arXiv:2601.05509v1 [cs.MA] 9 Jan 2026

探索如何破坏共享策略多智能体强化学习中的合作

Yi-Ning Weng[a]、Hsuan-Wei Lee[b,*]

[a]台湾大学会计学系，台湾[b]美国利哈伊大学健康学院

摘要

动态博弈中的多智能体强化学习（MARL）面临着根本性挑战，这是由非平稳交互和探索噪声所致，而参数共享等常见设计选择往往会加剧这些挑战。尽管参数共享通常用于提高样本效率和可扩展性，但其在探索过程中对合作行为的影响仍鲜为人知。本研究在动态囚徒困境环境中，对采用参数共享的深度Q网络（DQN）进行了系统研究，重点考察探索与收益严苛度如何共同影响学习稳定性和合作结果。

我们发现，随着探索增加，共享DQN中的合作会稳健地崩溃。当收益强烈偏向背叛时，这种崩溃最为严重。这种崩溃并非偶发，而是在参数空间中形成了结构化的失效区域，其失效边界会随博弈严重程度系统性地移动。仅凭收益激励或学习不稳定性无法解释这一现象。合作崩溃伴随着学习到的表征退化：合作行为与背叛行为在特征空间中变得纠缠，动作价值差距也随之系统性缩小。系统最终收敛到稳定但合作程度较低的解。

我们发现，部分可观测性和表征耦合是关键驱动因素。分组策略学习将参数共享限制在智能体子集内，能够持续缓解合作崩溃。加入学习进度和探索信号的状态增强显著延缓了崩溃，并在某些状态下使其部分逆转。网络拓扑结构也很重要：空间结

*通讯作者。电子邮箱： hsl324@lehigh.edu (Hsuan-Wei Lee)

构通过局部协调支持合作，而随机网络则会消除合作状态。这些干预证实，失败源于表征，而非算法。

我们的研究揭示了合作情境中参数共享深度Q网络的一种根本性失效模式：共享表征在探索过程中无法维持稳定的合作。然而，采用表征感知的状态设计可以显著提升鲁棒性。这些结果对于设计合作至关重要的可扩展多智能体强化学习系统具有重要启示。

关键词： 多智能体强化学习、共享策略学习、深度Q网络、动态博弈、表征学习

---

1. 引言

多智能体强化学习（MARL）为多个人工智能体通过反复交互进行学习的系统建模提供了一个框架。其应用涵盖分布式控制、网络化系统以及大规模协调问题；在这些场景中，个体决策者必须适应环境反馈和其他参与者不断演变的行为。探索是强化学习面临的一项根本挑战。在多智能体情境中，这一挑战更加严峻，因为并发学习的智能体会导致每个智能体所经历的有效环境呈现非平稳性 [1, 2, 3]。尽管已有研究提出结构化探索机制，以提高单智能体DQN中的采样效率 [4], ，但这些机制与多智能体情境下参数共享之间的相互作用仍缺乏充分认识。因此，理解学习架构如何与非平稳性和探索相互作用，对于构建可扩展且稳健的多智能体强化学习系统至关重要 [5, 6, 7, 8, 9, 10]。

参数共享已成为应对多智能体强化学习可扩展性问题的标准方法 [11, 12, 13]。通过使用共享策略或价值函数训练多个智能体，这种设计降低了样本复杂度，并使大规模智能体种群中的学习成为可能。许多基于价值的协作式多智能体强化学习方法采用集中式训练与分散式执行，并在训练期间纳入有关队友行为的信息 [14, 15]。尽管具有这些优势，参数共享仍施加了一项根本性约束：所有智能体都通过同一条表征与优化路径相互耦合 [16, 17]。无论局部交互情境有何差异，每个智能体都会向同一个网络贡献梯度，并依据该网络做出决策。作为

因此，共享策略会受到来自众多交互智能体的聚合探索噪声和非平稳反馈的影响 [18]。参数共享通常被视为一种无害的工程选择。然而，人们对其在探索过程中的学习稳定性和合作所产生的影响仍缺乏深入理解，尤其是在智能体面临异质且不断变化的交互情境时。

这一理解上的缺口在合作情境中尤为关键。持续协调需要个体动作与集体结果之间保持稳定的反馈。探索或协调失误造成的微小扰动，可能沿着交互过程传播并破坏合作结构[19, 20, 21]。在探索过程中，共享策略的多智能体强化学习中经常会观察到合作崩溃。然而，人们很少将其作为系统性现象加以研究。对此有三种可能的解释：不可避免的激励冲突、训练不稳定性，或共享表征处理非平稳多智能体交互时存在的根本局限。本研究旨在区分这些可能性。网络结构也会影响合作学习。先前研究表明，网络拓扑结构和自适应重连会影响学习系统中合作的出现与稳定性 [22, 23, 24, 25, 26, 27, 28, 29]。我们考察交互结构如何调节所发现的合作崩溃。

我们通过系统研究动态囚徒困境中的参数共享深度Q网络来探讨这些问题。该环境可作为一种诊断工具，用于对激励结构、探索强度和网络拓扑结构进行受控调整 [30, 31, 32,33, 34]。我们将合作率用作学习稳定性和协调质量的指标，而非优化目标。我们的主要发现是，随着探索增加，参数共享深度Q网络中会出现稳健的合作崩溃。这种崩溃发生在参数空间的广大区域内，并形成结构化的失效边界；这些边界会随博弈严重程度系统性地变化。

为理解这种失败，我们考察了共享网络的内部动态。我们发现，合作崩溃与学习到的表征退化同时发生。合作行为与背叛行为在潜在空间中日益纠缠。在价值层面，动作价值差距呈系统性收缩。这表明系统收敛到了稳定但低合作的吸引子，而非训练不稳定性。这些发现排除了两种简单的解释：激励失配和过度的探索噪声。相反，失败根源在于共享表征如何处理

非平稳多智能体学习。

如果表征耦合导致了失败，那么降低耦合程度应该会有所帮助。我们使用分组策略学习来检验这一预测。分组策略学习将参数共享限制在智能体子集内。这种方法始终能提高稳定性并延缓崩溃，证实全局表征共享会导致失败。接下来，我们检验失败是否源于学习动态的部分可观测性。我们在状态中加入用于编码学习进度和探索强度的信号。这种扩展会延缓所有探索水平下的合作崩溃。在某些情形下，它还能部分扭转崩溃，使更广泛的参数范围内恢复稳定合作。这些干预表明，状态和表征设计对共享策略多智能体强化学习至关重要。仅靠调节探索无法实现鲁棒性。我们还考察网络拓扑结构如何影响合作崩溃。空间结构化网格通过提供局部协调支架来支持合作。随机拓扑结构则会消除这一益处，使合作崩溃更加严重。这揭示了一个边界条件：参数共享的 DQN 需要环境结构，才能在探索过程中维持合作。仅靠学习动态并不足够。

本研究识别并刻画了合作情境下参数共享深度Q网络的一种基本失效模式。我们表明，表征感知状态设计能够缓解这一失效。研究结果将宏观层面的合作失败与内部表征动态联系起来，为设计更稳健的共享策略多智能体强化学习系统提供了指导。

2. 方法

我们研究一个多智能体强化学习系统，其中智能体使用深度Q网络（DQN）学习在动态囚徒困境博弈（PDG）中采取动作。该模型基于将强化学习与演化博弈动态相结合的成熟框架构建，旨在考察由同步适应和探索所引发的非平稳交互下的学习行为[35]。

一群智能体嵌入固定的交互网络中，并反复与相邻智能体进行局部成对博弈。在每个离散时间步，每个智能体从两个可选动作中选择一个：合作（C）或背叛（D）。交互是对称的，并在整个网络中并发进行。每轮交互结束后，智能体根据局部博弈结果获得收益，并更新其决策

策略，使其基于累积经验进行调整。这种设置会内生地产生非平稳学习环境，因为每个智能体所面临的有效动态会随着时间推移而变化，以响应其他智能体的并发学习与探索。以下各小节将正式说明本文采用的动态博弈环境、共享策略学习架构、智能体的状态表示及其部分可观测性、探索机制和学习过程。在全文中，合作（C）在二进制状态表示中编码为 0，背叛（D）编码为 1。

2.1. 囚徒困境的收益结构与交互网络

我们采用一种困境强度参数化方法，可独立控制受剥削风险和背叛诱惑 [36, 24, 37, 38]。收益矩阵定义如下：

$$\begin{pmatrix} R & S \\ T & P \end{pmatrix} = \begin{pmatrix} 1 & -D_r \\ 1 + D_g & 0 \end{pmatrix}, \tag{1}$$

其中，$R = 1$ 是相互合作的奖励，$P = 0$ 是相互背叛的惩罚，$T = 1 + D_g$ 是背叛诱惑，$S = -D_r$是因遭受剥削而获得的受骗者收益。

参数 $D_g = T - R$ 和 $D_r = P - S$ 控制困境强度的两个方面，二者的取值范围均为 0 到 1。具体而言，$D_g \in [0, 1]$ 体现了单方面背叛所带来的诱惑优势，对应胆小鬼博弈型困境的成分；而 $D_r \in [0, 1]$ 衡量遭受剥削时所受惩罚的严重程度，反映了猎鹿博弈型困境的要素。当 $(D_r,\ D_g)$ 在 $[0, 1]^2$范围内变化时，这种参数化方法涵盖了双人社会困境的全部空间。

在这一框架中，D_r 主要控制被剥削的风险，而 D_g 调节背叛的奖励激励。(D_r, D_g) 平面中的不同区域对应不同类型的社会困境：当$D_r > D_g$ 时，系统呈现出强调协调收益的猎鹿博弈式特征；而 $D_g > D_r$ 则产生由背叛诱惑主导的胆小鬼博弈式动态。与先前关于合作动态对 D_r 的变化比对 D_g [39, 40]的变化更为敏感的研究结果一致，我们将分析重点放在对角线约束 $D_g = D_r$上，并在受限范围内系统地改变 D_r 。

智能体被排列在大小为 $L \times L$的二维方格晶格上，共有 $N = L^2$ 个智能体。我们采用周期性边界条件。每个智能体

与其四个最近邻（冯·诺依曼邻域）进行交互。每次运行期间，网络结构保持固定。这种空间结构会引发重复的局部交互，同时限制远距离信息传播，从而为研究局部交互约束下的合作学习提供了一个可控环境。研究表明，网络上的空间聚类和局部交互模式会显著影响博弈论情境中合作行为的演化 [41, 42, 43]。在每个时间步，每个智能体都会同时与全部四个邻居进行博弈。每个智能体的奖励是这四次交互所得的平均收益。环境按离散时间步演化。在每一步中，智能体观察局部状态、同时选择动作、获得奖励并更新策略。网络拓扑和收益矩阵保持固定。

除网格网络外，我们还测试了三种替代拓扑结构：模块化4正则、随机4正则和小世界网络。所有拓扑结构的平均度均保持为四，以确保公平比较。除非另有说明，不同拓扑结构下的收益结构、状态表示、学习算法和训练方案均保持一致，仅交互图有所不同。交互拓扑对合作结果的影响将在结果部分中考察。

2.2. 共享策略深度*Q*网络架构

所有智能体共享一个深度Q网络（DQN）。整个种群在学习和动作选择时都使用同一个动作价值函数。本文后续将这种架构称为*共享DQN*。采用参数共享是为了提升可扩展性和样本效率，使所有智能体都能通过同一个函数逼近器，将各自的局部观测映射为动作价值估计。

设 s_i^t 表示智能体 i 在时间步 t观测到的局部状态，$a \in \{C, D\}$ 表示可用动作。共享DQN对一个共同的动作价值函数进行近似

$$Q_\theta(s_i^t, a), \tag{2}$$

由 θ 参数化，所有智能体都使用它来选择动作。智能体会经历不同的局部情境，但在决策和学习时都使用同一个共享网络。

网络包含一个由 96 个 ReLU单元组成的隐藏层，以及一个线性输出层。输出层为两种动作（合作和

背叛）生成 Q值。为了稳定训练，会维护一个采用相同架构的独立目标网络 Q_{θ^-} ，并定期将其参数与在线网络的参数同步。在每个时间步，所有智能体都会生成经验元组（s， a， r， s'）。所有智能体的这些经验都会存入同一个共享经验回放缓冲区。系统从该缓冲区中均匀抽取小批量，使其中混合了来自所有智能体的经验。因此，梯度更新会汇总整个种群的信号，使所有智能体通过一个共享网络相互耦合。

共享DQN采用时序差分目标进行训练。对于一个转移 (s_t， a_t， r_t， s_{t+1})，单步目标定义为

$$y_t^{(1)} = r_t + \gamma\, Q_{\theta^-}\Big(s_{t+1}, \arg\max_{a'} Q_\theta(s_{t+1}, a')\Big), \tag{3}$$

其中， γ 是折扣因子， $y_t^{(1)}$ 表示单步时序差分目标值， a' 表示下一状态下可用动作的索引集合。该公式遵循双重深度Q网络原则，在在线网络和目标网络之间解耦动作选择与动作评估，以减少过高估计偏差。

为纳入长期信用分配，我们还采用n步回报。当可以获得来自同一智能体的一段有效转移序列时， n步目标定义为

$$y_t^{(n)} = \sum_{k=0}^{n-1} \gamma^k r_{t+k} + \gamma^n Q_{\theta^-}\Big(s_{t+n}, \arg\max_{a'} Q_\theta(s_{t+n}, a')\Big). \tag{4}$$

在我们的实现中，只要能够可靠地构建此类序列，我们就采用 n步回报，其中 $n = 5$；否则则退回到单步更新。这种混合策略可确保在异步经验收集过程中实现一致的学习。

我们使用带权重衰减 [44, 45]的AdamW优化器优化网络参数。损失函数定义为预测动作价值与目标动作价值之间的Huber损失，从而增强模型对时序差分误差中离群值的鲁棒性。训练期间，将梯度裁剪至固定范数，以提高数值稳定性。

我们还将一种分组变体作为诊断基线。在这种变体中，智能体被划分为 10 个互不重叠的组。每组都维护一个具有相同架构和学习过程的共享DQN。智能体被随机分配到各组（而非根据空间位置分配）。我们将这一

基线称为 分组 *DQN*。同一组内的智能体共享参数，但不同组之间不共享；环境和训练过程的所有其他组成部分均保持一致。这是一种诊断工具，用于检验全局表征共享是否导致了我们观察到的失败。

总而言之，这些架构定义了完全分散式执行方案，并在群体或种群层面进行集中式学习；本研究分析的主要学习架构是完全共享策略 DQN。

2.3. 状态表示与部分可观测性

每个智能体根据局部状态表示做出决策，该表示编码了与其直接邻居近期交互的信息。具体而言，智能体 i 在时间步 t 的基线状态定义为一个定长二进制向量

$$s_i^t = \left(a_{j_1}^{t-1}, a_{j_2}^{t-1}, a_{j_3}^{t-1}, a_{j_4}^{t-1}, a_i^{t-1}\right), \tag{5}$$

其中，j_1、.....、j_4 表示交互网络上智能体 i 的四个最近邻，$a \in \{0, 1\}$ 表示上一个时间步采取的动作，即合作或背叛。此状态仅包含最近一个时间步的信息，不包含 $t-1$ 之前的历史信息、全局信息或学习进度信号。

这种最简状态在空间演化博弈中十分常见。然而，当智能体同时进行学习时，它会导致部分可观测性。状态转移取决于邻居的策略，而这些策略会在训练过程中发生变化。因此，从每个智能体的视角来看，环境是非平稳的。重要的是，状态表示不编码任何有关训练阶段、探索强度或全局学习动态的信息。因此，即使底层群体层面的行为和探索机制随时间变化，采用共享策略的智能体仍必须仅凭局部交互信号来推断适当的动作。

在后续实验中，我们系统地为这一基线状态添加额外的低维信号，使共享表征能够获取潜在的时间信息或学习进度信息。这些增强状态通过将标量特征附加到原始的交互式状态上构造，同时保持学习架构、优化过程和环境动态不变。这种设计使我们能够在不引入额外算法修改的情况下，分离考察状态可观测性对学习结果的影响。

2.4. 探索机制

动作选择由基于 Softmax（玻尔兹曼）策略的随机探索机制控制。给定共享DQN生成的动作价值估计，每个智能体按照以下概率选择动作

$$\pi(a \mid s) = \frac{\exp(Q_\theta(s,a)/\tau)}{\sum_{a'} \exp(Q_\theta(s,a')/\tau)}, \tag{6}$$

其中，$\tau > 0$ 表示控制探索强度的温度参数。为确保数值稳定性，在计算 Softmax 之前，我们通过减去 $\max_a Q_\theta(s,\ a)$ 来应用标准的 logit平移技巧，这不会改变$\pi(a \mid s)$。较大的 τ 会使动作采样更加均匀，而较小的值则会使概率质量集中在价值较高的动作上。

温度参数 τ 会在训练过程中按照固定计划逐步退火。具体而言，τ 初始化为 τ_{init} ，并在预先设定的环境步数内线性递减至最终值 τ_{final} 。退火阶段结束后，在剩余的训练过程中，τ 保持固定为 τ_{final} 。所有智能体均全局同步地应用这一退火计划。

重要的是，在这种设置下，探索是在群体层面进行控制，而不是由每个智能体单独控制。在任意给定的时间步，所有智能体共享相同的温度参数，从而确保整个种群中动作选择的随机程度保持一致。这种设计使探索强度能够作为学习系统中的单一全局控制变量。

为便于对不同退火日程进行定量比较，我们用标量 B 概括有效的探索机制。该标量定义为训练初始阶段（此期间探索最为活跃）的平均温度。具体而言，

$$B = \frac{1}{K} \sum_{t=1}^{K} \tau(t), \qquad K = \left\lfloor \frac{T_{\text{anneal}}}{2} \right\rfloor,$$

其中，T_{anneal} 表示温度退火步数的总数。该标量以紧凑的方式刻画了给定退火计划所诱导的整体探索机制，同时抽象掉了 $\tau(t)$的细粒度时间细节。除非另有说明，本文报告的所有结果均以该探索强度参数为索引。

2.5. 仿真协议与实现细节

现在，我们介绍贯穿所有实验所采用的仿真方案与实现细节，并总结相关的学习过程和超参数设置。本文报告的所有结果均在完全同步多智能体交互协议下获得，并通过共享经验回放缓冲区和共享价值网络进行集中式学习。

除非另有说明，我们设定 $L = 30$，因此有 $N = 900$ 个智能体。每项实验包含固定数量的离散环境步，并分为两个不同阶段：采用退火探索的训练阶段，以及禁用学习的独立评估阶段。每次运行共包含 $T = 10^5$ 个环境步，其中训练阶段包含 $T_{\text{train}} = 95,\ 000$ 步，随后评估阶段包含 $T_{\text{eval}} = 5,\ 000$ 步。训练期间，所有智能体都会观测各自的局部状态，并根据式（6）定义的 softmax（玻尔兹曼）探索策略选择动作，探索温度 τ 会随时间退火。智能体之间的交互会生成经验元组，这些元组存储在共享经验回放缓冲区中，并用于通过时序差分学习更新共享Q网络。为改善信用分配，只要存在有效的转移序列，就采用 n 步回报；否则则回退到一步目标。目标网络参数按固定间隔与在线网络同步。

训练阶段结束后，将完全禁用学习更新。随后，系统进入评估阶段，在此期间，探索温度固定为 τ_{eval}，且不进行任何梯度更新或回放采样。所有报告的性能指标均仅在这一评估期间计算，以确保对学习得到的策略进行无偏评估。为量化集体行为，我们使用基于动作的序参量来衡量合作，其定义为种群所执行的合作动作占全部动作的比例。具体而言，在每个环境步 t，我们计算

$$c(t) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}[a_i(t) = C] \,, \tag{7}$$

其中，$\mathbb{I}[\cdot]$ 表示指示函数，$a_i(t) \in \{C, D\}$ 表示智能体 i 在时间 t选择的动作。该定义直接反映实际行为，并避免了对采用随机策略或混合策略的智能体进行分类时产生的歧义。

报告的合作水平是通过在评估阶段对 $c(t)$取平均得到的，

$$\bar{C} = \frac{1}{T_{\text{eval}}} \sum_{t=1}^{T_{\text{eval}}} c(t), \tag{8}$$

其中，$T_{\text{eval}} = 5,000$ 个环境步数对应于采用固定探索温度 τ_{eval} 且禁用学习的最终评估期。这一时间平均量作为我们的主要序参量，用于表征瞬态动力学消退后的稳态合作行为。

所有报告的结果均为多个独立模拟运行的平均值，这些运行使用了 30 个不同的随机种子。每次运行均采用相同的超参数、网络架构和探索计划，唯一的区别是初始随机种子；该种子决定网络初始化、动作采样和环境随机性。除非另有说明，图中展示的每个数据点均对应于基于这些独立随机种子计算出的平均合作水平，从而确保所报告的模式反映的是稳健的学习行为，而非个别随机效应。

算法1逐步说明了完整的训练和评估流程，而表1总结了实验中使用的网络架构、优化设置和所有超参数。除非另有明确调整，本研究始终固定这些设置。

算法1 共享策略DQN的两阶段交互协议。1：对于 $t = 1$、2、.....、T，执行 2：观察局部状态 $\{s_i^t\}_{i=1}^N$。3：如果 $t \leq T_{\text{train}}$，则执行 4：使用共享策略和退火探索$\tau(t)$选择动作 $\{a_i^t\}$。5：否则执行 6：使用共享策略和固定探索 τ_{eval}选择动作 $\{a_i^t\}$。7：结束条件判断 8：执行动作并观察奖励 $\{r_i^t\}$ 和下一状态 $\{s_i^{t+1}\}$。9：如果 $t \leq T_{\text{train}}$，则执行 10：将转移存入共享经验回放缓冲区。11：从回放样本中更新共享Q网络。12：定期更新目标网络并退火 $\tau(t)$。13：结束条件判断 14：如果满足 $t > T_{\text{train}}$，则记录评估指标。15：结束循环

表1：学习超参数和实现细节。

参数	值
网络架构	1 个隐藏层，96 个单元（ReLU 激活）
优化器	AdamW
AdamW 学习率	1×10^{-4}
AdamW 权重衰减	1×10^{-4}
AdamW 贝塔参数（β_1，β_2）	（0。9，0。999）
AdamW 艾普西隆 ϵ	1×10^{-8}
AdamW AMSGrad	否
AdamW 最大化	否
折扣因子 γ	0.99
回放缓冲区大小	90,000
小批量大小	256
目标网络更新	每 2000 个环境步数
n步回报	$n = 5$ （回退到 1步）
损失函数	Huber损失
梯度裁剪	ℓ_2 范数 ≤ 0.5
探索策略	Softmax（玻尔兹曼）
退火步数	95,000
评估温度 τ_{eval}	0.10

3. 结果

我们按层次组织结果，逐步揭示为何共享策略的深度Q网络（DQN）在探索过程中会在合作多智能体环境中失效。我们以合作水平作为学习稳定性和协调质量的指标，并系统地将合作与内部表征动态、价值景观及环境结构联系起来。

我们确立了五项主要发现。第一，随着探索程度和收益严苛度增加，合作会系统性地崩溃（图1）。第二，这种崩溃遵循可通过实证识别的阈值，且共享架构与分组架构之间存在差异（图 2）。第三，能够暴露学习进度的状态增强可以延缓合作崩溃，或使其部分逆转（图3）。第四，合作崩溃与学习到的表征退化（图4）以及动作价值差距收缩（图5）同时发生。第五，网络拓扑会调节这些效应：空间结构能够缓冲崩溃，而随机拓扑则会消除合作（图6）。

除了行为结果之外，我们还分析了共享网络学习到的内部表征，并表明合作崩溃与合作行为和背叛行为之间特征可分性的退化同时发生（图4）。在价值层面，我们进一步揭示，随着探索程度增加，动作价值差距会系统性收缩，这表明学习会趋向低合作吸引子，而非表现出不稳定或发散的学习动态（图5）。最后，我们考察了交互拓扑如何调节这些效应，结果显示，具有空间结构的环境能够提供隐性的协调支架，从而缓冲探索引发的溢出效应；相比之下，随机化拓扑基本会消除合作状态（图6）。总体而言，这些结果从多个层面连贯地刻画了非平稳多智能体交互下共享DQN的一种根本性失效模式。

3.1. 探索与收益参数中的合作崩溃

图1展示了两个维度上的合作水平：探索强度 B 和收益严苛度 D_r。每个单元格表示评估阶段30个随机种子的均值。通过这种表示方式，我们可以直观地观察合作行为如何随着探索和激励不对称性的共同增加而演变。随着 B 或 D_r 增加，合作水平会单调下降。参数空间的右上方形成了一片范围较大的低合作区域。这一转变是连续的，而非突然发生的。这表明合作崩溃具有规律性，而不是由随机初始化导致的零星失败。

随着探索增加，崩溃区域平滑扩展。在较低的 B下，即使收益有利于背叛，合作仍保持在较高水平。超过临界值 B后，随着 D_r 增加，合作会迅速下降。失败并非由任何单一因素决定，而是取决于这些因素之间的相互作用。

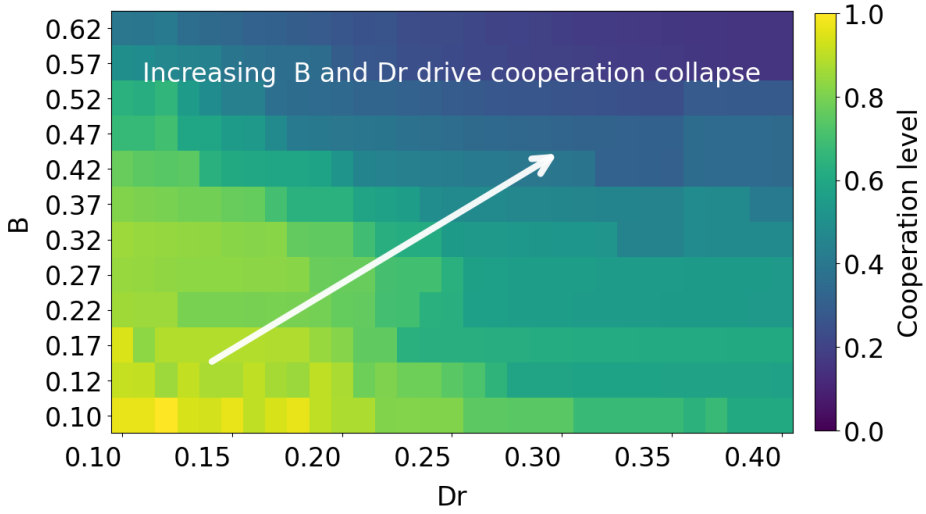

图1：共享策略深度Q网络中的合作水平随探索强度B和收益严苛度 D_r的变化。每个单元格表示评估期间30个随机种子的均值。随着任一参数增加，合作都会系统性下降，并在右上方形成一片范围广阔的低合作区域。平滑的边界表明这是有规律的崩溃，而非偶发的失败。

3.2. 实证崩溃阈值与表征共享的作用

我们使用每个探索水平对应的经验阈值 $D_r^*(B)$来概括崩溃情况。该阈值是合作仍高于标准水平时， D_r 值的最大值（在我们的测试范围 $[0.10, 0.40]$内）。我们对 30 个随机种子取平均，以确保阈值稳健。

图 2 展示了两种学习架构对应的崩溃阈值。图 2(a) 对应完全共享策略的 DQN，其中所有智能体都更新同一个共享价值网络。图 2(b) 展示了分组基线：智能体被随机划分为十个互不重叠的组，每组各自维护一个独立的共享价值函数。这种分组设计限制了表征共享的范围，同时保留了组内协调，为完全共享和完全独立学习提供了中间对照。

对于共享策略深度Q网络，提取出的阈值 $D_r^*(B)$ 会随着探索的增加而单调下降。更高程度的探索会逐步缩小能够维持合作的收益条件范围，使崩溃阈值在 B 范围内平滑且系统性地变化。重要的是，这一趋势在所探索的参数空间中始终一致出现，而非表现为孤立或不规则的崩溃。

分组 DQN 则呈现出不同的模式。在较低的 B 下，阈值像共享DQN 一样下降。然而，在较高的 B 下，阈值趋于稳定并有所回升。将表示共享限制在较小的组内，可以减轻探索的影响。两种架构采用不同的合作阈值来界定崩溃：共享DQN 为 0.55，分组 DQN 为 0.15。这是必要的，因为分组 DQN 的基线合作水平要低得多（见图 2）。选择这两个阈值，是为了确保它们在我们测试的 D_r 范围 $[0.10, 0.40]$ 内与曲线相交。这些不同的阈值不会影响定性比较，因为我们比较的是趋势（阈值如何随 B 变化），而非绝对值。

综上，这些结果表明，共享DQN中的合作崩溃可以通过崩溃阈值得到结构化的经验刻画；而当表征共享受到限制时，在完全共享架构中观察到的单调恶化会显著减弱。这一对比凸显了表征聚合在塑造探索诱发的非平稳性如何在系统层面显现方面所发挥的核心作用。

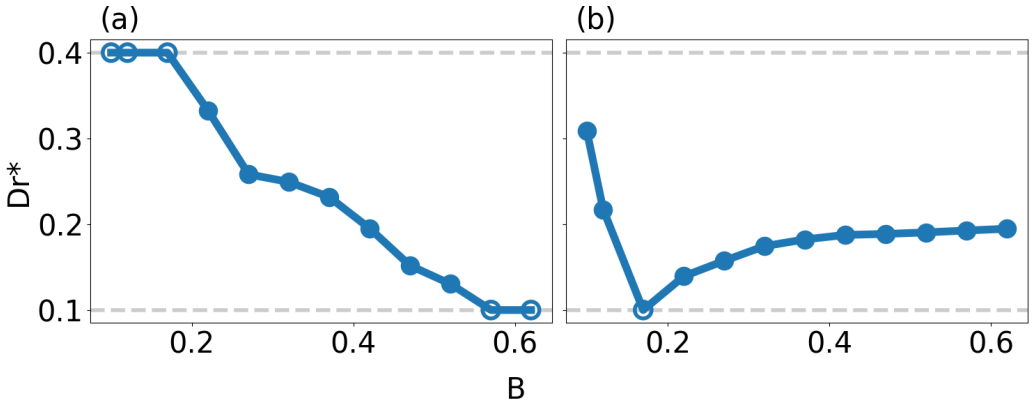

图 2：崩溃阈值 $D_r^*(B)$ 随探索强度 B 的变化（对 30 个随机种子取平均）。（a）共享 DQN：所有智能体使用同一个网络。阈值随 B 单调下降（在合作水平 $= 0.55$ 处识别）。（b）分组 DQN：将 90 个智能体分为 10 组，每组使用独立网络。阈值与 B 呈非单调关系，在较高 B 下有所回升（在合作水平 $= 0.15$ 处识别）。不同的阈值反映了不同的基线合作水平。虚线标示测试的 D_r 范围 $[0.10, 0.40]$。空心标记表示阈值位于测试范围之外。

3.3. 状态增强延缓合作崩溃

合作崩溃是否是因为智能体无法观察学习动态? 我们通过在状态中加入有关探索和训练进度的信号来检验这一点。在这一基线下，随着B 增加，合作持续下降。

基线共享DQN采用最简状态：智能体四个邻居的最近一次动作，以及自身的最近一次动作。这是一个5维二进制向量，不包含有关时间、训练进度或探索水平的信息。在这一基线设置下，随着 B 增加，合作持续下降（图3中的红色曲线）。相较于基线，不同形式的状态增强对探索条件下的合作产生了不同影响。基线共享DQN中的合作会随着探索强度 B 增加而单调下降，反映出协调逐步瓦解。相比之下，使用时间探索信号 τ 增强状态后，在低至中等探索水平下，合作略有改善；但随着 B 进一步增加，合作迅速恶化，最终在高探索水平下低于基线。这表明，仅凭时间信息不足以稳健地稳定共享学习。

使用训练进度信号增强状态，则会呈现出性质不同的模式。在一个狭窄的中等探索水平范围内，合作显著提升；但这种改善高度依赖于具体情形且不稳定，在更高的 B下，合作再次急剧下降。相比之下，同时纳入时间信号和训练进度信号的联合增强，带来了最稳健的稳定效果。与基线相比，联合增强下的合作随着探索强度B 增加而保持稳定且略有上升的趋势，同时对探索诱导的退化明显不那么敏感。这一表现表明，联合编码多种潜在时间结构对于在探索条件下稳定共享表征至关重要。

综合来看，这些结果提供了因果证据，表明在基线共享DQN中观察到的合作崩溃，与非平稳学习动态的部分可观测性密切相关。通过将潜在的时间信号显式暴露给共享表征，状态增强显著提升了在探索过程中的鲁棒性，这表明该失败并非完全由收益激励或算法不稳定性驱动，而是由状态信息不足所导致的表征局限性引起的。

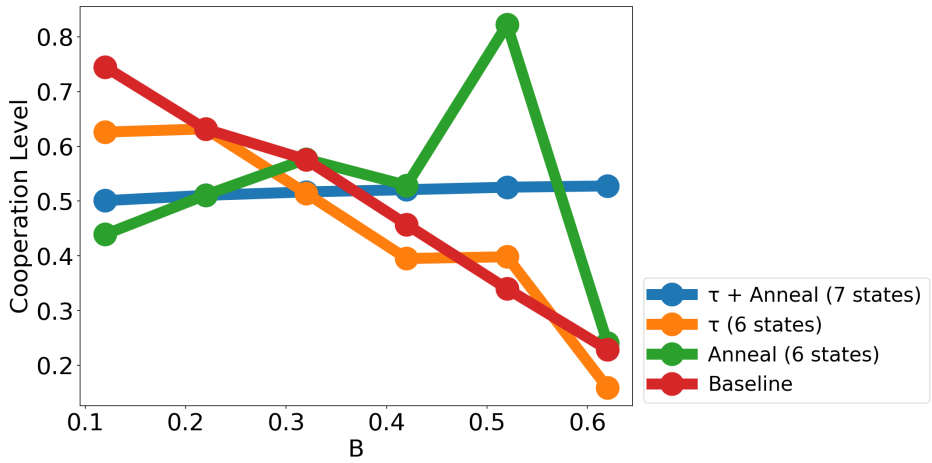

图3：不同探索强度 B下，状态增强对共享策略DQN合作结果的影响。所有曲线均表示在30个独立随机种子上取平均得到的合作水平。基线曲线对应原始共享DQN，其状态仅编码局部交互信息，包括每个智能体自身最近的动作以及其邻居最近的动作。我们将该基线与增强变体进行比较，后者在原始状态向量中附加了额外的标量信号：时间探索信号 τ、粗粒度的训练进度（退火）指标，以及二者的联合加入。所有配置均采用相同的学习算法、网络架构、奖励结构和交互拓扑；不同之处仅在于共享价值函数可获取的信息。相较于基线，联合增强表现出最稳健的行为：随着探索强度 B 增加，合作水平呈现稳定且缓慢上升的趋势，同时显著降低了对探索诱导退化的敏感性。相比之下，单一信号增强产生的影响更加有限，并且取决于所处阶段。

3.4. 探索下共享网络中的表示坍塌

减少表征共享或采用状态增强可以延缓合作崩溃。现在，我们考察共享网络学习到的内部表示。合作的变化是否对应于学习到的特征变化？为此，我们分析共享DQN在不同探索强度 B 下与合作动作和背叛动作相关的隐藏层激活。图4使用UMAP对这些表示进行了可视化。UMAP是一种非线性降维技术，在此仅用于定性检查表示几何。各点根据智能体所选择的动作进行着色。为了定量刻画隐藏表示的组织结构，我们基于对隐藏层激活进行的无监督双簇划分计算轮廓系数。

示。具体而言，对于每个探索水平 B，使用带有 $k = 2$的 k均值聚类将隐藏表示划分为两个簇，不使用动作标签。对于每个表示向量 i，轮廓系数定义为

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}},$$

其中， a（i）表示 i 与同一簇内其他点之间的平均距离， b（i）表示 i与最近邻簇中各点之间的最小平均距离。每个 B 对应的轮廓分数，是对所有采样表示的 s（i）取平均后得到的。

轮廓系数随 B呈现非单调演化。在中等B下，随着表示聚为两组，分数有所上升。然而，这并不意味着动作判别能力有所提升，而是反映出系统正收敛至背叛主导状态，其中大多数状态都映射为相似特征。在强探索下，这种结构会逐渐瓦解。潜在空间变得愈发弥散，合作表示收缩至稀疏且不稳定的区域，而背叛表示则主导流形。相应地，在较高 B下，轮廓系数下降，表明隐藏空间中连贯的簇结构有所丧失。重要的是，尽管训练在数值上保持稳定且表面上已收敛，这种崩塌仍会发生，因此可排除优化失败或发散是其成因。

轮廓系数的非单调变化与行为层面观察到的合作率单调下降形成对比。这一差异凸显了隐藏表示的可聚类性与动作判别可分性之间的关键区别。虽然中等程度的探索可能会暂时提升主导性背叛倾向模式周围的表示一致性，但并不会增强网络支持合作决策的能力。

结合图3中的干预结果，这些发现表明，共享DQN的主要脆弱性存在于表征学习层面。当探索诱发的非平稳性在智能体之间汇总，却没有明确的时间编码或结构性解耦时，共享网络便无法维持稳定且动作对齐的内部表示，最终损害持续合作。

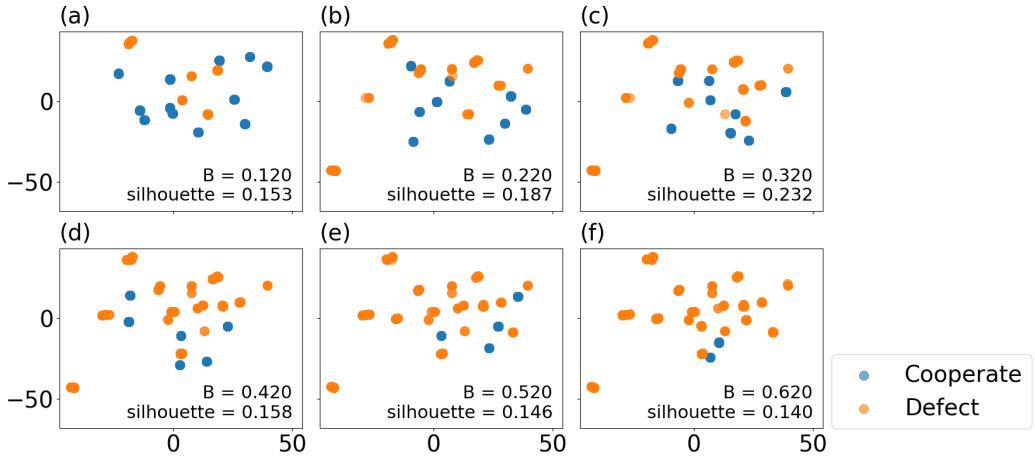

图4：随着探索强度不断增加，共享策略深度Q网络学习到的隐藏层表示的可视化结果 B。每个面板展示了智能体状态所对应隐藏激活的二维 UMAP 投影，点的颜色表示所选动作（合作或背叛）。在探索程度较低时，合作动作与背叛动作在投影后的潜在空间中占据部分不同的区域。在中等探索强度下B，表示的几何聚类性有所增强；而在强探索下，潜在空间变得愈发分散，合作表示则不断收缩并失去稳定性。基于隐藏表示的无监督双簇划分计算得出的轮廓系数，量化了簇结构的这种非单调演化。重要的是，轮廓系数反映的是几何聚类性，而非动作判别可分性；其在中等探索强度下 B 的短暂上升，并不意味着对合作决策的支持有所改善。

3.5. 价值景观收缩与向低合作吸引子的收敛

前述分析表明，共享DQN中的合作崩溃伴随着内部表示的退化。我们现在探究，这种崩溃是否与不稳定或发散的学习动态有关，还是反映了系统收敛到一个性质迥异且更压缩的价值景观。本节报告的所有价值统计数据，均通过对多个独立模拟运行取平均值计算得出；这些模拟运行使用了30个不同的随机种子，以确保观测到的价值景观收缩反映的是稳健的收敛行为，而非特定随机种子造成的伪象。

图5展示了两个随探索强度 B 变化的互补统计量，其中背叛损失固定为 $D_r = 0.25$：(a) 平均动作价值幅度（Q均值），定义为对动作、智能体和遇到的状态取平均后的动作价值绝对值均值

；(b) 合作动作与背叛动作之间的平均动作价值差距（Q差距）。Q差距定义为两个动作的估计 Q 值之差的绝对值均值。这两个量均在关闭学习的留出评估阶段计算，以确保所报告的统计量反映的是收敛后价值函数的几何结构，而非瞬态训练动态。

随着 B 增大，Q差距呈现出明显且单调的下降趋势，并在探索水平较高时趋近于零。这一趋势表明，合作动作与背叛动作之间的价值区分能力正系统性地减弱。相比之下，Q均值并未单调下降：在最初有所降低后，它稳定在一个有界范围内，并呈现轻微的非单调变化。关键的是，在所有探索机制下，Q均值始终有限且表现稳定，没有发散或振荡不稳定的迹象。

综合来看，这些观测排除了训练不稳定性或数值发散是合作崩溃主要驱动因素的可能。相反，增加探索的主导效应是使相对价值景观收缩：尽管绝对动作价值幅度仍然存在，但动作之间的差异优势逐渐消失。因此，共享DQN能够稳定地收敛到稳定但价值区分能力较弱的策略，在这些策略下，合作动作不再具有稳固的价值优势。

这种收敛对应于低合作吸引子，其特征是动作价值差异被压缩，而非学习动态紊乱。在这种机制下，探索会重塑学得的价值几何，却不会破坏优化的稳定性，从而导致策略选项之间变得无差别。

结合图4中的表征层面分析，这些结果表明，共享DQN中的合作崩溃源于价值判别能力与潜在结构的协同退化，而非收敛失败或数值稳定性问题。

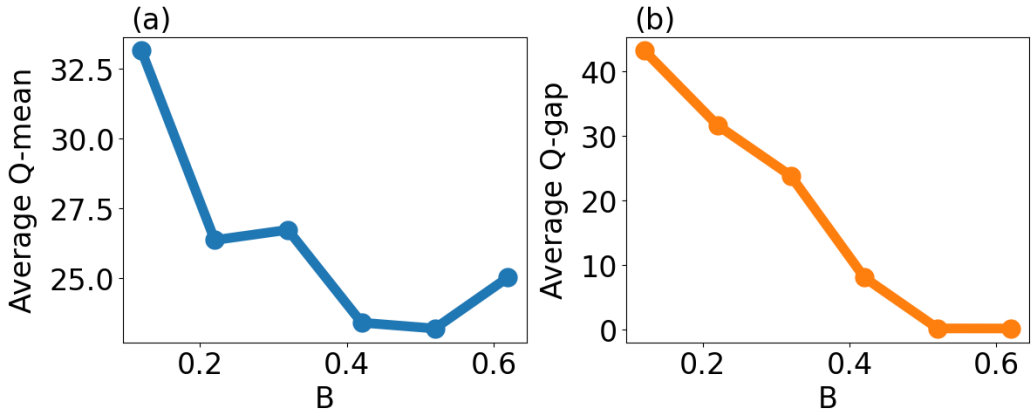

图5：在留出评估阶段（关闭学习）评估的共享策略深度Q网络的平均动作价值统计随探索强度 B的变化情况，背叛损失固定为 $D_r = 0.25$。所有统计量均对 30 个独立随机种子的结果取平均值。(a) 平均动作价值幅度（Q均值），计算方法是对动作、智能体和遇到的状态取动作价值绝对值的平均值。(b) 平均动作价值差距（Q差距），定义为合作动作与背叛动作的估计 Q值之间的绝对差值的平均值。随着探索增加，Q差距单调下降并趋近于零，而 Q均值保持有限且有界，这表明相对价值差异在收缩，而非学习动态不稳定或发散。

3.6. 探索诱发的合作崩溃的拓扑结构鲁棒性

接下来，我们评估共享策略深度Q网络中探索诱发的合作崩溃在不同交互拓扑下的稳健性，包括网格、模块化4正则、随机4正则和小世界网络。对于所有考察的拓扑结构，我们将平均节点度固定为4，因此合作动态的差异仅归因于结构组织方式，而非连接密度。本节报告的所有合作曲线均取自多次独立模拟运行的平均值，每次运行使用30个不同的随机种子，以确保观察到的跨拓扑差异反映的是稳健的结构效应，而非特定随机种子造成的波动。

图6展示了在固定诱惑参数 $D_r = 0.25$ 下，平均合作水平随探索强度 B 的变化，并根据训练收敛后的评估阶段数据计算得出。为确保不同拓扑之间的比较公平，所有网络的智能体总数均与网格基线相同，区别仅在于连接模式。在网格、模块化4正则和随机4正则拓扑中，我们观察到一致的定性模式：合作水平在较低的探索水平下仍

保持相对较高（$B \leq 0.2$），随后随着 B 增加而显著且单调地下降。值得注意的是，这三种局部均质拓扑的合作曲线在整个 B 范围内几乎重合，这表明局部连接方式的适度变化——无论是规则、模块化还是随机连接——都不会实质性地改变共享DQN的崩溃动力学。

相比之下，小世界拓扑呈现出截然不同的行为。随着 B 增加，各探索强度下的合作水平始终普遍偏低，变化也很小。与其他拓扑结构不同，即使探索程度较低，也未能形成合作行为。这表明，在全局共享价值函数下，远程连接的存在会阻碍稳定局部协调的形成。

总而言之，图6表明，共享策略深度Q网络中的合作崩溃基本不受局部结构变化的影响，但在促进快速全局混合的拓扑结构下会显著加剧。这些结果支持这样一种解释：观察到的崩溃反映了探索条件下共享表征学习的内在局限，而非特定网络构造造成的假象。这也补充了我们此前的证据，即交互图组织和适应机制会深刻影响由学习驱动的社会困境中的合作状态形成 [24]。

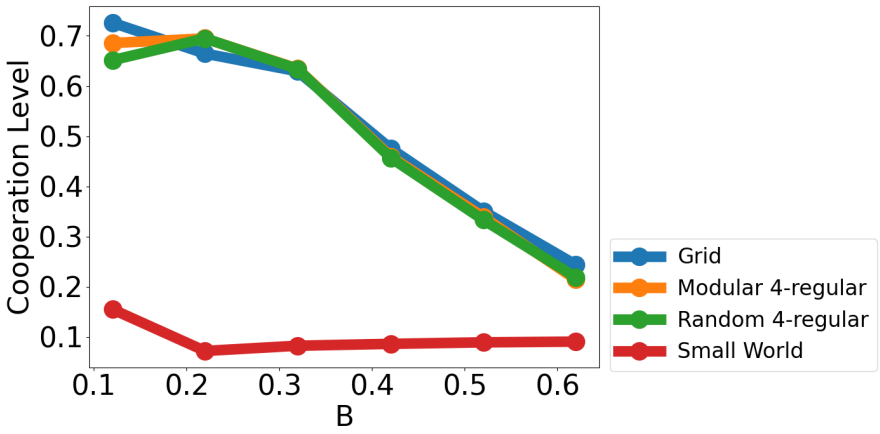

图6：在不同交互拓扑下，共享策略深度Q网络中平均合作水平随探索强度 B 的变化，包括网格、模块化4正则、随机4正则和小世界网络。所有曲线均取30个独立随机种子的平均值。所有网络均以平均度为4构建，并具有相同的智能体总数，从而确保对不同交互结构进行受控比较。所有结果均在固定诱惑参数 $D_r = 0.25$ 下进行评估，并在训练收敛后的评估阶段计算。所有拓扑结构均使用与网格网络相同的智能体总数，以隔离交互结构与种群规模的影响。尽管网格、模块化和随机拓扑呈现出几乎重叠的合作曲线，并随着探索强度增加而平滑下降，但小世界拓扑在 B 的整个范围内始终保持较低的合作水平，表明在全局混合交互结构下，共享DQN对探索的脆弱性更高。

4. 讨论

4.1. 哪些方面会失败，哪些不会

共享DQN中的合作崩溃并非源于不稳定学习或数值发散。在所有测试条件下，学习过程都表现稳定。价值估计始终有界，训练可靠收敛，评估指标也很稳定。问题不在于系统是否收敛，而在于它收敛到什么状态。共享DQN总能找到稳定的解，但这些解无法维持合作。这一区别很重要。性能退化常被归咎于噪声过大、调参不佳或优化问题。我们的结果排除了这三种可能。即使学习率、损失函数、优化器和缓冲区大小各不相同，合作崩溃依然存在（见附录）。动作价值差距平稳且单调地收缩，并非毫无规律地波动。

这种失效并非仅仅关乎激励。即使中间收益保持不变，随着探索增加，合作仍会崩溃。仅改变激励无法解释这一模式。这排除了简单的均衡转变。合作崩溃源于学习动态本身。这与早期研究的发现相呼应：独立学习者在缺乏协调时难以取得成功 [46]。

我们将合作崩溃重新界定为一种结构性失效模式。它源于共享表征在探索过程中如何处理非平稳多智能体学习。在合作马尔可夫博弈中，独立学习与协-调之间的矛盾已得到广泛记录 [47],；在这类博弈中，使用标准单智能体算法的智能体往往无法收敛到全局最优解。

4.2. 共享表征作为结构性瓶颈

共享策略学习存在结构性矛盾。尽管所有智能体经历着不同的局部情境，学习速率也各不相同，它们仍共同贡献并使用同一个价值函数。这提高了可扩展性和样本效率。然而，它并未抵消非平稳性，反而汇集了各智能体的非平稳性。这种对学习信号的汇集，代表着多智能体深度强化学习面临的一项根本挑战：从每个智能体的视角来看，环境都变得非平稳 [48]。每个智能体的探索都会扰动其邻居所处的环境。这些扰动经由共享经验回放缓冲区传入梯度更新。共享网络会接收到来自处于不同学习阶段的智能体之间相互冲突的信号。随着探索增加，这些冲突会加剧，并且无法在单一表征中得到化解。

我们的表征层面分析揭示了这种聚合带来的后果。合作崩溃与动作对齐的潜在结构退化以及动作价值差异系统性收缩同时发生，这表明系统正收敛到稳定但区分能力较弱的价值景观。这种收敛并不意味着优化失败。相反，它反映了表征干扰：共享网络必须在固定容量内兼顾来自不同智能体且彼此不兼容的信号 [49]。

这种失效是表征层面的瓶颈。在参数空间中会出现明确的崩溃阈值，类似于耦合动力系统中的相变 [50]。这表明，该失效是架构特征，而非调参伪象。当一种表征无法在不同交互情境下为各智能体保留可区分特征时，就会发生崩溃。

4.3. 部分可观测性会加剧失败

部分可观测性会加剧表征瓶颈。标准状态仅编码近期动作，并假设环境是平稳的。在并发学习过程中，这一假设并不成立。

同一局部状态在不同训练阶段可能意味着不同的事情。训练初期，四个合作邻居可能表示合作稳定。训练后期，同一观测则可能表示背叛状态中的暂时噪声。缺乏时间上下文时，网络无法区分这两种情况。这会加剧表征干扰。状态增强提供了因果证据。在状态中加入探索水平和训练进度，可在不改变算法的情况下提高可观测性。这能延缓所有探索水平下的合作崩溃，并在某些状态下部分逆转崩溃。该干预确实行之有效，证明了状态信息至关重要。

然而，状态增强是一种诊断工具，而非解决方案。它虽能减少冲突，却无法消除耦合。所有智能体仍共用一个网络。部分可观测性会加剧这种结构性局限，但并非唯一原因。

4.4. 为什么缓解措施有所帮助，却无法消除失败

缓解策略能够减轻失败，却无法彻底解决问题。分组学习限制了参数共享。状态增强提高了可观测性。两者都能扩大合作得以维持的范围，但都无法消除根本限制。

结构性约束依然存在。即使使用增强状态，所有智能体仍共用一个价值函数。不同智能体产生的梯度依然会相互干扰。经验回放还会带来额外问题：存储的转移反映的是其他智能体过时的策略 [51]。分组架构能够降低耦合，但无法将其消除。它们以小规模共享取代全局共享。在高强度探索下，即使采取缓解措施，系统仍会崩溃。共享表征最终会超出其容量限制。无论采用多少状态增强或部分解耦，都无法突破这一限制。

这些发现阐明了，为何在强非平稳性条件下，合作崩溃应被视为共享策略学习的一项根本局限，而非可通过渐进式调优或辅助输入加以解决的缺陷。缓解策略能够扩大鲁棒性裕度，但无法消除全局参数共享所造成的表征瓶颈。

4.5. 对可扩展多智能体强化学习设计的启示

我们的研究结果对多智能体强化学习系统设计有四点启示。第一，参数共享并非没有代价。它能够提升样本效率，但也会聚合非平稳性。设计者必须明确管理可扩展性与表征一致性之间的权衡。

第二，稳健性要求观察学习动态，而不仅仅是环境状态。标准的马尔可夫状态假设平稳性，但在并发学习下这一假设会失效。以训练阶段或探索水平为条件来设定策略，能够提升稳定性。

第三，表征模块化至关重要。该失败源于强迫单一表征处理所有智能体和学习阶段。模块化、分层或部分解耦的架构，可能更好地局部化非平稳性。

最后，多智能体强化学习中的收敛诊断不应局限于奖励轨迹。我们的结果表明，稳定收敛可能与价值几何和动作判别的定性退化并存。因此，监测表征结构和相对价值景观对于评估共享策略系统的稳健性至关重要。

4.6. 该诊断的范围、局限性与适用性

我们研究了共享策略深度Q网络，将其作为一种代表性的基于价值的多智能体强化学习方法。许多发现可能适用于其他采用完全参数共享的架构。不过，采用集中式评论家、对手建模或模块化表征的方法或许能够避免这些失效。

动态囚徒困境在此用作诊断环境，而非全面的基准。其简单性便于对探索、激励和拓扑结构进行受控操纵，从而能够分离出表征层面机制。尽管不同任务的定量结果可能有所不同，但这里发现的定性矛盾——共享表征、非平稳性与探索之间的矛盾——并非这款博弈所特有。

重要的是，本研究所记录的合作崩溃并非数值不稳定或未能收敛所致。在所有情形下，共享DQN都能可靠地收敛到定义明确的价值函数。因此，此处所指出的局限在于定性层面，涉及所学解的性质，而非解是否存在。

我们的结果揭示了一种未得到充分重视的风险：共享表征难以应对非平稳多智能体学习。了解这一失效模式，有助于设计更稳健的多智能体强化学习系统。

数据可用性

所有仿真和分析所用的源代码均已公开发布于作者的 GitHub。仓库中包含原始仿真数据、网络参数和可复现性脚本。使用所提供的代码和有记录的随机种子，即可复现所有结果。

参考文献

[1]Zhang-Wei Hong、Tzu-Yun Shann、Shih-Yang Su、Yi-Hsiang Chang、Tsu-Jui Fu 和 Chun-Yi Lee。用于深度强化学习的多样性驱动探索策略。神经信息处理系统进展，31，2018年。[2] Weiming Shen、Douglas H Norrie 和 Jean-Paul Barthès。面向并发智能设计与制造的多智能体系统。CRC出版社，2019年。[3] Jianye Hao、Tianpei Yang、Hongyao Tang、Chenjia Bai、Jinyi Liu 和 Zhaopeng Meng。深度强化学习中的探索：从单智能体到多智能体领域。*IEEE*神经网络与学习系统汇刊，35(7):8762–8782，2024年。[4]Wenhui Huang、Cong Zhang、Jingda Wu、Xiangkun He、Jie Zhang 和 Chen Lv。通过偏好引导的随机探索实现采样高效的深度强化学习。*IEEE*神经网络与学习系统汇刊，35(12):18553–18564，2024年。[5]Pablo Hernandez-Leal、Michael Kaisers、Tim Baarslag 和 Enrique Munoz De Cote。多智能体环境中的学习综述：应对非平稳性。*arXiv*预印本 *arXiv:1707.09183*，2017年。[6] Andrei Marinescu、Ivana Dusparic 和 Siobhán Clarke。在固有非平稳环境中的基于预测的多智能体强化学习。*ACM*自主与自适应系统汇刊（*TAAS*），12(2):1–23，2017年。[7] Annie Xie、James Harrison 和 Chelsea Finn。在持续结构化非平稳性中进行深度强化学习。载于国际机器学习会议，第11393–11403页。PMLR，2021年。

[8]Christian Alexander Steinparz、Thomas Schmied、Fabian Paischer、Marius-Constantin Dinu、Vihang Prakash Patil、Angela Bitto-Nemling、Hamid Eghbal-zadeh 和 Sepp Hochreiter。通过反应式探索应对终身强化学习中的非平稳性。载于*终身学习智能体会议*，第441–469页。PMLR，2022年。[9] Zihe Liu、Jie Lu、Junyu Xuan 和 Guangquan Zhang。在真实非平稳环境中学习潜在且不断变化的动态。*IEEE知识与数据工程汇刊*，2025年。[10] Diogo S Carvalho、Pedro A Santos 和 Francisco S Melo。在渐近收敛非平稳环境中的强化学习：分层结构与学习到的表征。人工智能，第104382页，2025年。[11] Filippos Christianos、Georgios Papoudakis、Muhammad A Rahman 和 Stefano V Albrecht。通过选择性参数共享扩展多智能体强化学习。载于国际机器学习会议，第1989–1998页。PMLR，2021年。[12] Dapeng Li、Na Lou、Bin Zhang、Zhiwei Xu 和 Guoliang Fan。面向多智能体强化学习的自适应参数共享。载于*ICASSP 2024-2024 IEEE声学、语音与信号处理国际会议（ICASSP）*，第6035–6039页。IEEE，2024年。[13] Xuan Zhang、Hengxi Zhang、Huaze Tang、Le Liang、Ling Cheng、Xinlei Chen、Wenbo Ding 和 Xiao-Ping Zhang。用于多目标车联网资源分配的可扩展平均场多智能体强化学习框架。*IEEE智能车辆汇刊*，2024年。[14] Shanqi Liu、Weiwei Liu、Wenzhou Chen、Guanzhong Tian、Jun Chen、Yao Tong、Junjie Cao 和 Yong Liu。通过考虑队友的动作来学习多智能体协作。*IEEE神经网络与学习系统汇刊*，35(8):11553–11564，2024年。[15] Yuhang Pei、Tao Ren、Yuxiang Zhang、Zhipeng Sun 和 Matys Cham-peyrol。用于多智能体强化学习高效分散式执行的策略蒸馏。《神经计算》，626:129617，2025年。

[16] Sangho Lee、Youngjae Yu、Gunhee Kim、Thomas Breuel、Jan Kautz 和 Yale Song。用于视频表征学习的参数高效多模态变换器。*arXiv*预印本 *arXiv:2012.04124*，2020年。[17]Marco Fumero、Florian Wenzel、Luca Zancato、Alessandro Achille、Emanuele Rodolà、Stefano Soatto、Bernhard Schölkopf 和 Francesco Locatello。利用稀疏且共享的特征激活进行解缠表征学习。神经信息处理系统进展，36:27682–27698，2023年。[18] Sindhu Padakandla、Prabuchandran KJ 和 Shalabh Bhatnagar。面向非平稳环境的强化学习算法。*AppliedIntelligence*，50(11):3590–3606，2020年。[19] Spiros Kapetanakis 和 Daniel Kudenko。异构合作多智能体系统中的强化学习协同。载于自适应智能体与多智能体系统研讨会，第119–131页。施普林格，2003年。[20] Meng-Lin Li、Shaofei Chen 和 Jing Chen。自适应学习：一种用于合作多智能体系统的新型分散式强化学习方法。*IEEE Access*，8:99404–99421，2020年。[21] Weize Chen、Yusheng Su、Jingwei Zuo、Cheng Yang、Chenfei Yuan、Chi-Min Chan、Heyang Yu、Yaxi Lu、Yi-Hsin Hung、Chen Qian 等。Agentverse：促进多智能体协作并探索涌现行为。载于*ICLR*，2024年。[22] Anuran Pal 和 Supratim Sengupta。网络重连促进愿景学习模型中的合作。*Chaos: An InterdisciplinaryJournal of Nonlinear Science*，32(2)，2022年。[23] Hsuan-Wei Lee、Szu-Ping Chen 和 Feng Shi。通过基于强化学习的重连策略增强动态网络中的合作。*New Journal of Physics*，27(1):013025，2025年。[24] Yi-Ning Weng 和 Hsuan-Wei Lee。基于Q学习的自适应重连，用于异构网络中的协作控制。*arXiv*预印本 *arXiv:2509.01057*，2025年。

[25] Francisco C Santos、JF Rodrigues 和 Jorge M Pacheco。图的拓扑结构对合作演化起决定性作用。英国皇家学会学报*B*辑：生物科学，273(1582):51–55，2006年。[26] Brandon J Kinne。网络动力学与国际合作的演化。《美国政治学评论》，107(4):766–785，2013年。[27] David G Rand、Martin A Nowak、James H Fowler 和 Nicholas A Christakis。静态网络结构能够稳定人类合作。《美国国家科学院院刊》，111(48):17093–17098，2014年。[28]Babak Fotouhi、Naghmeh Momeni、Benjamin Allen 和 Martin A Nowak。具有社群结构的大型网络上的合作演化。《英国皇家学会界面》，16(152):20180677，2019年。[29] Ruey-An Shiu 和 Hsuan-Wei Lee。通过演化博弈塑造网络结构以增强合作。《英国皇家学会学报*A*辑：数学、物理与工程科学》，481(2322)，2025年。[30] Martin A Nowak 和 Robert M May。演化博弈与空间混沌。《自然》，359(6398):826–829，1992年。[31] György Szabó 和 Gabor Fath。图上的演化博弈。《物理学报告》，446(4-6):97–216，2007年。[32] Chao Yu、Minjie Zhang、Fenghui Ren 和 Guozhen Tan。空间社会困境中的情感多智能体强化学习。*IEEE*神经网络与学习系统汇刊，26(12):3083–3096，2015年。[33] Friederike Mengel。风险与诱惑：关于囚徒困境博弈的元研究。《经济学杂志》，128(616):3182–3209，2018年。[34] Hsuan-Wei Lee、Nishant Malik 和 Peter J Mucha。在自适应网络上共同演化的囚徒困境博弈。复杂网络期刊，6(1):1–23，2018年。[35] Sindhu Padakandla、K. J. Prabuchandran 和 Shalabh Bhatnagar。面向非平稳环境的强化学习算法。*AppliedIntelligence*，50:3590–3606，2020年。

[36]Hsuan-Wei Lee 和 Yi-Ning Weng。粒度化 Q 学习适应提升了多智能体囚徒困境中的集体福利。*Chaos, Soli-tons & Fractals*，199:116642，2025年。[37] Yini Geng、Yifan Liu、Yikang Lu、Chen Shen 和 Lei Shi。强化学习解释了多种条件合作。*Applied Mathe-matics and Computation*，427:127182，2022年。[38] Zhen Wang、Satoshi Kokubo、Jun Tanimoto、Eriko Fukuda 和 Keizo Shigaki。深入了解所谓的空间互惠。*Physical ReviewE*，88:042145，2013年。[39] Jun Tanimoto 和 Hiroki Sagara。双人对称博弈中困境的出现与弱占优策略存在之间的关系。*Biosystems*，90:105–114，2007年。[40] Zhen Wang、Satoshi Kokubo、Marko Jusup 和 Jun Tanimoto。演化博弈中困境强度的普适标度关系。*Physics ofLife Reviews*，14:1–30，2015年。[41] MN Kuperman 和 S Risau-Gusman。网络中的聚类系数与合作成功之间的关系。*Physical ReviewE—Statistical, Nonlinear, and Soft Matter Physics*，86(1):016104，2012年。[42] Hsuan-Wei Lee、Colin Cleveland 和 Attila Szolnoki。何时有成本迁移有助于提升合作。*Chaos: An InterdisciplinaryJournal of Nonlinear Science*，32(9)，2022年。[43] Zehua Si、Takayuki Ito 和 Hsuan-Wei Lee。具有迁移和互动多样性的空间公共物品博弈中的合作演化。*Applied Mathematics and Co mputation*，506:129544，2025年。[44] Ricardo Llugsi、Samira El Yacoubi、Allyx Fontaine 和 Pablo Lupera。比较 Adam、Adamax 和 AdamW 优化器，用于实现基于神经网络的安第斯城市基多天气预报。载于*2021*年 *IEEE* 第五届厄瓜多尔技术分会会议 *(ETCM)*，第 1–6 页。I EEE，2021年。[45] Mohamed Reyad、Amany M Sarhan 和 Mohammad Arafa。用于深度神经网络优化的改进 Adam 算法。*Neural Comput-ing and Applications*，35(23):17095–17112，2023年。

[46] Ming Tan。多智能体强化学习：独立智能体与合作智能体。载于第十届国际机器学习会议论文集，第 330–337 页，1993年。[47] Laetitia Matignon、Guillaume J Laurent 和 Nadine Le Fort-Piat。合作马尔可夫博弈中的独立强化学习者：关于协调问题的综述。《知识工程评论》，27(1):1–31，2012年。[48] Pablo Hernandez-Leal、Bilal Kartal 和 Matthew E Taylor。多智能体深度强化学习综述与评析。《自主智能体与多智能体系统》，33(6):750–797，2019年。[49]Craig Atkinson、Brendan McCane、Lech Szymanski 和 Anthony Robins。伪复习：在不发生灾难性遗忘的情况下实现深度强化学习。《神经计算》，428:291–307，2021年。[50] Hsuan-Wei Lee 和 Vincent Cheng-Sheng Li。耦合流行病—行为系统中的相边界与临界转变。《数学生物学通报》，87:179，2025年。[51]Jakob Foerster、Nantas Nardelli、Gregory Farquhar、Triantafyllos Afouras、Philip HS Torr、Pushmeet Kohli 和 Shimon Whiteson。稳定化深度多智能体强化学习中的经验回放。载于国际机器学习会议，第 1146–1155 页。PMLR，2017年。

附录 A. 补充材料

附录 *A.1.* 对种群规模的敏感性

图 A.1 通过比较三种网格尺寸 30×30、 40×40 和 50×50 下的平均合作水平随探索强度 B 的变化，考察了共享策略深度Q网络对种群规模的敏感性。所有结果均在固定的评估阶段进行评估，此时学习已禁用；不同系统尺寸使用相同的收益参数（$D_r = 0.25$）和学习超参数。

在三种种群规模下，我们观察到一致的定性模式：在较低的探索水平下，合作保持相对较高，并随着 B 的增加而单调下降。这一下降的位置和形态在不同网格尺寸下都非常稳定，仅在较小的 B 下存在轻微的定量差异。具体而言，较大的种群在低探索机制下表现出略低的合作水平，这反映了局部交互情境中更高的异质性，但总体的崩溃轨迹并未改变。

重要的是，随着探索增加，合作崩溃的开始及其发展过程并不敏感于系统规模。在中等和较高 B 下，曲线几乎重合，这表明主要结果中观察到的探索诱导退化并非有限尺寸伪影，也不是由小种群效应驱动的。相反，它反映了非平稳多智能体学习中共享策略深度Q网络的一项内在特性。

这些结果证实，正文中报告的结构性合作崩溃在不同种群规模下依然存在，这支持了所观察到的失效模式具有普遍性和可扩展性。因此，图 A.1 是一项稳健性检验，表明根据默认 30×30 设置得出的结论同样适用于规模大得多的系统。

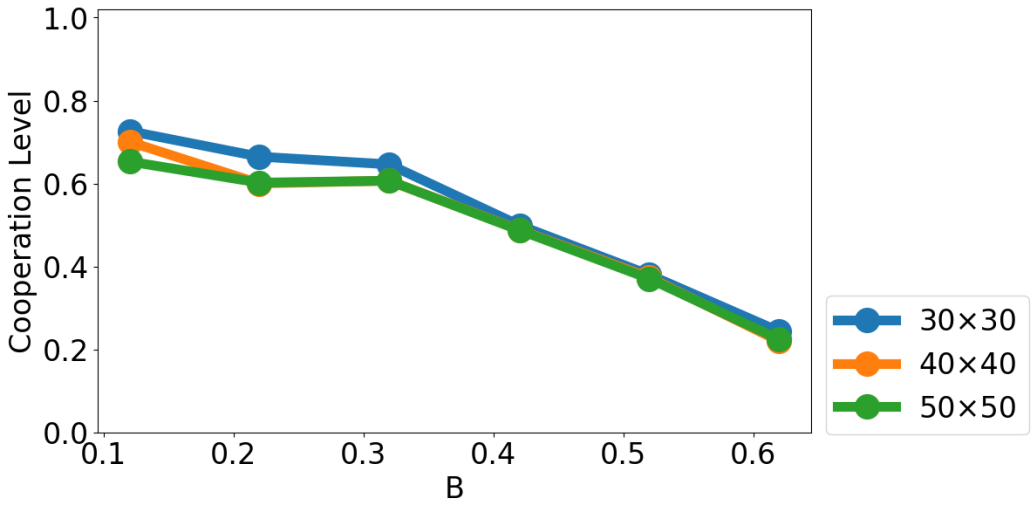

图 A.1：共享策略深度Q网络下合作结果对种群规模的敏感性。图中展示了评估阶段的平均合作水平随探索强度 B 变化的情况，涵盖三种网格尺寸： 30×30、 40×40 和 50×50。所有结果均在训练收敛后、禁用学习的条件下计算，使用固定的收益参数 $D_r = 0.25$，且不同系统规模采用相同的学习超参数和探索计划。在所有种群规模下，合作水平都随B增加而单调下降，且各曲线呈现出高度一致的崩溃轨迹。只有在低探索水平下才出现轻微的定量差异，而在中、高 B下，合作崩溃的起始时间和发展过程均保持一致。这表明，共享策略深度Q网络中观察到的探索诱发型合作崩溃对种群规模具有稳健性，并非由有限规模效应导致。

附录 *A.2.* 对学习超参数的敏感性

图 A.2 展示了针对共享策略深度Q网络关键学习超参数的系统性敏感性分析，包括 (a) 隐藏层大小、(b) 经验回放缓冲区容量、(c) 折扣因子 γ、(d) 损失函数和 (e) 优化器。对于所有图板，平均合作水平均在固定的评估阶段进行评估，此阶段禁用学习，且收益参数 $D_r = 0.25$保持不变。除非另有说明，学习算法、交互拓扑和仿真方案的所有其他组成部分均与正文实验所用设置相同。

在所有超参数变化中，都呈现出一致的定性模式：随着有效探索强度 B 增加，合作水平单调下降。尽管不同的架构和优化选择会导致低探索强度下的合作水平出现明显的定量差异，

但相对于 B 的排序和崩塌轨迹保持不变。对 30 个随机种子取平均后，没有任何超参数配置会改变这种对 B 的单调依赖，也无法使合作恢复到弱探索条件下观察到的高合作水平。

面板 (a) 显示，增大隐藏层大小主要影响低 B下的绝对合作水平：在低探索机制下，较大的网络表现出更高的合作水平。然而，随着探索增加，所有网络规模的合作水平均趋于相近的低水平，这表明表征能力的提升无法阻止探索导致的崩溃。面板 (b) 表明，经验回放缓冲区容量会改变合作水平的整体尺度，但不会从定性上改变其对B的依赖关系：较大的缓冲区在低探索水平下始终带来更高的合作水平，但随着 B 增加，合作水平仍会单调下降。

面板 (c) 表明，改变折扣因子 γ 会使弱探索条件下的整体合作水平发生变化。然而，不同折扣因子下，对探索的定性敏感性保持一致；当 B足够大时，所有设置下的合作水平均会降至较低水平。这说明，时间信用分配上的差异无法缓解强探索所导致的崩溃。

面板 (d) 和 (e) 进一步证实，损失函数或优化器的选择都不会从根本上改变观察到的动态。在低探索水平下，SmoothL1 损失和 AdamW 始终比均方误差和 Adam 带来更高的合作水平，而 RMSprop 在所有 B下产生的合作水平均较低。重要的是，随着探索增加，所有优化器下的合作水平都会单调下降，进一步说明观察到的崩溃并非由某种特定的优化方法导致。

综上，图 A.2 表明，在广泛采用的学习超参数范围内，共享策略深度Q网络中由探索引发的合作崩溃具有稳健性。这些结果证实，正文中所识别的失效模式并非架构容量、回放记忆设计、折扣因子或优化选择所造成的偶然现象，而是反映了在高度非平稳的多智能体交互下，共享表征学习固有的局限性。

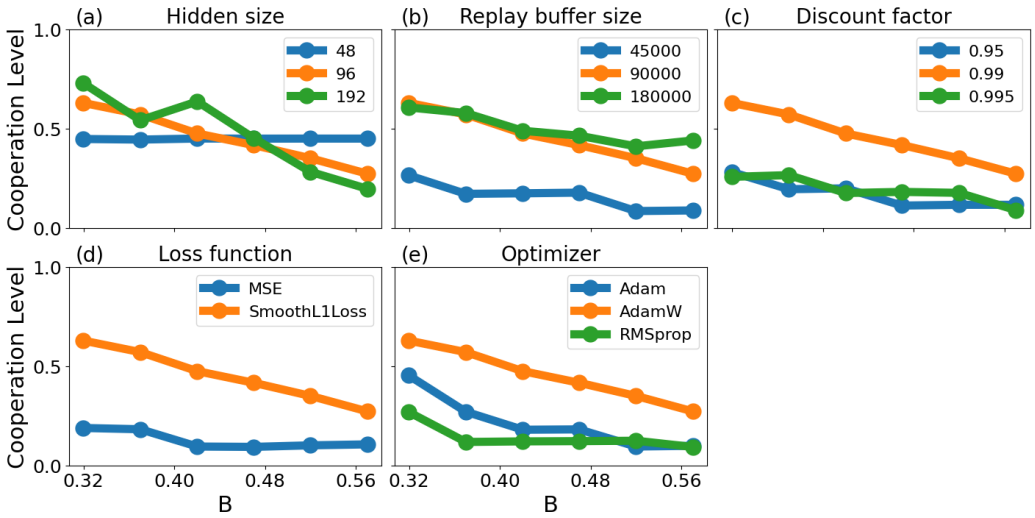

图 A.2：固定收益条件下，共享策略深度Q网络的合作结果对学习超参数的敏感性。图中展示了评估阶段的平均合作水平随探索强度 B 变化的情况，分别考察了 (a) 隐藏层大小、(b) 经验回放缓冲区容量、(c) 折扣因子 γ、(d) 损失函数和 (e) 优化器的影响。除非另有说明，所有结果均在训练收敛后、禁用学习的条件下计算，收益参数固定为 $D_r = 0.25$，且交互拓扑、探索计划和仿真方案均保持一致。在所有超参数设置下，合作水平都随着 B 的增加而单调下降。尽管不同的架构和优化选择会改变低探索强度下的绝对合作水平，但没有任何配置会改变其对 B 的定性依赖关系，也无法使合作恢复到弱探索下观察到的高合作状态。

附录 *A.3.* 贪婪评估下的合作

图 A.3 报告了在固定诱惑参数 $D_r = 0.25$ 下，平均合作水平随探索强度 B 变化的情况。对于 B 的每个取值，所报告的合作水平均取 30 个随机种子的平均值。需要特别指出的是，所有测量均在评估阶段进行，此时通过设置 $\tau_{\text{eval}} = 1$ 禁用探索，使智能体相对于其学到的Q函数采取贪婪动作。

如图所示，尽管评估期间没有探索噪声，合作仍会随着 B 的增加而单调下降。在探索强度较低时，学习得到的策略能够维持相对较高的合作水平。然而，随着 B 的增加，合作逐渐恶化，最终降至较低水平，这表明强探索的不利影响是通过学习动态而非

测试时的随机动作选择产生的。

这一结果表明，共享策略 DQN 中观察到的合作崩溃并非评估期间探索行为造成的假象。相反，训练期间的高强度探索从根本上改变了智能体学到的价值表征，导致其策略即使在完全满足贪心条件的情况下进行评估，也倾向于背叛。这一趋势在 30 个随机种子下均表现稳健，进一步表明，该现象反映的是强探索条件下学习架构的系统性局限，而非暂时波动或收敛不足。

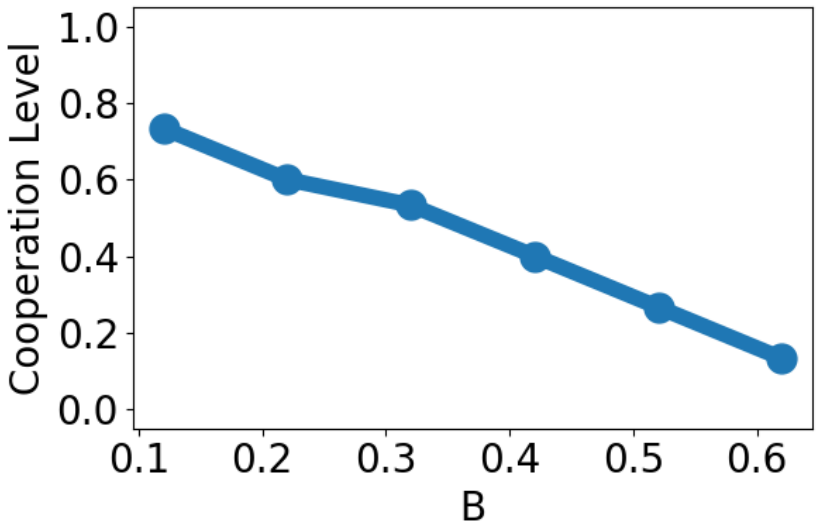

图 A.3：在诱惑参数 $D_r = 0.25$ 固定时，平均合作水平随探索强度 B 的变化。对于 B 的每个取值，结果均为 30 个随机种子的平均值。所有测量均在评估阶段进行，并通过设置 $\tau_{\text{eval}} = 1$ 关闭探索，使智能体根据学到的 Q函数采取贪心行动。