随机矩阵理论中提出的一种高效概率模型，已广泛应用于机器学习的诸多领域（Gong等，2014年；Perez-Nieves等，2021年；Parker-Holder等，2020年），例如推荐系统（Chen等，2018年）和视频摘要（Gong等，2014年）。高性能模型 DPP 可以将复杂的概率计算转化为简单的行列式计算，然后利用核矩阵的行列式计算各个子群的概率。近期研究（如 Dai等，2022年；Yang等，2020年）已将 DPP 模型融入强化学习方法中。Dai等（2022年）利用 DPP 模型引入内在奖励，以增强强化学习方法的探索能力。与此同时，Yang等（2020年）利用 DPP 鼓励强化学习进化算法中的智能体保持多样性，从而改进现有的强化学习算法。

在学习过程中，多人格网络可以被视为具有多种人格。每个人格网络都可看作一种独特的基准人格。衡量多个基础人格之间的多样性，对于在人格空间中构建多样化的基础人格集合至关重要。为了有效探索任务空间中的人格范围，我们在 MoP-SAN 中集成了促进多样性的 DPP 模块，以对这些基础人格进行正则化。这有助于高效探索和优化多样化的人格集合，从而提升 MoP-SAN 的整体性能。

我们可以通过 DPP 模块施加的多样性约束这一内在奖励，来衡量人格的多样性，并选出多样性人格子集。Y表示一个包含许多个性的集合， y 表示 Y的一个子集，其中包含能够使多样性最大化的 k 个人格。由于这些人格网络共享相同的观测输入，且特定人格网络$per_{t,\ i}^2$ 的输出是一个动作分布，因此可以根据动作空间上的动作分布来衡量基础人格之间的差异。我们将 y 的核矩阵记为 L_y。 L_y 的行列式值可以表示人格集合 y的多样性。为了构造集合 y，我们需要从人格空间中选出 k个人格，使 y的核矩阵行列式值最大化。人格集合 y 可以视为人格空间中使多样性最大化的一组基础人格。

$$y^* = \arg\max_y P(Y = y) = \arg\max_y \det\left(L_y\right), \tag{5}$$

由于 m　矩阵 L_y 是半正定的，因此在每个时间步 t 都存在矩阵 B_t，使得　t:

$$L_y = B_t B_t^T, \tag{6}$$

B_t 且内在奖励 r_t^{dpp} 可定义如下，其中 k 为人格数量:

$$B_t = \left[\upsilon_\eta\left(per_{t,1}^2\right), \upsilon_\eta\left(per_{t,2}^2\right), \upsilon_\eta\left(per_{t,3}^2\right), \ldots, \upsilon_\eta\left(per_{t,k}^2\right)\right]^T, \tag{7}$$

$$r_t^{\text{dpp}}\left(per_{t,1}^2, per_{t,2}^2 \ldots per_{t,k}^2; \eta\right) = \log\det\left(B_t {B_t}^T\right). \tag{8}$$

其中，υ_η 表示由参数 η参数化的特征向量。

我们致力于构建一些独特的人格向量作为多人格网络的基础人格，以组合整个人格空间。因此，我们提出的 DPP 模块与 MoP 模型能够在协作任务中快速适应任意未见过的合作伙伴，并实现泛化。

3.6 SAN学习

我们的方法对 SAN 智能体的策略参数 θ^1 和 MoP 模型参数（φ，η）进行迭代优化。总体优化目标是最大化累积折扣回报，该回报取决于 MoP 模型 $m_{\varphi,\ \eta}\left(a_t^2 \mid o_t^2,\ c_t^2\right)$ 以及 SAN 智能体的脉冲策略 $\rho_\theta^1\left(a_t^1 \mid o_t^1,\ a_t^2\right)$:

$$\theta^{1^*}, \varphi^*, \eta^* = \max_{\theta,\varphi,\eta} \sum_{t=0}^{\infty} \mathbb{E}_{a_t^1, a_t^2}\left[\gamma^t\left(r\left(\mathbf{o}_t, \mathbf{a}_t\right) + \alpha\overline{\mathcal{H}}\left(\pi\left(\mathbf{a}_t \mid \mathbf{o}_t\right)\right)\right)\right], \tag{9}$$

SAN 智能体的目标是通过与合作伙伴协作来最大化外在奖励 r_t^{ex}。我们可以如下计算 SAN 的梯度:

$$\begin{aligned} \nabla_\theta J\left(\rho_\theta^1\right) = \mathbb{E}_{a_t^1, a_t^2}\Big[&\nabla_\theta \log(\rho_\theta^1(a_t^1 | o_t^1, a_t^2))\big(G^{ex}(\mathbf{o}_t, \mathbf{a}_t) \\ &- b_1\left(o_t^1, \mathbf{a}_t\right) - \alpha\log(\rho_\theta^1(a_t^1 \mid o_t^1, a_t^2))\big)\Big] \\ a_t^1 \sim \rho_\theta^1\left(a_t^1 \mid o_t^1, a_t^2\right), &a_t^2 \sim m_{\varphi,\eta}\left(a_t^2 \mid o_t^2, c_t^2\right) \end{aligned}, \tag{10}$$

其中，b_1 是基线函数。我们可以通过所有可能动作的期望回报来估算基线函数 b_1，如下所示:

$$b_1\left(o_t^1, a_t^1\right) = \sum_{a_t^1 \in \mathcal{A}} \rho_\theta^1\left(a_t^1 \mid o_t^1, a_t^2\right) G^{ex}\left(\mathbf{o}_t, \mathbf{a}_t\right). \tag{11}$$

其中，$G^{\text{ex}}\left(\mathbf{o}_t,\ \mathbf{a}_t\right)$ 表示 SAN 和 MoP 的折扣外在回报。这里我们使用游戏得分作为外在奖励 r_t^{ex}。

3.7 MoP学习

我们在工作中引入 DPP 约束，与近期研究（Dai等，2022年）类似，将 DPP 多样性度量视为内在奖励。我们采用双层优化框架（Dai等，2022年）来优化 MoP模型及其 DPP模块，从而最大化内在奖励和外在奖励。

我们的目标可以定义如下：

$$\max_{\eta} J^{\mathrm{ex}}\left(\varphi', \eta\right) \text{ s.t. } \varphi' = \underset{\varphi}{argmax} J^{\mathrm{mix}}(\varphi, \eta), \tag{12}$$

对于这个优化问题，我们可以将其视为一个斯塔克尔伯格博弈。我们将 DPP 奖励作为内在奖励。混合奖励是内在奖励与外在奖励之和。混合奖励可以写作如下形式：

$$r_t^{\mathrm{mix}} = r_t^{\mathrm{ex}} + \beta r_t^{\mathrm{dpp}}\left(a_1, a_2 \ldots a_k; \eta\right), \tag{13}$$

其中，β 是内在奖励的权重系数。r_t^{ex}是环境提供的标准奖励，在时间步 t，SAN智能体在环境状态 s_t下执行动作 a_t^1，MoP执行动作 a_t^2，而 r_t^{dpp} 是针对合作伙伴的 DPP 约束多样性奖励。梯度 $\nabla_\varphi J^{\mathrm{mix}}$ 可按如下方式计算：

$$\begin{aligned} \nabla_\varphi J^{\mathrm{mix}} = \alpha \cdot \nabla_\varphi \log m_{\varphi,\eta}\left(a_t^2 \mid o_t^2, c_t^2\right) \Big(& G^{\mathrm{mix}}\left(\mathbf{o}_t, \mathbf{a}_t\right) \\ & - b_2\left(o_t^2, \mathbf{a_t}\right) - \alpha \log\left(m_{\varphi,\eta}\left(a_t^2 \mid o_t^2, c_t^2\right)\right)\Big) \end{aligned}, \tag{14}$$

其中，$G^{\mathrm{mix}}\left(\mathbf{o}_t, \mathbf{a}_t\right)$ 表示我们 MoP-SAN 的折扣混合回报。梯度 $\nabla_\eta J^{\mathrm{ex}}$ 可使用链式法则计算：

$$\nabla_\eta J^{\mathrm{ex}} = \nabla_{\varphi'} J^{\mathrm{ex}} \nabla_\eta \varphi', \tag{15}$$

with

$$\begin{aligned} \nabla_\eta \varphi' &= \nabla_\eta \alpha G^{\mathrm{mix}}\left(\mathbf{o}_t, \mathbf{a}_t\right) \nabla_\varphi \log m_{\varphi,\eta}\left(a_t^2 | o_t^2, c_t^2\right) \\ &= \alpha\beta \sum_{l=0}^{\infty} \gamma^l \nabla_\eta R_{\eta,t+l}^{\mathrm{d}} \nabla_\varphi \log m_{\varphi,\eta}\left(a_t^2 | o_t^2, c_t^2\right) \end{aligned}, \tag{16}$$

我们可以利用重要性采样来提高算法的样本效率：

$$\nabla_{\varphi'} J^{\mathrm{ex}} = \nabla_{\varphi'} \left(\frac{m_{\varphi',\eta}\left(a_t^2 | o_t^2, c_t^2\right)}{m_{\varphi,\eta}\left(a_t^2 | o_t^2, c_t^2\right)}\right) G^{\mathrm{mix}}\left(\mathbf{o}_t, \mathbf{a}_t\right), \tag{17}$$

$$\begin{aligned} \nabla_\eta J^{\mathrm{ex}} &= \nabla_{\varphi'} J^{\mathrm{ex}} \nabla_\eta \varphi' \\ &= \nabla_{\varphi'} \left(\frac{m_{\varphi',\eta}\left(a_t^2 | o_t^2, c_t^2\right)}{m_{\varphi,\eta}\left(a_t^2 | o_t^2, c_t^2\right)}\right) G^{\mathrm{mix}}\left(\mathbf{o}_t, \mathbf{a}_t\right) \alpha\beta \cdot \\ &\quad \sum_{l=0}^{\infty} \gamma^l \nabla_\eta R_{\eta,t+l}^{\mathrm{dpp}} \nabla_\varphi \log m_{\varphi,\eta}\left(a_t^2 | o_t^2, c_t^2\right) \end{aligned}. \tag{18}$$

因此，SAN和MoP模型中策略参数的迭代学习最终使整个系统收敛，从而支持后续的多智能体强化学习任务。

4 实验结果

4.1 环境设置

我们的实验环境是 Overcooked（Carroll 等人，2019年），这是一个主要的人工智能-人类零样本协作基准。与先前的工作（Carroll 等人，2019年；Shih 等人，2021、2022）类似，我们

我们基于 PantheonRL（Sarkar 等，2022）开展了“simple”地图实验。PantheonRL 是一个用于人机协作的 PyTorch 框架。在该环境中，两名玩家协作完成烹饪任务，即在有限时间内尽可能多地制作洋葱汤，以获得更高奖励。玩家可以从六种动作中选择一种并同时执行，包括上、下、左、右、空操作或交互。

制作洋葱汤时，必须遵循特定的顺序。玩家必须将三个洋葱放入锅中，烹煮20步。然后，玩家将锅中的洋葱汤倒入盘子，并将菜肴端到指定位置。完成这一过程后，玩家可以获得一定的奖励（20）。在具有挑战性的任务中，玩家无法独自完成这项任务。只有通过良好的合作，玩家才能取得高分；这要求玩家先推断合作伙伴的个性，并预测合作伙伴的动作。

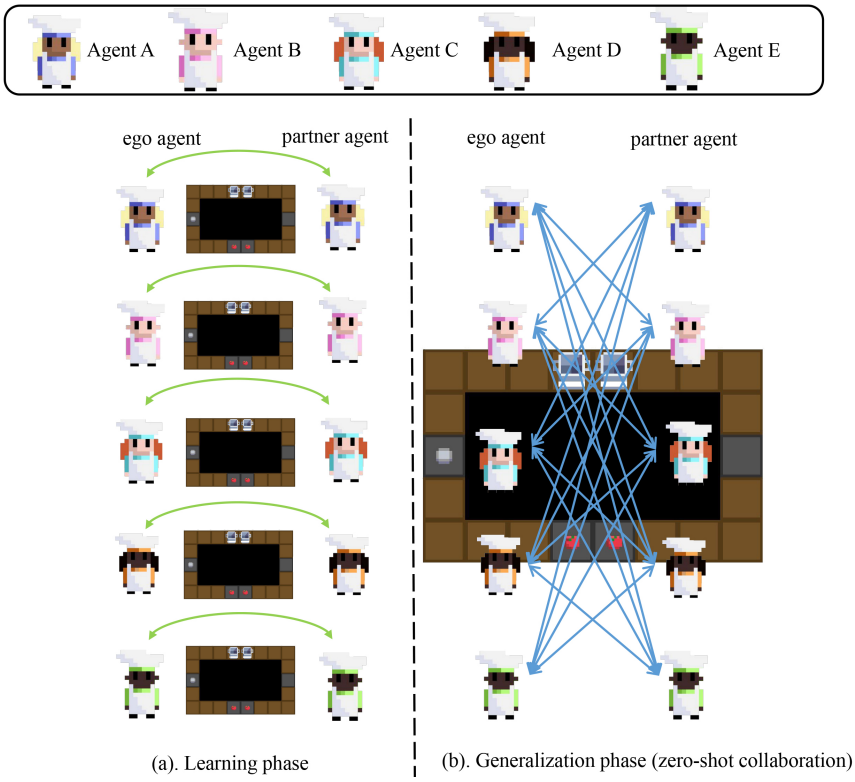


图3. 我们的 MoP-SAN 在学习阶段和泛化阶段（零样本协作）中的实验设置详情。智能体 A-E 分别对应使用不同随机种子的智能体，其名称为 A-E。(a) 在学习阶段，自我智能体与特定伙伴智能体组成一对，共同完成此任务，并通过迭代优化进行训练。每对中的自我智能体和伙伴智能体名称相同。学习阶段共有五对智能体：(A, A)、(B, B)、(C, C)、(D, D) 和 (E, E)。(b) 在泛化阶段，自我智能体需要以零样本的方式与所有未见过的伙伴智能体合作。例如，自我智能体 A 将与另一个名称不同的未见过的伙伴智能体（B、C、D 或 E）合作，进行零样本协作测试。

4.2 我们的基线方法和 MoP-SAN 的配置

有几种基线方法。其中一种是标准的深度神经网络近端策略优化基线（Schulman 等人，2017），这是一种重要的多智能体强化学习方法，在许多场景中表现出色。在这种方法中，两者都

自我智能体和伙伴智能体都是同质的近端策略优化智能体，这种方式在强化学习中也称为自博弈（Silver 等人，2018）。

另一个重要的基线是 SAN PPO基线。由于我们方法中的自我智能体也是 SAN PPO，因此在下文的实验描述中，我们将 SAN PPO基线称为 SAN 基线。值得一提的是，我们首次将 SAN 版本的 PPO 引入多智能体协作任务 Overcooked。对于 SAN 基线，在我们的合作环境中，自我智能体是 SAN PPO，合作伙伴则是标准PPO。

我们的实验设置详情如图3所示。如图1和图3所示，在我们的 MoP-SAN 中，同一对中的 SAN 智能体和 MoP 具有相同的名称，并在学习阶段通过迭代优化进行联合训练。例如，我们的 SAN A 作为自我智能体、 MoP A 作为合作伙伴，将在学习阶段协作以取得好成绩。在泛化阶段，名称相同的 SAN 和 MoP 将组合成 MoP-SAN，作为自我智能体。我们将通过与不同的未见合作伙伴协作，评估所提出的 MoP-SAN 模型的泛化能力，这意味着同一对中的自我智能体和伙伴智能体名称不同。

我们的训练实验运行五十万步，并在多个游戏中开展泛化实验（零样本协作），以计算所有实验在泛化阶段的平均得分。个性数量为 12，上下文大小为 5。对于 MoP-SAN 中的上下文编码器，如果合作伙伴历史轨迹的长度小于上下文大小，我们会用 0 填充。我们使用一个具有两个注意力头的单层 Transformer 作为上下文编码器，其内部维度为 256，q、k、v 的维度为 64。对于填充的 0，我们在 Transformer 中将其屏蔽。我们的 MoP-SAN 模型采用演员-评论家框架，演员基于 SAN，与先前工作类似（Zhang 等人，2022）。演员网络为（64，tanh，64，tanh，6）；评论家网络为（64，tanh，64，tanh，1）。所有方法均从分类分布中采样动作。在这些方法中，我们使用 Adam优化器，学习率为 0.0003。奖励折扣因子为 $\gamma = 0.99$，批量大小为64。内在奖励 β 的权重系数为 0.5，经验回放缓冲区的最大长度为 2048。我们使用梯度裁剪来防止梯度爆炸和梯度消失。

4.3 MoP-SAN 更强的泛化能力

图4是表示三种方法在Overcooked任务中获得的泛化得分和学习得分的直方图。直方图中的折线图显示了不同方法平均得分的趋势。红点表示所有相应智能体的平均得分，阴影区域表示三种方法相应结果的标准差。

左侧图表中该方法的平均得分，是所有与未见过的合作伙伴进行的泛化测试的平均得分。如图3所示，我们的 MoP-SAN 方法在 A 中的平均得分为 142，这意味着四项未见过的测试（A-B、A-C、A-D、A-E）的平均得分为 142。我们的方法的平均得分为 142.25，这意味着二十项未见过的测试（A-B、A-C、A-D、A-E、B-A、B-C、B-D、B-E、C-A. . .）的平均得分为 142.25。图5展示了所有与未见过的合作伙伴进行的泛化测试的详细得分。每对组合在学习阶段和泛化阶段的详细得分可见补充材料。

图4表明，在零样本协作过程中，针对未见过的合作伙伴，我们提出的 MoP-SAN 模型优于所有基线方法，展现出更强的合作稳健性和稳定性。需要进一步强调的是，在泛化测试中，我们的 MoP-SAN 方法不仅显著优于 SAN 基线，也优于 DNN 基线，这有力地证明了我们的方法在零样本协作中对合作伙伴异质性具有强大的泛化能力。

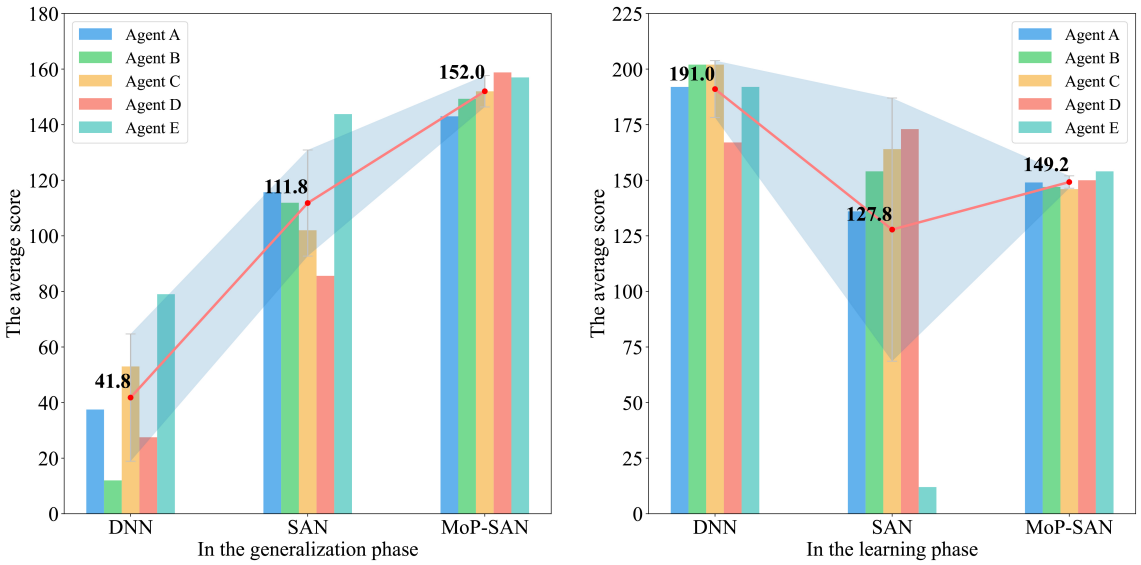


图4. 基线模型与 MoP-SAN 模型在泛化和学习阶段的得分比较。左图显示，我们的 MoP-SAN 在泛化表现方面优于其他基线方法，具有较强的泛化能力，能够与未见过的合作伙伴完成合作任务。右图显示，我们的 MoP-SAN 改善了 SAN 在学习阶段表现不佳且方差较大的问题。智能体A-E表示使用不同随机种子的不同智能体。

学习阶段的平均得分见图4右侧的图。尽管我们的 MoP-SAN 方法主要关注在没有任何合作伙伴先验知识的情况下进行零样本泛化测试，但学习阶段的得分仍能反映与特定合作伙伴的协作表现。与学习阶段的 SAN 基线相比，我们的 MoP-SAN 得分更高，且方差更小。

4.4 MoP-SAN 的零样本协作表现显著更优

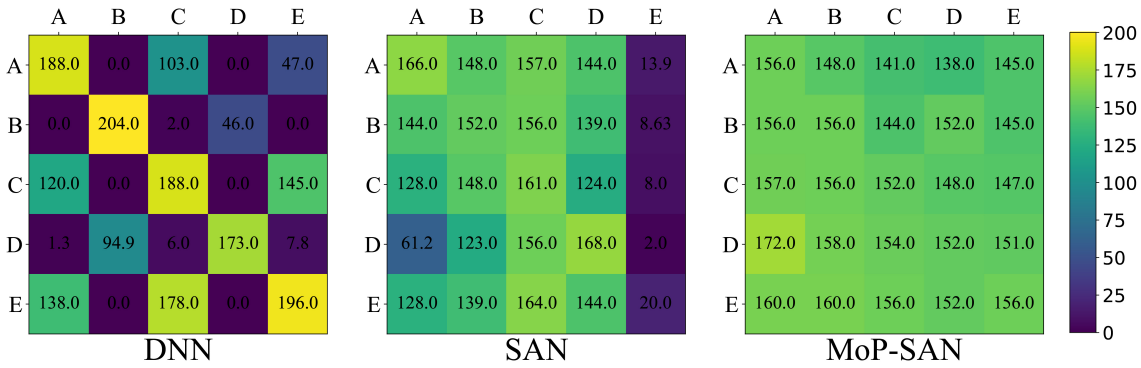


图5. 色温图展示了基线方法和我们的 MoP-SAN 的详细泛化得分。颜色差异体现了得分差异。与深度神经网络和 SAN 基线相比，我们提出的 MoP-SAN 得分更高、方差更小，效果更令人满意。

我们在零样本协作测试中的实验结果反映了不同方法对合作伙伴异质性的泛化能力。图5为色温图，展示了具体的实验。

三种方法在泛化测试中的数据。图5中的色温图分别对应 DNN 基线、SAN 基线和我们的 MoP-SAN 模型。行代表自我智能体，列代表合作伙伴。例如，我们的 MoP-SAN 在（第一行、第三列）中的得分，表示 MoP-SAN A 与未见过的合作伙伴 C 之间的零样本协作得分。对角线上的得分表示相应配对在学习阶段取得的得分，这些得分不计入泛化阶段的零样本协作得分数据。可以看出，颜色差异越明显，该方法的方差越大。

如图5所示，在能耗较低的零样本泛化测试中，对于大多数合作伙伴组合，我们的多尺度生物合理性 MoP-SAN 得分显著高于其他基线方法，方差也更小；面对风格各异、未见过的合作伙伴时，它也取得了良好的泛化结果。尽管深度神经网络在部分泛化测试实验中得分较高，但其方差较大，平均得分较低。此外，SAN 基线的平均得分高于 DNN 基线，方差也更小。这些结果表明，我们的 MoP模型能够完成伙伴建模，并帮助 SAN智能体取得更高的协作得分和更强的泛化能力。

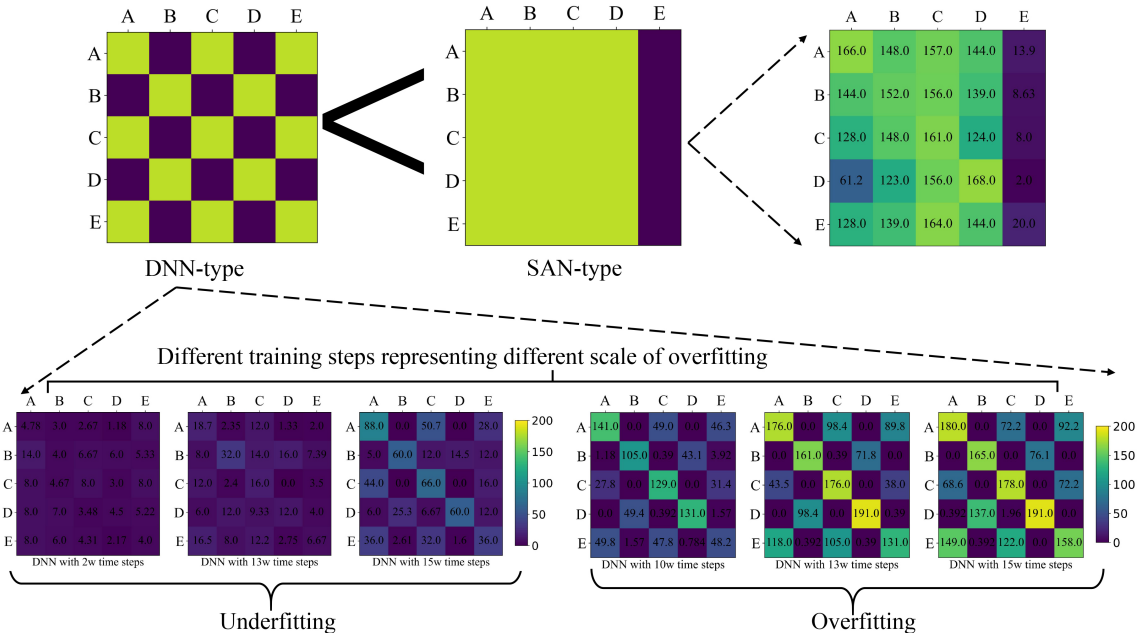


图6。该图详细展示了深度神经网络和 SAN 的泛化分析实验，并给出了深度神经网络在不同训练步数下的泛化测试结果，这些步数代表了不同程度的过拟合。随着训练步数增加，深度神经网络的泛化表现逐步提升。深度神经网络的泛化测试结果呈现出与深度神经网络类型相似的模式，而 SAN 的结果也呈现出与 SAN 类型相似的模式。通过比较这两种模式，我们可以看出，SAN 具有更强的泛化能力和鲁棒性。

SAN 为何能够取得比深度神经网络更好的泛化结果，这一问题引起了我们的关注。为了进一步验证深度神经网络在泛化测试中表现不佳是否是由过拟合导致的，我们对深度神经网络进行了一系列分析实验。我们保存了深度神经网络从欠拟合到“过拟合”这一学习过程中的检查点，并进行了未见过的合作伙伴泛化测试。结果表明，随着训练步数的增加，深度神经网络的泛化性能逐步提升。我们在这些测试结果中发现了类似的模式，并将其命名为深度神经网络型。

同样，在 SAN 的泛化测试结果中，我们也发现了类似的模式，并将其命名为 SAN 型。与深度神经网络型相比，SAN 型在面向未见过的合作伙伴的泛化场景中表现出更强的泛化能力和合作能力。这些结果表明，“过拟合”并非深度神经网络在泛化测试中表现不佳的主要原因。我们认为，深度神经网络在与未见过的合作伙伴进行泛化测试时表现不如 SAN，是因为 SAN 具有更强的抗噪能力和鲁棒性。在合作式强化学习中，与未见过的合作伙伴进行泛化测试可视为噪声扰动测试，因此，在我们的泛化实验中，SAN 的表现优于深度神经网络。

4.5 更大的人格规模有助于提升协作性能

此外，我们进行了一些消融实验，以确认 MoP-SAN 中不同模块和参数的有效性。表 1 的实验结果表明，随着人格数量的增加

表 1. 我们的方法中不同人格数量对应的平均分数。

智能体	A	B	C	D	E	avg
我们的 w/个性 6	0.2	0.4	0.4	0	1.6	0.52 (±0.63)
我们的个性 w/个性 8	1.8	123	0	0.4	0	25.04 (±54.77)
我们的 w/人格 10	7.6	151	1.6	**157**	114	86.24 (±76.35)
我们的 w/ 个性 12	**149**	**154**	**150**	146	**146**	**149 (±3.32)**

随着学习能力的提升，我们的 MoP-SAN 模型的学习能力逐渐增强，方差也逐渐减小。这些结果还表明，多样化的人格在多智能体协作任务中发挥着至关重要的作用。

从表 1 可以看出，当基础人格数量较少时，一些组合的合作得分非常低。这可能是因为这些基础人格无法组合起来表达合作伙伴个性的所有维度。随着基础人格数量增加，现有基础人格对当前合作伙伴个性的表达能力也随之增强，从而带来更好的表现。

认知心理学中的人格理论认为，将个性分解为更细粒度的特质，是提高对人类行为进行预测和解释能力的有效方法。我们的实验结果进一步验证了这一观点。通过采用更多的个性类别，我们能够实现更精确的人格刻画，从而更准确地预测合作伙伴的个性，并与合作伙伴开展更高效的协作，最终取得更高的分数。

4.6 更丰富的上下文信息有助于更好地预测个性

表2显示，随着合作伙伴上下文信息的增加，我们的 MoP-SAN 在学习阶段的得分越来越高，这表明合作伙伴信息对于我们的 MoP-SAN 模型完成合作任务至关重要。完全没有合作伙伴信息时，结果最差。这是因为合作伙伴信息是 PE 模块预测合作伙伴个性的输入。缺少此类信息时，个性预测就会变得随机，导致自我智能体与伙伴智能体在完成制作洋葱汤等任务时无法高效协作。合作伙伴信息有限可能导致个性预测不准确，从而不利于提升合作得分。

表2. 我们的方法在不同上下文大小下的平均分数。

智能体	A	B	C	D	E	avg
我们的方法，上下文 0	4.4	2.8	2.6	1	1.2	2.4 (±1.38)
我们的方法，上下文 1	11.6	0.2	1	17.2	85.2	23.04 (±35.48)
我们的模型（上下文 / 3）	136	148	143	140	140	141.4 (±4.45)
我们的模型（上下文 / 5）	**149**	**154**	**150**	**146**	**146**	**149 (±3.32)**

表2中的这些结果还表明，合作伙伴上下文信息的存在是我们解决此任务的关键。我们发现，合作伙伴信息的存在能让模型在学习阶段取得更好的结果，并在零样本协作泛化实验中获得更好的泛化结果。

4.7 由 DPP 控制的个性多样性

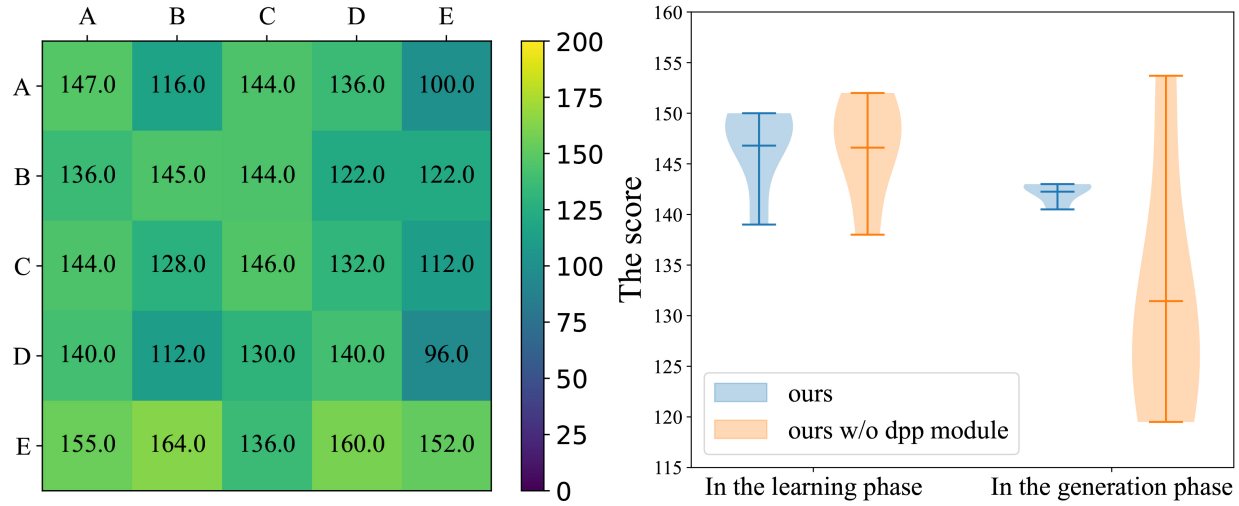


图7。左侧色温图展示了我们未使用DPP模块的方法的详细泛化得分。右侧小提琴图直观地比较了DPP模块消融实验中的得分，其中分别展示了我们未使用DPP模块的方法和我们的方法在学习阶段与泛化阶段的得分。

DPP消融实验的结果证明了DPP模块的有效性，该模块能够在泛化实验中取得更好的结果。我们进一步通过图7中的色温图和小提琴图分析消融实验结果。小提琴图中展示了最大值、最小值和平均值曲线，阴影表示数据分布，其大小代表相应方法的方差。如图7右侧的小提琴图所示，在泛化测试中，我们的方法明显优于未使用DPP的方法；此外，我们的MoP-SAN的方差小于未使用DPP的MoP-SAN。图5中的第三幅图c展示了我们的MoP-SAN的色温图。对比图7左侧的色温图与图5中的图c可知，得益于DPP模块，我们的MoP-SAN模型具有更好的泛化性能和更小的方差。

这一结果表明，在个性数量相同的情况下，加入 DPP 能够约束 MoP 中的基础人格，从而使这些基础人格尽可能多地覆盖人格空间。

这是可能的。这种全面覆盖使 PE模块 更加稳健，能够更准确地预测未见过的合作伙伴的个性，从而取得更好的分数。

5 结论

本文着重通过融入多尺度生物学启发来增强传统演员网络，包括采用脉冲编码的局部尺度神经元动力学，以及蕴含心智理论理念的全局尺度人格理论。我们提出的人格混合增强脉冲演员网络（MoP-SAN）算法，能够在能耗极低的情况下，显著提升多智能体强化学习协作场景中的泛化能力和适应性。

随后，我们通过实验验证了 MoP-SAN。实验结果表明，人格理论中的两步过程对于预测未见过的合作伙伴动作至关重要。与 SAN 和 DNN 基线相比，MoP 改进的 SAN 展现出更令人满意的学习能力和泛化性能。据我们所知，我们是首个将 SAN 和 MoP 应用于多智能体强化学习协作任务的团队。这一整合成果使我们更有信心在未来借鉴更多神经科学和认知心理学方面的启发，以设计新一代多智能体强化学习算法。

致谢

作者谨向 Yali Du、Dengpeng Xing、Zheng Tian 和 Duzhen Zhang 致谢，感谢他们此前通过宝贵的讨论给予的帮助。

作者贡献

B.X、J.S、T.Z 和 X.L. 提出了构想；X.L.、Z.N. 开展了实验并分析结果；X.L.、J.R、L.M 参与了问题定义。他们共同撰写了论文，并批准了提交版本。

利益冲突声明

作者声明，本研究是在不存在任何可能被视为潜在利益冲突的商业或财务关系的情况下开展的。

资金支持

本研究由中国科学院战略性先导科技专项（XDA27010404、XDB32070100）、上海市科技重大项目（2021SHZDZX）和中国科学院青年创新促进会资助。

参考文献

Anglim, J. 和 Horwood, S.（2021）。新冠肺炎疫情和大五人格对主观幸福感与心理幸福感的影响。社会心理学与人格科学12，1527–1537 Aru, J.、Labash, A.、Corcoll, O. 和 Vicente, R.（2023）。注意差距：深度学习方法在心智理论方面的挑战。人工智能评论，1–16 Bellec, G.、Scherr, F.、Subramoney, A.、Hajek, E.、Salaj, D.、Legenstein, R. 等（2020）。脉冲神经元循环网络学习困境的解决方案。《自然·通讯》11，1–15

Boyd, R. 和 Richerson, P. J.（2009）。文化与人类合作的演化。英国皇家学会哲学汇刊*B*辑：生物科学364，3281–3288 Carroll, M.、Shah, R.、Ho, M. K.、Griffiths, T.、Seshia, S.、Abbeel, P. 等（2019）。学习了解人类对人类-人工智能协作的效用。神经信息处理系统进展32 Cattell, H. E. 和 Mead, A. D.（2008）。十六种人格因素问卷（16PF）。*SAGE*人格理论与评估手册 2，135–159 Chen, L.、Lu, K.、Rajeswaran, A.、Lee, K.、Grover, A.、Laskin, M. 等（2021）。决策Transformer：通过序列建模实现强化学习。神经信息处理系统进展 34，15084–15097 Chen, L.、Zhang, G. 和 Zhou, E.（2018）。用于行列式点过程的快速贪心映射推断，以提升推荐多样性。神经信息处理系统进展31 Dai, T.、Du, Y.、Fang, M. 和 Bharath, A. A.（2022）。用于深度强化学习的多样性增强的内在动机。神经计算468，396–406 De Raad, B.（2000）。大五人格因素：人格的心理词汇学方法。（Hogrefe与Huber出版社）Florian, R. V.（2007）。通过调节脉冲时序依赖的突触可塑性进行强化学习。神经计算19，1468–1502 Fr´emaux, N.、Sprekeler, H. 和 Gerstner, W.（2013）。使用基于脉冲神经元的连续时间演员-评论家框架进行强化学习。*PLoS*计算生物学9，e1003024 Frith, C. 和 Frith, U.（2005）。心智理论。当代生物学15，R644–R645 Gallagher, H. L. 和 Frith, C. D.（2003）。“心智理论”的功能成像。认知科学趋势7，77–83 Gong, B.、Chao, W.-L.、Grauman, K. 和 Sha, F.（2014）。用于监督式视频摘要的多样化序列子集选择。神经信息处理系统进展27 Kulesza, A.、Taskar, B. 等（2012）。用于机器学习的行列式点过程。机器学习基础与趋势®5，123–286 Lou, X.、Guo, J.、Zhang, J.、Wang, J.、Huang, K. 和 Du, Y.（2023）。Pecan：利用策略集成实现上下文感知的零样本人类-人工智能协作。*arXiv*预印本 *arXiv:2301.06387*McCrae, R. R. 和 Costa Jr, P. T.（2008）。人格五因素理论。Meng, L.、Wen, M.、Le, C.、Li, X.、Xing, D.、Zhang, W. 等（2023）。离线预训练多智能体决策Transformer。机器智能研究20，233–248 Mnih, V.、Kavukcuoglu, K.、Silver, D.、Graves, A.、Antonoglou, I.、Wierstra, D. 等（2013）。使用深度强化学习玩Atari游戏。*arXiv*预印本 *arXiv:1312.5602*O’Connor, M. C. 和 Paunonen, S. V.（2007）。大五人格对高等教育阶段学业表现的预测作用。人格与个体差异43，971–990 Parker-Holder, J.、Pacchiano, A.、Choromanski, K. M. 和 Roberts, S. J.（2020）。基于群体的强化学习中的有效多样性。神经信息处理系统进展33，18050–18062 Patel, D.、Hazan, H.、Saunders, D. J.、Siegelmann, H. T. 和 Kozma, R.（2019）。将强化学习策略转换为脉冲神经网络平台后，其鲁棒性得到提升，并应用于Atari Breakout游戏。神经网络120，108–115 Perez-Nieves, N.、Yang, Y.、Slumbers, O.、Mguni, D. H.、Wen, Y. 和 Wang, J.（2021）。对开放式游戏中学习所需的行为多样性进行建模。载于国际机器学习会议（PMLR），8514–8524

Rabinowitz, N., Perbet, F., Song, F., Zhang, C., Eslami, S. A., and Botvinick, M. (2018). 机器心智理论. 收录于国际机器学习会议(PMLR), 4218–4227 Rand, D. G. and Nowak, M. A. (2013). 人类合作. 认知科学趋势17, 413–425 Rashid, T., Samvelyan, M., De Witt, C. S., Farquhar, G., Foerster, J., and Whiteson, S. (2020). 深度多智能体强化学习的单调价值函数分解. 机器学习研究期刊21, 7234–7284 Roth, M., Marsella, S., and Barsalou, L. (2022). 心智理论中的偷工减料 Ruan, J., Du, Y., Xiong, X., Xing, D., Li, X., Meng, L., et al. (2022). GCS：基于图的多智能体强化学习协调策略. *arXiv*预印本 *arXiv:2201.06257*Ryckman, R. M. (2012). 人格理论(圣智学习) Sarkar, B., Talati, A., Shih, A., and Sadigh, D. (2022). PantheonRL：用于动态训练交互的多智能体强化学习库. 收录于*AAAI*人工智能会议论文集. vol. 36, 13221–13223 Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. (2015). 信赖域策略优化. 收录于国际机器学习会议(PMLR), 1889–1897 Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). 近端策略优化算法. *arXiv*预印本 *arXiv:1707.06347*Schultz, D. P. and Schultz, S. E. (2016). 人格理论(圣智学习) Shih, A., Ermon, S., and Sadigh, D. (2022). 多智能体博弈的条件模仿学习. 收录于*2022*年第*17*届*ACM/IEEE*人机交互国际会议（*HRI*）(I EEE), 166–175 Shih, A., Sawhney, A., Kondic, J., Ermon, S., and Sadigh, D. (2021). 惯例在人机自适应协作中的关键作用. *arXiv*预印本 *arXiv:2104.02871*Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., et al. (2018). 一种通过自博弈掌握国际象棋、日本将棋和围棋的通用强化学习算法. 《科学》 362, 1140–1144 Silver, D., Sc hrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., et al. (2017). 无需人类知识掌握围棋. 《自然》550, 354–359 Strouse, D., McKee, K., Botvinick, M., Hughes, E., and Everett, R. (2021). 在没有人类数据的情况下与人类协作. 神经信息处理系统进展34, 14502–14515 Sunehag, P., Lever, G., Gruslys, A., Czarnecki, W. M., Zambaldi, V., Jaderberg, M., et al. (2017). 协作式多智能体学习的价值分解网络. *arXiv*预印本 *arXiv:1706.05296*Tabrez, A., Luebbers, M. B., and Hayes, B. (2020). 人机团队协作中的心智建模技术综述. 《当代机器人报告》1, 259–267 Tang, G., Kumar, N., and Michmizos, K. P. (2020). 在神经形态硬件上对深度神经网络和脉冲神经网络进行强化协同学习，以实现节能无地图导航. 收录于*2020*年*IEEE/RSJ*智能机器人与系统国际会议（*IROS*）(IEEE), 6090–6097 Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). 注意力机制即你所需. 神经信息处理系统进展30 Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., et al. (2019). 使用多智能体强化学习达到《星际争霸II》宗师级水平. 《自然》575, 350–354 Wang, Y., Zhong, F., Xu, J., and Wang, Y. (2021). TOM2C：基于心智理论的目标导向多智能体通信与协作. *arXiv*预印本 *arXiv:2111.09189*Yang, R., Xu, H., Wu, Y., and Wang, X. (2020). 通过软模块化实现多任务强化学习.神经信息处理系统进展33, 4767–4777 Yu, C., Velu, A., Vinitsky, E., Wang, Y., Bayen, A., and Wu, Y. (2021). PPO在协作式多智能体博弈中出人意料的有效性. *arXiv*预印本 *arXiv:2103.01955*

Yuan, L., Gao, X., Zheng, Z., Edmonds, M., Wu, Y. N., Rossano, F. 等（2022）。原位双向人机价值对齐。*Science Robotics*7, eabm4183 Zhang, D., Zhang, T., Jia, S., 和 Xu, B.（2022）。用于强化学习的多尺度动态编码改进脉冲演员网络。载于 *AAAI*人工智能会议论文集。第 36 卷, 59–67 Zhao, R., Song, J., Haifeng, H., Gao, Y., Wu, Y., Sun, Z. 等（2021）。用于零样本人机协作的最大熵群体训练。*arXiv*预印本 *arXiv:2112.11701*

arXiv:2305.05898v1 [cs.AI] 10 May 2023

补充材料

1 补充表格和图

1.1 图

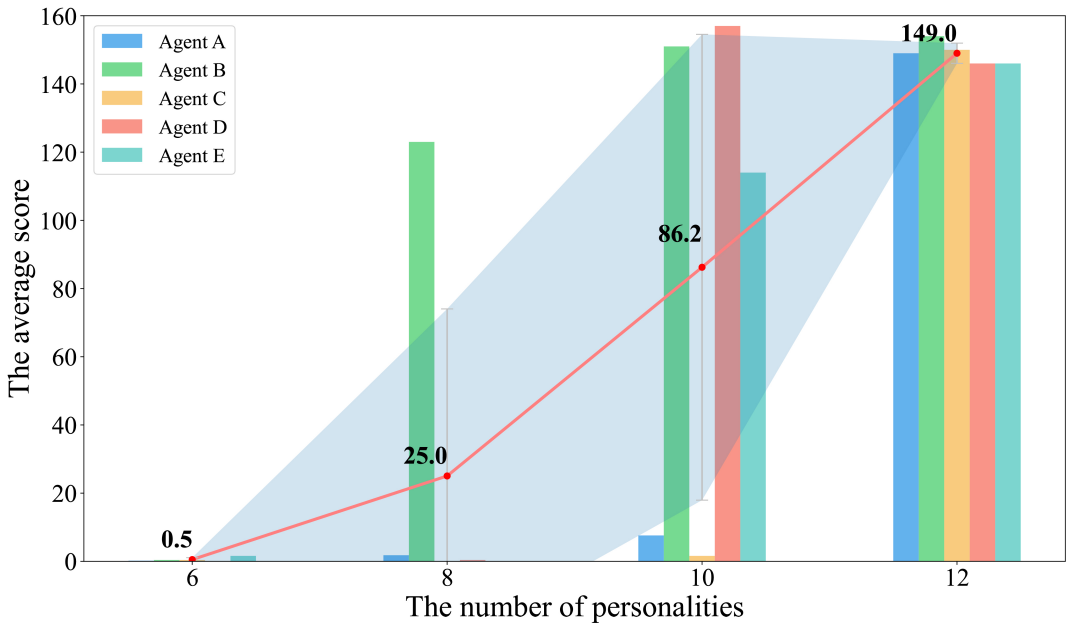


图 S1. 该图展示了不同人格数量（6、8、10、12）对应的详细得分。A-E 表示使用不同随机种子的不同智能体。在我们的方法中，随着人格数量的增加，得分显著提高。

1.2 表格

表 S1. 所有方法的平均奖励。

方法	A	B	C	D	E	avg
DNN 基线	**192**	**202**	**202**	167	**192**	**191.0 (±14.32)**
SAN 基线	136	154	164	**173**	12	127.8（±66.18）
MoP-SAN（我们的方法）	149	147	146	150	154	149.2（±3.32）

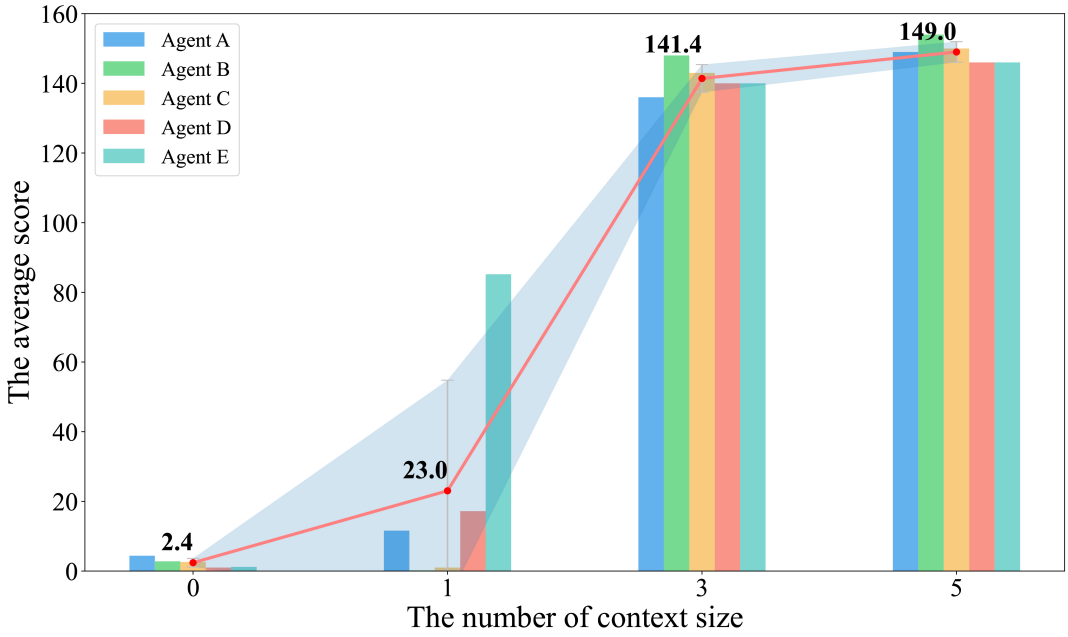


图 S2. 该图展示了不同上下文规模（0、1、3、5）下的详细得分。A–E 表示使用不同随机种子的不同智能体。在我们的方法中，得分会随着上下文数量的增加而提高。

表 S2. 泛化阶段的整体泛化奖励均值。

方法	A	B	C	D	E	avg
DNN 基线	37.5	12	53	27.5	79	41.8 (±25.59)
SAN 基线	115.7	111.9	102	85.6	143.8	112.2 (±19.11)
MoP-SAN（我们的方法）	**143**	**149.3**	**152**	**158.8**	**157**	**152.0 (±6.32)**

表 S3. 我们不使用 Transformer 上下文编码器的方法与我们的方法的平均分数。

方法	A	B	C	D	E	avg
我们的方法（不含/上下文编码器）	4.4	2.8	2.6	1	1.2	2.4（±1.38）
ours	**149**	**154**	**150**	**146**	**146**	**149**（±3.32）

表 S4。 我们的方法（不含 DPP模块）和我们的方法的平均泛化得分。

方法	A	B	C	D	E	avg
我们的方法（不含/ DPP模块）	124	131	129	119.5	153.7	131.4 (±13.22)
ours	**143**	**149.3**	**152**	**158.8**	**157**	**152.0 (±6.32)**