参赛学生姓名：	辛朋锦
中学：	中国人民大学附属中学
省份：	北京市
国家/地区：	中国/北方赛区
指导老师姓名：	赵彤
指导老师单位：	中国科学院大学数学科学学院
论文题目：	从潜在角色到行为分工：RoboCup 2D 足球智能体中的动态角色学习与验证

2026 S. T. Yau High School Science Award
仅用于2026丘成桐中学科学奖论文公示

从潜在角色到行为分工：RoboCup 2D 足球智能体中的动态角色学习与验证

辛朋锦，赵彤

摘要

在机器人足球比赛中，共享策略有助于提高同构智能体的样本效率，但当多个智能体面对相似的局部目标时，也可能产生相似的行为响应，从而导致局部行为重叠与空间聚集。针对 RoboCup 2D 进攻任务中智能体角色随球权与空间关系动态变化的特点，本文在多智能体近端策略优化（MAPPO）框架上引入动态软角色模块，并将侧抑制的竞争机制引入归一化角色表示，以促进足球智能体的局部动态角色分化。首先，利用门控循环网络（GRU）从智能体的局部观测及历史状态中提取连续角色表示；其次，根据同一时刻局部队友之间的表示相似性，在单位球面的切空间中实施侧抑制（Tangent-space Lateral Inhibition, TLI）, 并通过归一化回缩对角色表示的方向进行自适应调整。为验证所提出机制的有效性，本文首先在 2v1 场景中开展受控实验，随后在 3v3 场景下比较加入 TLI、关闭 TLI 以及采用余弦平方去相关正则三种配置，并从真实评估回合中的角色占用、角色切换、动作构成和空间分布四个方面分析角色分化特征。在 2v1 场景中，加入 TLI 后三次独立训练的平均进球率由 88.7% 提升至 93.5%；在 3v3 场景中，加入 TLI 的配置在进球率和角色—动作互信息方面均取得更高的表现，而更强的几何去相关并未相应带来更优的任务表现。实验中，进一步将完整配置持续训练至约 200 万环境步，在最终 200 局评估中获得 42.0% 的进球率，并观测到任务性能保持持续性稳定。上述结果表明，相较于不显式建模角色竞争的 MAPPO 框架，本文方法在共享 Actor 内提供了动态角色表示、同周期表示竞争与角色条件策略调制的显式通路；其队友均值聚合形式不依赖固定的三智能体规模，具备向更大队伍扩展的结构条件。

关键词：多智能体强化学习；动态角色学习；侧抑制；共享策略；仿真足球

2026 S. T. Yau High School Science Award
仅用于2026年成绩评审学术交流

Abstract

In robot soccer, shared policies can improve the sample efficiency of homogeneous agents. However, when multiple agents face similar local objectives, they may also produce similar behavioral responses, resulting in local behavioral overlap and spatial aggregation. To address the dynamic variation of agent roles with ball possession and spatial relationships in RoboCup 2D offensive tasks, this paper introduces a dynamic soft-role module into the Multi-Agent Proximal Policy Optimization (MAPPO) framework and incorporates a competitive lateral-inhibition mechanism into normalized role representations to promote local dynamic role differentiation among soccer agents. First, a gated recurrent unit (GRU) is employed to extract continuous role representations from each agent's local observations and historical states. Second, based on the representational similarity among teammates at the same time step, Tangent-space Lateral Inhibition (TLI) is applied in the tangent space of the unit hypersphere, and the directions of role representations are adaptively adjusted through normalization-based retraction.

To evaluate the effectiveness of the proposed mechanism, controlled experiments are first conducted in a 2v1 scenario. Subsequently, three configurations are compared in a 3v3 scenario: TLI enabled, TLI disabled, and cosine-squared decorrelation regularization. Role differentiation is further analyzed from four aspects in actual evaluation episodes: role occupancy, role switching, action composition, and spatial distribution. In the 2v1 scenario, introducing TLI increases the average goal-scoring rate over three independent training runs from 88.7% to 93.5%. In the 3v3 scenario, the configuration with TLI achieves higher performance in both goal-scoring rate and role-action mutual information, whereas stronger geometric decorrelation does not correspondingly lead to better task performance. The complete configuration is further trained for approximately two million environment steps, achieving a goal-scoring rate of 42.0% over the final 200 evaluation episodes, while maintaining sustained stability in task performance.

These results indicate that, compared with the MAPPO framework without explicit role competition, the proposed method provides explicit pathways within the shared Actor for dynamic role representation, same-step representation competition, and role-conditioned policy modulation. In addition, its teammate mean-aggregation mechanism does not depend on a fixed three-agent team size, providing a structural basis for extension to larger teams.

Keywords: multi-agent reinforcement learning; dynamic role learning; lateral inhibition; shared policy; simulated soccer

2026 S. T. Yau High School Science Award
仅用于2026年成都市学科竞赛论文公示

目录

1. 引言 1
2. 相关工作 3
2.1 多智能体动态角色学习 3
2.2 表征竞争与角色分工 3
2.3 从潜在角色到行为分工 4
3. 问题定义与预备知识 4
3.1 合作型部分可观测任务建模 4
3.2 MAPPO 与集中训练—分散执行 5
3.3 动态角色变量与研究目标 5
4. 方法：三层递进的动态角色学习框架 5
4.1 情境感知软角色建模 6
4.2 球面切空间竞争性侧抑制 6
4.3 角色条件残差策略调制 7
5. 实验设置 8
5.1 任务、环境与证据组织 8
5.2 比较模型与评价指标 8
6. 实验结果与分析 8
6.1 从受控协作任务验证角色竞争 8
6.2 3v3 中的侧抑制机制对照 9
6.3 3v3 最终任务表现与持续训练稳定性 10
6.4 角色随局势变化，而不是永久绑定球员 11
7. 讨论与结论 11
参考文献 13
附录：理论推导、实现细节与评价定义 17
A. 符号与切空间侧抑制的几何说明 17
B. 网络结构与训练实现 18
C. 评价指标与统计口径 20

2026 STYSW High School Science Award
仅用于2026年成都市中学生科学论文公示

1. 引言

在机器人足球比赛中，单个动作是否合理只是协作的一部分，当持球者需要推进或寻找射门机会时，无球者则需要调整位置；球权或者空间位置发生变化后，原有分工也应随之重组。若把职责固定到具体某个球员编号上，可能限制策略对局势变化的适应能力。因此，本文将角色视为由局部观测和历史状态共同决定的动态功能倾向，而不是预先指定的位置标签。

参数共享能够让同构智能体复用感知和决策知识，是 MAPPO 等合作学习方法的重要基础 [1]，但也可能使智能体在相似目标下形成相近的内部响应，表现为重复追逐同一机会或长期占据相近区域，从而增加行为冗余的风险。即使网络额外输出角色变量，如果缺少促使角色分化的机制，这些变量仍可能收敛到相似模式。

已有研究从多种角度建模智能体角色：ROMA 通过连续角色嵌入和角色条件网络学习涌现角色 [5]，RODE 依据动作效果构造角色动作子空间 [6]，CDS 与 R3DM 则分别从行为多样性和未来轨迹建模协作差异 [7,8]。这些方法说明，角色既可以作为内部表示，也可以进一步影响策略结构。面向机器人足球，本文保留环境给出的完整合法动作集合，不为角色预先划定动作子集，而是在共享策略内部形成随状态变化的软角色，并利用竞争机制减少角色响应冗余。

从角色学习的优化过程看，至少需要区分两类性质不同的失效现象。第一类是表征坍缩（Representation Collapse）：多个智能体的角色表示在训练过程中逐渐趋同，使原本用于刻画功能差异的潜在空间退化为高度相似甚至近似单一的表示。此时，即使网络结构中保留了多个角色通道，也难以形成稳定可辨的角色坐标。针对这一问题，已有角色学习与表示学习研究通常采用正交约束、多样性目标或其他去相关正则，试图维持潜在表示之间的可分性。

第二类问题更为隐蔽，本文将其称为装饰性角色（Decorative Roles）：角色表示在几何空间中已经能够区分，但这些差异没有充分传递到动作策略，不同角色条件下的决策响应仍然接近。此时，仅观察表示相似度、聚类结构或有效角色数，模型可能表现出 “角色已经形成” 的外观，却难以在动作构成和实际交互中观察到相应的功能差异。因此，角色是否具有意义不能只由潜在空间的几何结构判断，还需要检查角色差异是否真正作用于策略输出和行为分布。

这两类失效发生在不同层次：表征坍缩对应潜在角色空间内部的退化，而装饰性角色对应从角色表示到动作决策未能充分传递。本文关注角色表示的几何差异是否能够体现为可观察的行为差异：通过同周期角色之间的竞争调整相关表示，并结合角色条件策略调制以及动作、空间和任务指标，考察角色差异与实际行为之间的关系。实验中进一步观察到，更强的几何去相关并不必然带来更高的行为差异或任务表现，这也说明 “表示分开” 与 “形成有效分工” 需要分别评价。

这种竞争机制受到侧抑制启发，当多个单元对相近输入产生重叠响应时，局部相互作用能够削弱冗余并突出差异 [9,10]。将这一思想用于角色学习，并不是预先指定某个机器人承担固定职位，而是在同一局势下对相似角色表示施加竞争，使共享策略更容易探索不同的功能方向。最终形成何种足球行为，仍由任务奖励和环境交互共同决定。

2026 Tyust High School Science Award
仅用于2026年全国青少年科技创新大赛论文公示

由于角色向量被归一化到单位球面，其方向比长度更能表征角色差异。为此，本文将相似性相关的更新投影到球面切空间，再通过归一化回到单位球面，得到切空间侧抑制（Tangent-space Lateral Inhibition，TLI）。与通过额外损失项直接约束角色表示之间的几何关系不同，TLI 将同周期智能体之间的前向竞争嵌入角色生成过程，并把具体行为分工交由共享策略学习。

本文以切空间侧抑制为核心方法设计，结合动态软角色表示与角色条件策略调制，构建以下三个相互衔接的模块：

（1）情景感知动态软角色表示。在共享循环策略的隐藏状态上构造连续角色坐标与软角色概率，使角色由局部观测历史和当前局势共同决定，而不是与智能体身份或预设足球职位固定绑定。该机制为同一智能体随球权、位置和交互历史变化而调整功能倾向提供了统一表示。

（2）球面切空间竞争性侧抑制，缓解软角色表示坍缩。针对参数共享条件下同周期角色表示容易出现冗余响应的问题，在单位球面角色空间中引入基于余弦相似性的切向竞争，并通过归一化回缩保持角色表示的球面约束。该机制不直接指定角色语义或战术位置，而是通过调整相关角色表示，为不同智能体探索差异化的功能倾向提供竞争性归纳偏置。

（3）提出角色行为特化机制，将角色差异直接映射到动作空间。将软角色概率与可学习角色原型进行连续混合，并以残差形式注入共享策略，使潜在角色差异能够直接调节合法动作之间的偏好，同时保持环境动作掩码与共享策略主体不变，从而实现从角色表示到行为分工的连续传递。

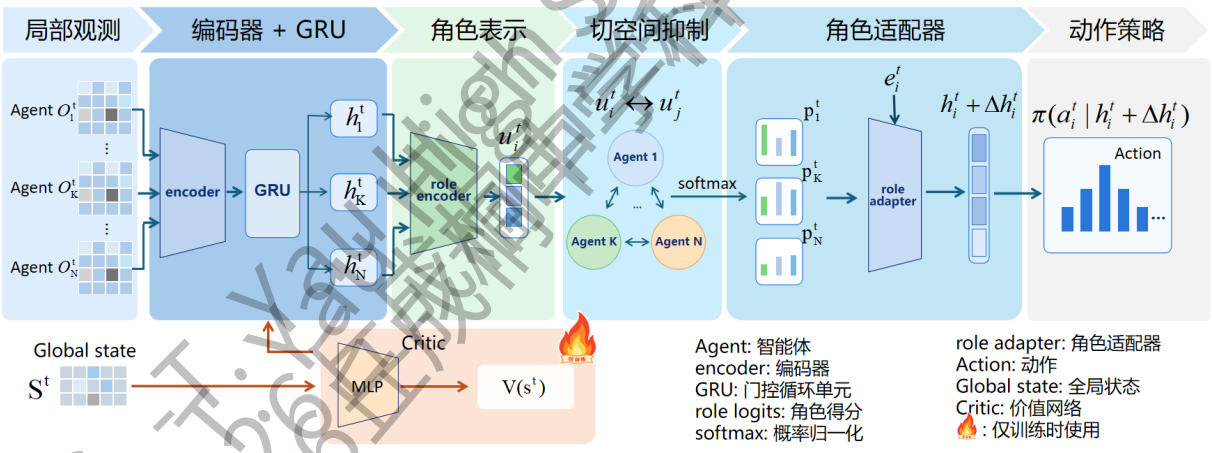


图 1　动态软角色与切空间侧抑制的总体框架。

实验结果为上述方法设计提供了分层证据。2v1 受控实验中，加入 TLI 后三训练种子平均进球率由 88.7% 提升至 93.5%；3v3 匹配比较中，保留 TLI 的平均进球率和角色–动作互信息分别为 36.5%±4.3% 和 0.387，高于关闭 TLI 的 33.7%±6.0% 和 0.291，也高于 cos2 正则替代条件的 32.0%±2.3% 和 0.289。与此同时，cos2 条件获得更低的平均绝对余弦，却未得到更好的任务和行为指标，说明角色几何分离强度不能单独代表有效分工。进一步持续训练至约 200 万环境步后，最终 200 局进球率为 42.0%，且中后期多个检查点未出现连续下降。2v1 提供较简单协作场景下的受控机制证据，3v3 则进一步检验更复杂交互条件下任务、表示与行为指标是否保持一致。

2026 S.-T. Yau High School Science Award
2026 年 S.-T. 丘成桐中学科学奖
仅用于 2026 年丘成桐中学科学奖公示

2. 相关工作

2.1 多智能体动态角色学习

在合作型多智能体强化学习中，参数共享能够减少模型规模并提高经验复用效率，但也使同构智能体更容易学习相近的决策模式。为在共享策略中形成具有差异的协作行为，已有研究逐渐引入潜在角色作为中间变量。ROMA 通过学习潜在角色嵌入并利用角色条件网络生成策略，使智能体能够依据局部轨迹形成动态角色；RODE 则进一步从动作效果出发，将动作空间组织为不同的角色动作子空间，从而降低复杂协作任务中的决策搜索难度 [5,6]。这类方法表明，角色无需与智能体身份固定绑定，而可以作为随状态变化的潜在条件参与多智能体决策。

除显式角色建模外，部分工作从行为多样性的角度缓解共享策略下的同质化。CDS 利用智能体身份与交互轨迹之间的关系维持个体行为差异，在共享策略的同时保留个体特征；R3DM 则通过动力学模型将角色表示与未来状态变化联系起来，使角色发现不仅依赖当前观测，还受到预期行为后果的约束 [7,8]。这些研究从不同角度说明，潜在角色的价值不只是为智能体附加一个离散或连续标签，更重要的是使不同智能体在团队任务中形成具有功能差异的决策模式。

足球智能体的角色分工具有明显的状态与局势依赖性，射门、盘带、移动和传球等动作的可用性会随球权、位置和比赛状态不断变化，如果角色与固定球员身份绑定，或角色动作子空间限制了当前局势所需的动作选择，可能限制智能体对局势变化的响应。基于这一特点，本文保留基线环境定义的完整合法动作集合，不预先规定角色的战术语义或动作子空间，而将角色作为连续条件信息注入共享策略，以调节合法动作之间的相对偏好，使角色能够随局部观测及历史状态动态调整。

2.2 表征竞争与角色分工

动态角色模型提供了表达差异的空间，但角色变量的存在并不意味着多个智能体会自然使用不同的表示方向。特别是在共享参数和共同团队奖励下，不同智能体面临相似的局部目标时，可能逐渐产生相似的角色表示，使潜在角色空间退化为高度重叠的响应，即本文所关注的表征坍缩问题。因此，如何在保留共享策略的同时，减少同一局势下角色表示之间的冗余，是动态角色学习需要解决的一个关键问题。

已有表征学习研究通常通过正交约束、多样性正则或去相关目标减少潜在特征冗余。例如，Barlow Twins 通过约束跨视图特征的相关矩阵减少重复信息 [12]；在强化学习和技能发现中，多样性目标也常用于推动策略或潜变量覆盖不同区域。此类方法通过附加损失约束表示的相关结构。对于多智能体角色而言，降低表示相关性并不必然带来有效的行为分工，因此还需要结合动作行为和团队任务表现进行评价。另一类与本文思路相关的研究来自竞争学习与侧抑制。经典侧抑制描述相邻感受单元之间通过抑制性相互作用突出响应差异，竞争学习则进一步说明局部响应竞争能够促使不同单元形成不同的输入选择性 [9,10]。这里具有启发意义的并不是具体生物神经回路，而是“重叠响应之间引入竞争”这一计算原则。对于机器人足球而言，这一原则可以自然对应于同一局部情境下多个智能体之间的角色竞争：根据同周期智能体角色表示之间的相关性，竞争机制对表示方向施加调整，从而为不同功能倾向的形成提供额外的归纳偏置。本文进一步将角色表示归一化到单位球面，使角色差异主要由向量方向描述。在这一几何约束下，欧氏空

2026 RoboCup Junior Soccer Open Simulation League Science Award
仅用于2026年度青少组仿真组科学奖论文公示

间中的任意扰动还可能同时改变方向和模长；球面优化中的切空间投影与回缩则为保持单位范数约束下的方向更新提供了自然工具 [11]。因此，相比只在训练目标中加入全局去相关约束，本文关注的是如何在每次前向决策过程中，对同周期智能体的重叠角色表示施加局部竞争。

2.3 从潜在角色到行为分工

角色表示能够在潜在空间中形成差异，并不意味着这种差异一定会进入实际决策。即使两个角色向量在几何空间中具有较大距离，下游共享策略仍可能对它们产生相近的动作响应，从而形成本文所称的“装饰性角色”：角色表示在潜在空间中可以被区分，但这种差异对策略输出和实际行为影响有限。与表征坍缩不同，这类失效并不发生在角色空间内部，而发生在角色信息向动作策略传递的过程中。

行为多样性与技能发现研究为区分这两个层次提供了参考。DIAYN 通过最大化技能变量与访问状态之间的信息关联学习可区分技能，DvD 则从策略种群角度研究如何获得具有差异的行为模式 [13,14]。这些研究说明，潜变量是否有意义不能只由其本身的分布或几何结构决定，还需要考察不同潜变量是否对应可观察的行为变化。对于足球智能体动态角色而言，这种行为差异既可能表现为不同的动作构成，也可能表现为场上位置、角色持续时间以及随球权变化产生的角色切换。

基于这一认识，本文将角色分工分为多个层次进行观察。首先，通过潜在空间指标判断智能体的角色表示是否出现趋同，从而检验表征竞争是否真正作用于角色空间；其次，通过角色占用和角色转移统计描述不同角色状态在比赛过程中的使用方式，观察角色是否能够随局势变化进行动态调整；最后，通过角色条件下的动作构成和空间分布分析，检验潜在角色差异是否进一步映射为真实比赛回合中的行为差异。这种分层评价避免将“表示距离变大”直接解释为“协作能力增强”，也使角色学习的评价从单纯的潜在表征分离进一步延伸到实际行为分工。

3. 问题定义与预备知识

3.1 合作型部分可观测任务建模

RoboCup 2D 进攻任务可形式化为合作型部分可观测随机博弈。设受控智能体集合为 $I = \{1, \dots, N\}$，全局状态空间为 S，智能体 i 的局部观测空间与动作空间分别记为 O_i 和 A_i。时刻 t，环境处于全局状态 s^t，智能体 i 仅获得局部观测 o_i^t，并依据环境给出的合法动作掩码 m_i^t 从当前可用动作集合 $A_i^t = \{a \in A_i \mid m_i^t(a) = 1\}$ 中选择动作 a_i^t。所有受控智能体共享团队奖励 r^t，联合动作决定状态转移。给定折扣因子 γ，策略学习的目标是最大化团队期望累计回报 $J(\theta) = \mathbb{E}[\sum_t \gamma^t r^t]$。

由于足球环境具有显著的部分可观测性，单步观测不足以完整描述球权变化、队友运动以及此前决策对当前局势的影响。因而，智能体的决策信息更适合表示为局部交互历史 $\tau_i^t = (o_i^0, a_i^0, \dots, o_i^t)$，并由循环隐藏状态 h_i^t 对其进行压缩。与此同时，合法动作集合随比赛状态动态变化。本文将动作合法性视为环境约束：角色模块只能重新分配当前合法动作之间的偏好，而不能为角色人为增加或删除动作。这一设定使角色模块仅调节合法动作之间的偏好，而不通过预设动作子空间规定角色分工。

2026 RoboCup Tianjin Jinnan Middle School Science Award
2026年RoboCup世界杯天津市津南中学科学奖论文奖

3.2 MAPPO 与集中训练—分散执行

本文采用 MAPPO 作为多智能体策略优化框架 [1–3]。在集中训练、分散决策的基本框架下，各智能体共享 Actor 参数 θ，但分别维护自己的循环状态；训练阶段由集中式 Critic $V_\phi(s^t)$ 读取全局状态估计团队价值，策略执行时则依据个体可获得的信息生成动作。PPO 通过重要性比率 $\rho_t(\theta)$ 对新旧策略变化进行裁剪，采用 $L_{\text{clip}} = \mathbb{E}[\min(\rho_t A_t, \text{clip}(\rho_t, 1-\varepsilon, 1+\varepsilon)A_t)]$ 限制单次更新幅度；GAE 以时序差分残差 $\delta_t = r_t + \gamma V(s^{t+1}) - V(s^t)$ 构造优势估计 $\hat{A}_t = \sum_l (\gamma\lambda)^l \delta_{t+l}$，从而在方差与偏差之间取得折中。

上述机制解决的是共享合作策略的稳定优化问题，但并不会自动赋予同构智能体不同的功能倾向。换言之，MAPPO 可以学习 “团队应该如何行动”，却没有显式机制回答 “在当前局势下不同智能体应以何种角色参与行动”。因此，本文保持 PPO/GAE 的基本形式不变，把研究变量集中在共享 Actor 内部的角色生成、角色间竞争以及角色对动作偏好的调制上。这样既保留统一的优化基础，也使有无角色竞争的实验差异能够被单独观察。

3.3 动态角色变量与研究目标

对每个智能体 i，在时刻 t 定义连续角色表示 $u_i^t \in S^{d-1}$ 以及 K 维软角色概率 $p_i^t \in \Delta^{K-1}$。二者由局部交互历史所形成的循环特征决定，而不是由球员编号直接指定。因而，角色在本文中表示状态相关的功能倾向：同一智能体可以随球权、位置和历史信息改变角色概率，不同智能体也可以在同一时刻具有相近的软角色分布。角色分化并不要求 K 个角色与 N 个智能体一一对应，也不以三个角色等频出现作为成立条件。

动态角色学习需要同时关注两类不同层面的失效。其一是表征坍缩：不同智能体的角色表示长期趋于一致，使潜在空间难以承载可区分的功能信息；其二是装饰性角色：角色表示在几何上已经存在差异，但这种差异没有形成稳定、可观察的动作或空间行为。基于这一区分，本文从任务、表示和行为三个层次评价角色学习：任务层考察进球率等协作结果，表示层考察角色占用与向量相关结构，行为层考察角色切换、动作构成和空间分布。由此，本文要解决的问题可以概括为：在不预设足球职位、不缩减环境合法动作集合的条件下，如何在共享策略中形成随局部情境变化的动态软角色，并通过同周期足球智能体之间的竞争减少冗余表示，使角色差异能够进一步体现为可观察的行为分工。

4. 方法：三层递进的动态角色学习框架

针对上文提出的 “角色如何形成、如何避免表示趋同、以及如何进一步作用于行为” 三个问题，本文在共享 MAPPO Actor 上构建三层递进的动态角色学习框架。三层结构分别负责角色表示、角色竞争与策略调制，在前向计算中依次连接，将个体局部交互历史与同周期队友角色表示共同用于动作生成。

（1）情境感知软角色建模。由共享编码器与 GRU 汇总局部观测历史，在循环特征上构造连续角色坐标，使角色随球权、空间位置和交互历史变化，而不是与智能体编号或固定足球职位绑定。

（2）球面切空间竞争性侧抑制。在归一化角色空间中，根据同周期队友角色表示的相似性施加切向竞争，对同周期智能体之间相关的角色表示进行调整，并通过回缩保持单位球面约束。

（3）角色条件残差策略调制。将竞争后的软角色概率映射为角色嵌入，并以残差方式注入共享策略，为

2026 S.T. Yau High School Science Award
仅用于2026年度学科竞赛论文公示

角色信息调节合法动作之间的偏好提供显式通路，而不改变环境动作掩码。

完整 Actor 前向流程可概括为：

$$o_i^t \to \text{Shared Encoder + GRU} \to h_i^t \to \text{Role Encoder} \to u_i^t \to \text{TLI} \to \hat{u}_i^t$$
$$\to W_r + \text{Softmax} \to p_i^t \to \text{Prototype Mixture} \to e_i^t$$
$$\to \text{Role Adapter} \to \tilde{h}_i^t \to \text{Action Head} \to \pi_i^t.$$

上述三层在功能上具有明确分工：第一层回答 “角色从何而来”，第二层回答 “多个角色表示如何避免长期趋同”，第三层回答 “角色差异如何进入动作策略”。这种递进关系并不预设三种角色必须等频出现，也不要求角色与智能体一一对应；角色的具体语义仍由任务奖励和环境交互共同塑造。下文依次给出三层结构的具体计算。

4.1 情境感知软角色建模

在每个决策时刻，智能体 i 的局部观测 o_i^t 首先经过共享编码器，并与该智能体独立维护的循环状态共同输入 GRU，得到包含当前情境与历史信息的隐藏表示 h_i^t：

$$h_i^t = \text{GRU}_\theta(f_\theta(o_i^t), h_i^{t-1}) \tag{1}$$

随后，角色编码器从共享循环特征中提取未归一化角色向量 z_i^t，并通过带数值保护的归一化得到单位球面上的连续角色坐标 u_i^t：

$$z_i = f_\omega(h_i), \qquad u_i = N_\epsilon(z_i) \tag{2}$$

其中，归一化操作仅约束角色表示的模长，角色信息主要由方向承载。u_i^t 不是预定义的足球职位，而是由当前局部信息与历史共同决定的功能坐标。该设计先为共享策略提供可变化的角色表示空间，至于不同智能体是否会形成不同方向，则交由下一层的同周期竞争机制处理。

角色编码阶段本身仍保持参数共享：不同智能体使用同一组编码参数，但由于观测历史与循环状态不同，可以产生不同的角色坐标。TLI 所需的队友角色表示在同一环境、同一决策周期内汇总；该表示交换与训练阶段 Critic 使用全局状态属于两条不同的信息路径。

4.2 球面切空间竞争性侧抑制

仅建立连续角色空间并不能保证角色自然分化。由于 Actor 参数共享且团队奖励一致，不同智能体的角色向量仍可能在训练中长期靠近。为此，第二层在同一决策周期的角色表示之间引入相似性相关的竞争。对单位角色向量 u_i 和 u_j，首先计算两者的余弦关系，并构造位于 u_i 切空间内的队友方向分量：

$$c_{ij} = u_i^\top u_j, \qquad v_{ij} = u_j - c_{ij} u_i \tag{3}$$

2026 CAAI Yau Hsiu High School Science Award
仅用于2026年度高中生科学论文公示

其中，$u_j - c_{ij}u_i$ 消除了队友表示在 u_i 径向上的分量，因此只保留沿单位球面的切向变化；再以相似度 c_{ij} 对该方向加权，并对其余 $N-1$ 个队友取平均，得到智能体 i 的竞争方向 g_i：

$$g_i = \frac{1}{N-1}\sum_{j\neq i} c_{ij} v_{ij} \tag{4}$$

$$\hat{u}_i = N_\epsilon(u_i - \alpha g_i) \tag{5}$$

本文采用竞争系数 $\alpha = 0.1$，并使用队友均值聚合。更新只作用于角色表示方向，不直接约束场上位置、球员距离或具体动作。与直接将角色向量推向最大距离不同，TLI 的目标是对当前重叠响应施加适度竞争，再由后续策略学习决定这些方向是否形成有用的行为差异。

由于竞争方向位于 u_i 的切空间，单步更新不会产生一阶径向收缩；归一化随后将更新结果回缩到单位球面。该性质解释了采用切空间形式的几何动机。其局部能量解释、作用强度以及完全同向或反向时的退化情形统一放在附录 A 中讨论。

4.3 角色条件残差策略调制

第三层负责把角色表示转化为策略可利用的条件信号。竞争后的角色向量 $\hat{u}_i^t$ 首先映射到 K 个角色通道，并通过温度 Softmax 得到软角色概率 p_i^t：

$$p_i = \mathrm{softmax}\left(\frac{W_r \hat{u}_i}{\tau}\right) \tag{6}$$

$$e_i = \sum_{k=1}^{K} p_{ik} C_k \tag{7}$$

C_k 为可学习原型，本文取 $K = 3$、$\tau = 1$。实际策略使用整个概率向量，而非仅选最大的角色。一个通道即使很少成为硬标签，也可能通过软混合参与动作生成，因此不要求三种角色等频出现。

由软角色概率加权得到角色嵌入 e_i^t 后，使用轻量残差适配器将其注入共享隐藏表示：

$$\tilde{h}_i = h_i + W_a e_i, \qquad l_i = b_\theta(\tilde{h}_i) \tag{8}$$

适配器采用较小初始化，使新增角色通路在训练初期只对原共享策略产生有限扰动。得到的 $\tilde{h}_i^t$ 再输入共享动作头。角色模块仅调整当前合法动作之间的相对偏好，不改变环境提供的动作集合；因此最终策略在动作掩码约束后的合法动作上完成归一化：

$$\pi_i(a) = \frac{m_i(a)\exp(l_i(a))}{\sum_{b:\, m_i(b)=1} \exp(l_i(b))} \tag{9}$$

2026 S. Tsay High School Science Award
仅用于2026正威桐高学科学奖论文公示

5. 实验设置

5.1 任务、环境与证据组织

实验基于 R2D-RL 对 RoboCup 2D 与 HELIOS 球员客户端的封装[15]，以 3v3 进攻任务为主要场景。三个受控智能体面对固定等级对手，局部输入为 100 维，包含 97 维环境观测与 3 维身份编码；全局价值输入为 136 维。策略选择 19 个高层离散动作，每回合上限为 300 环境步。身份编码用于区分受控智能体，不是预设的角色或职位标签；动态软角色由策略网络生成。

回合会因进球、球权丢失、出界或超时等事件结束。进球率定义为全部评估回合中己方进球回合的比例：

$$G = \frac{1}{n_{\mathrm{ep}}} \sum_{e=1}^{n_{\mathrm{ep}}} \mathbf{1}[\mathrm{goal}_e] \tag{10}$$

5.2 比较模型与评价指标

MAPPO 作为无显式角色的共享策略参照；核心受控比较在关闭 TLI 的 SR-MAPPO 与启用 TLI 的 TLI-SR-MAPPO 之间进行。普通余弦平方正则作为替代机制对照：在匹配训练中关闭 TLI 前向，并将同周期角色余弦平方均值加入损失，用于比较前向竞争与直接去相关约束的差别。

任务评价以进球率为主，并结合回合长度和同一定义下的回报；角色评价包括软 / 硬占用、有效角色数、转移与驻留统计；几何评价同时考察有符号余弦、绝对余弦、余弦平方及 TLI 前后的变化。动作构成和空间分布用于描述角色在真实回合中的行为差异，详细定义见附录 C。

固定检查点上的多回合统计主要反映评估抽样波动；多次训练的均值和标准差则描述共同初始化下的训练变化 [16]。

此外，为观察完整配置在持续训练下的任务稳定性，实验从既有完整配置检查点继续训练至约 2M个累计环境步。在 250k 至 1.75M 的若干检查点各进行 50 局阶段性评估，最终约 2M 检查点使用两批各 100 局的独立评估。

6. 实验结果与分析

6.1 从受控协作任务验证角色竞争

先考察 2v1 中的软角色与 TLI 对照。两者均训练至约 400k 环境步，并在三个训练种子上进行评估。软角色对照的平均进球率为 0.887±0.030，加入 α=0.1 的 TLI 后为 0.935±0.009，均值提高约 4.8 个百分点。该结果说明，在这一受控设置中，前向角色竞争与较好的任务表现可以并存。

表 1　2v1 受控任务的进球率。± 为种子间标准差。

方法	进球率
动态软角色（不含 TLI）	0.887 ± 0.030

2026 S.T. Yau High School Science Award
仅用于2026年成绩单学术论文公示

动态软角色 + TLI（α=0.1）	**0.935 ± 0.009**

2v1 在本文中是机制起点，而不是最终任务的替代。它提供一种工作条件下的正向证据，不能自动推出三个受控智能体、不同训练历史或任意比赛场景都有相同收益。为此，后续分析转向 3v3，并同时观察表示、行为和任务结果。

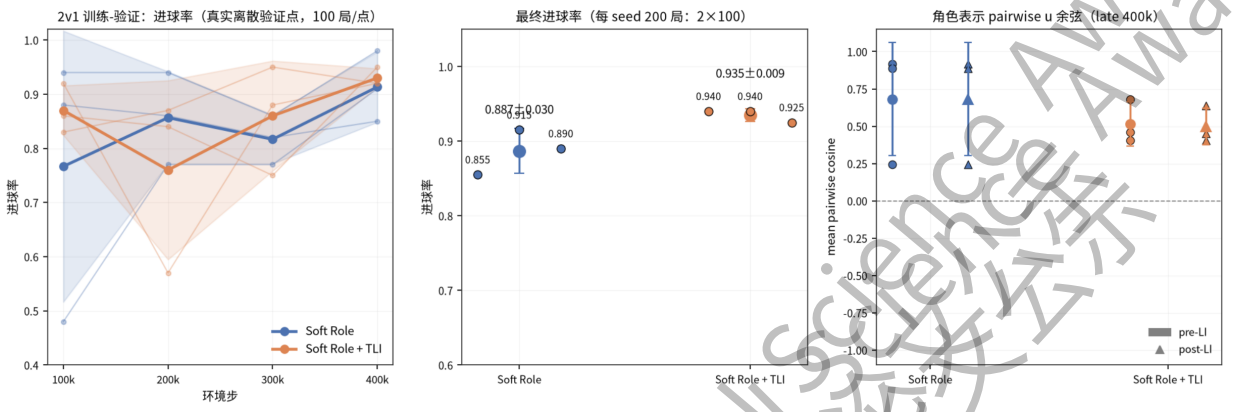


图 2　受控协作任务中的侧抑制验证。浅色线表示单个训练种子，粗线及误差条表示三种子汇总；右图给出 TLI 前后的角色表示相似度。

6.2 3v3 中的侧抑制机制对照

为避免将训练时长、初始化和检查点选择的影响混入侧抑制分析，本节采用共同前驱和相同追加训练预算，对微调阶段保留 TLI、关闭 TLI，以及以 cosine² 正则替代 TLI 的三种配置进行匹配比较。该实验的目的是在尽量控制其他训练因素后，观察 TLI 对任务表现和角色行为结构的影响。

各条件追加约 200k 环境步后，保留 TLI 的配置平均进球率为 0.365±0.043，关闭 TLI 的配置为 0.337±0.060；对应的经验角色—动作互信息分别为 0.387 和 0.291。保留 TLI 的配置在任务和行为指标上具有更高的点估计，但三种子波动与均值差距处于相近量级，因此这里将其视为方向性的受控证据，而不据此宣称 TLI 在 3v3 中具有稳定显著的独立性能优势。

表 2　3v3 匹配微调比较。每条件 3 个微调种子，每个最终检查点 200 局；各条件继承同一前驱。

微调配置	进球率	经验角色—动作 MI
保留 TLI	0.365 ± 0.043	0.387
关闭 TLI	0.337 ± 0.060	0.291
cosine² 正则替代 TLI	0.320 ± 0.023	0.289

值得注意的是，cosine² 正则提供了一个更直接的去相关参照。其平均绝对余弦为 0.545，低于保留 TLI 配置的 0.961，说明它获得了更强的几何分离；但其经验角色—动作互信息和进球率并未同步提高。这表明角色分化不能简单等同于潜在向量之间距离的最大化，表示几何、行为差异与任务完成率需要分别考察。

2026 S.T. Yau High School Science Award
仅用于2026年丘成桐中学科学奖论文公示

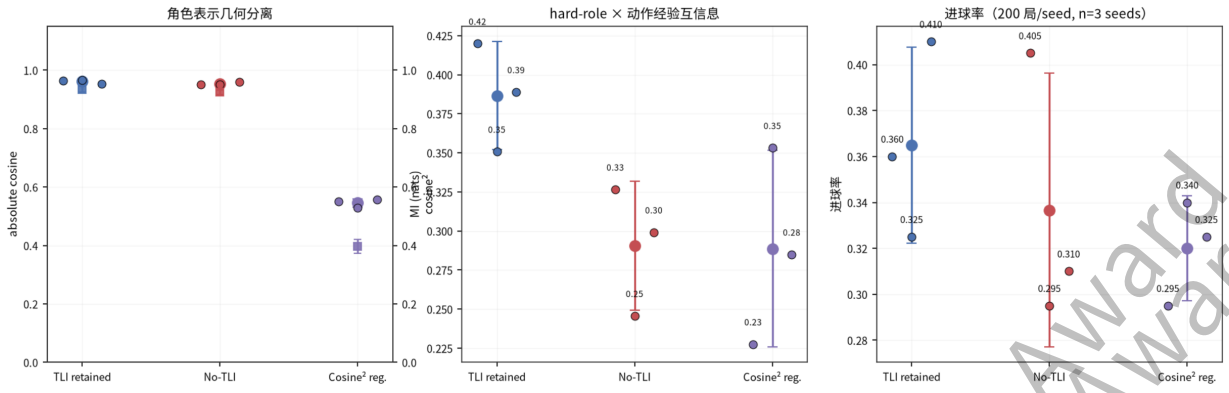


图 3　角色几何变化、实际行为关联与任务表现的对应关系。每个小点表示一个训练种子，误差条为三种子均值与标准差。

6.3 3v3 最终任务表现与持续训练稳定性

在完成侧抑制机制对照后，我们进一步考察当前完整训练配置能够达到的任务水平，以及持续优化是否会带来后期退化。图 4 给出了从既有检查点继续训练至约 2×10□ 个累计环境步的阶段性结果。250k 至 1.75M 的检查点均采用 50 局评估，进球率依次为 0.42、0.42、0.48、0.48、0.36、0.44 和 0.44。其中 1.25M 检查点出现单次回落，随后在 1.5M 和 1.75M 恢复至 0.44，未出现连续多个检查点共同下降的趋势。

在最终 step 1,999,800（约 2M）检查点上，两批各 100 局的独立评估进球率分别为 0.44 和 0.40，合并结果为 84/200=0.420，Wilson 95% 置信区间为 [0.354, 0.489]。两批合并后的平均回报为 1.006，平均回合长度约为 90.2 个环境步，球权丢失率为 0.52。由此可见，在当前训练配置和这条长程训练轨迹上，模型持续训练至约 2M 环境步后仍保持了与中期检查点相近的任务水平。

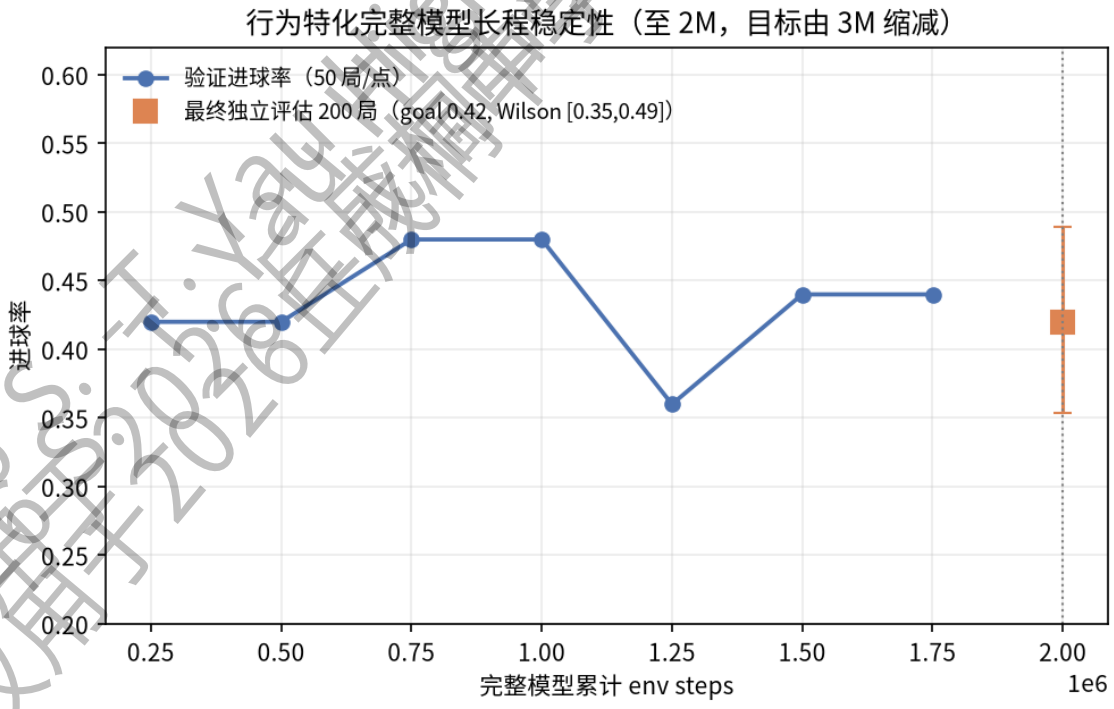


图 4　完整配置持续训练过程中的任务性能。250k–1.75M 检查点为 50 局阶段性评估；约 2M 检查点为两批各 100 局的独立评估，误差范围为最终 200 局进球率的 Wilson 95% 置信区间。

2026 S-T Yau Undergraduate Science Award
仅用于2026丘成桐青年科学家奖

6.4 角色随局势变化，而不是永久绑定球员

图 5 汇总了真实评估回合中的角色动作构成、软角色占用、单回合切换和场上空间分布。沿同一回合、同一智能体的连续时间步统计后，三个角色的自保持概率约为 0.69、0.96 和 0.94。这说明角色通常会持续一段时间，但并未永久绑定到某个球员。

为减少两个通道概率接近时 argmax 短暂翻转的影响，实验进一步采用 3 步和 5 步持续条件过滤角色切换，并结合驻留时长进行判断。过滤后仍存在非零切换，说明角色会随观测和循环状态变化，而不是单纯由智能体身份决定。

动作构成与空间分布进一步显示，不同角色通道在移动、盘带、拦截以及场上占位方面具有不同的统计倾向。由于实际策略使用完整的软角色概率，一个通道即使较少成为最大概率角色，仍可通过非零权重参与动作生成。因此，本文同时参考软占用和硬占用解释角色差异，并不将角色编号直接等同于固定足球职位。

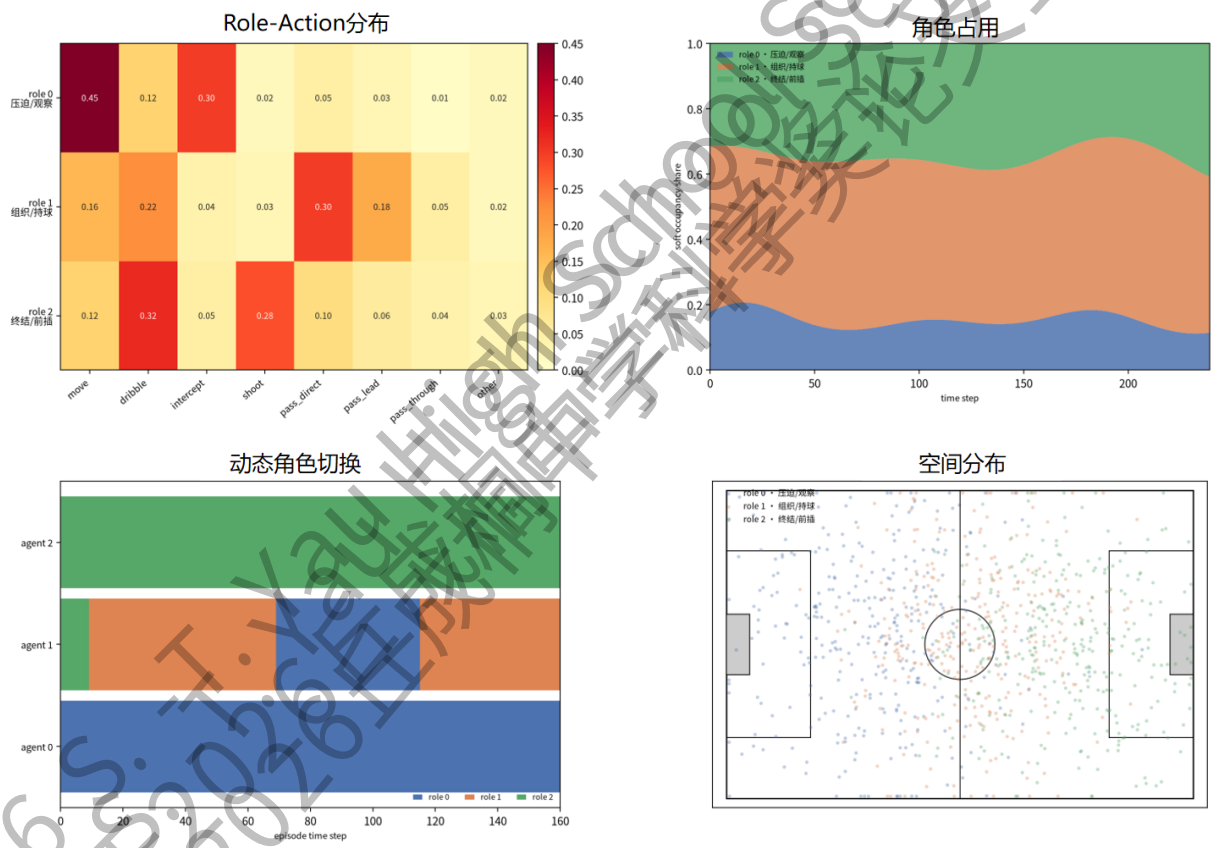


图 5　真实评估回合中的角色行为与动态变化。图中依次给出角色—动作构成、软角色占用、单回合角色切换以及场上空间分布。

7. 讨论与结论

本文围绕 RoboCup 2D 进攻任务中的动态角色分工问题，在共享 MAPPO 策略中构建软角色表示，并利用切空间侧抑制在同周期角色向量之间建立相似性相关的竞争。2v1 受控实验中，加入 TLI 后三种子平均进球率由 0.887 提高至 0.935，提升约 4.8 个百分点；在 3v3 的匹配机制对照中，保留 TLI 的平均进球率

2026 S. T. Yau High School Science Award
仅用于2026年度青少年科学奖论文公示

为 0.365±0.043，高于关闭 TLI 的 0.337±0.060 和 cosine² 正则的 0.320±0.023，对应的角色—动作互信息也分别为 0.387、0.291 和 0.289。虽然三种子波动仍需谨慎解释，但任务指标与行为指标均给出了支持 TLI 继续用于角色竞争的方向性证据。

持续训练实验进一步表明，完整配置在约 200 万个累计环境步的长程轨迹上没有出现持续性的任务性能崩塌，最终两批共 200 局的进球率为 0.420，Wilson 95% 置信区间为 [0.354, 0.489]。在 250k 至 1.75M 的阶段性评估中，进球率主要维持在 0.36–0.48 区间，说明当前配置在较长训练过程中仍能保持可用的进攻能力。这一结果与前述 2v1 和 3v3 受控比较相互补充：前者用于观察侧抑制机制的相对作用，持续训练则用于检查完整配置是否在持续优化中出现明显退化。

上述实验结果是针对于 RoboCup 2D 的 2v1 和 3v3 进攻场景。综合现有实验，可以认为切空间侧抑制为共享策略提供了一种可控的角色分化机制，在受控任务中获得了积极的任务表现，并能够与多角色占用、动态切换和角色相关动作偏好同时存在。后续工作可进一步扩大训练随机性、任务规模和高层协作行为，以检验该动态角色机制在更复杂足球场景中的泛化能力。

2026 S. T. Yau High School Science Award
仅用于2026年成桐中学科学奖论文公示

参考文献

[1] Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The Surprising Effectiveness of PPO in Cooperative Multi-Agent Games. Advances in Neural Information Processing Systems, 35:24611–24624, 2022.

[2] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal Policy Optimization Algorithms. arXiv:1707.06347, 2017.

[3] John Schulman, Philipp Moritz, Sergey Levine, Michael I. Jordan, and Pieter Abbeel. High-Dimensional Continuous Control Using Generalized Advantage Estimation. In ICLR, 2016.

[4] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In EMNLP, pages 1724–1734, 2014.

[5] Tonghan Wang, Heng Dong, Victor Lesser, and Chongjie Zhang. ROMA: Multi-Agent Reinforcement Learning with Emergent Roles. In ICML, PMLR 119:9876–9886, 2020.

[6] Tonghan Wang, Tarun Gupta, Anuj Mahajan, Bei Peng, Shimon Whiteson, and Chongjie Zhang. RODE: Learning Roles to Decompose Multi-Agent Tasks. In ICLR, 2021.

[7] Chenghao Li, Tonghan Wang, Chengjie Wu, Qianchuan Zhao, Jun Yang, and Chongjie Zhang. Celebrating Diversity in Shared Multi-Agent Reinforcement Learning. Advances in Neural Information Processing Systems, 34, 2021.

[8] Harsh Goel, Mohammad Omama, Behdad Chalaki, Vaishnav Tadiparthi, Ehsan Moradi Pari, and Sandeep Chinchali. R3DM: Enabling Role Discovery and Diversity Through Dynamics Models in Multi-agent Reinforcement Learning. arXiv:2505.24265v3, 2025.

[9] H. K. Hartline, H. G. Wagner, and Floyd Ratliff. Inhibition in the Eye of Limulus. Journal of General Physiology, 39(5):651–673, 1956.

[10] David E. Rumelhart and David Zipser. Feature Discovery by Competitive Learning. Cognitive Science, 9(1):75–112, 1985.

[11] Nicolas Boumal. An Introduction to Optimization on Smooth Manifolds. Cambridge University Press, 2023.

[12] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stephane Deny. Barlow Twins: Self-Supervised Learning via Redundancy Reduction. In ICML, PMLR 139:12310–12320, 2021.

[13] Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is All You Need: Learning Skills without a Reward Function. In ICLR, 2019.

2026 S. T. Yau High School Science Award
仅用于2026丘成桐中学科学奖论文公示

[14] Jack Parker-Holder, Aldo Pacchiano, Krzysztof M. Choromanski, and Stephen J. Roberts. Effective Diversity in Population Based Reinforcement Learning. Advances in Neural Information Processing Systems, 33, 2020.

[15] Haobin Qin, Baofeng Zhang, Hidehisa Akiyama, and Keisuke Fujii. R2D-RL: A RoboCup 2D Soccer Environment for Multi-Agent Reinforcement Learning. arXiv:2606.18786v2, 2026.

[16] Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C. Courville, and Marc G. Bellemare. Deep Reinforcement Learning at the Edge of the Statistical Precipice. Advances in Neural Information Processing Systems, 34, 2021.

2026 S.T.Yau High School Science Award
仅用于2026丘成桐中学科学奖论文公示

致谢

本论文的选题来源于我参加机器人足球比赛时的实际经历。2026 年初，在参加多场人形机器人足球比赛中，我发现虽然比赛中已经存在一些固定的角色分配，例如由不同机器人承担主攻、策应等任务，但实际运行时仍然经常出现角色冲突、球权处理不合理，以及有空位却不上前争抢等问题。这些现象让我开始思考，机器人之间是否能够根据场上的情况自主形成更加合理的分工与协同。我开始在相关的线上社区查阅资料，阅读多智能体强化学习，尤其是角色学习的相关论文，研究如何解决机器人足球比赛中多个智能体之间的分工与协同问题。

在此之前，我学习了 AP Calculus BC（考试成绩是 5 分），具备一些基础的高等数学知识。在阅读相关论文的过程中，我仍然遇到了一些的不能理解的数学概念和公式，我想起在中国科学院公众科学日活动中遇到的中国科学院大学数学科学学院的赵彤教授，就向他请教，赵教授热心解答了我的问题。在多次解答我的问题之后，赵教授答应做我本次研究的指导老师。在选题上，赵教授指导我把机器人足球比赛中遇到的问题抽象成为科学问题，以及如何研究和实验验证这些问题；在理论上，赵教授用通俗的语言解释抽象的数学概念，帮助我快速理解研究过程中遇到切空间、黎曼梯度等数学知识，并在公式的推导过程中，帮我指出方向；在对实验数据分析中，赵教授指导我如何选择参数、如何验证论文观点等，帮我把实验、数据和论文形成一体；在论文的写作过程中，赵教授指导我学习用学术语言来撰写论文，例如对于创新点和三层递进的动态角色学习框架的总结，赵教授让我理解了学术表达的边界限定，及其逻辑性和准确性；除上述指导内容之外，赵教授严谨的治学态度和对前沿领域的敏锐动察在整个的研究过程中一直激励着我，尤其是赵教授在繁忙的工作中抽出个人时间，无偿指导我完成本次研究，在此表示忠心感谢。

在研究过程中，线上的机器人和 MARL 的相关社区给了我很多的帮助，Github 中的实验基线（R2DRL)和多个开源的框架，让我的实验有良好的基础。我使用了 Deepseek 和 Chat GPT 5.6 来辅助学习数学和 MARL 相关知识，理解开源框架以及论文检索等。在研究过程中，遇到问题时我会使用 ChatGPT 进行辅助讨论和思路整理，在 IDE 中也会调用 Deepseek 和 Chat GPT 的大模型来辅助进行代码调试等。

最后，感谢我的父母从选题讨论、仿真实验到论文撰写一直陪我推进这项研究，在我遇到困难时，鼓励我不放弃，启发我整理思路继续实验。正是这些讨论、鼓励和一次次对实验过程的修正和论文修改，让我最终完成了本论文。

以下是我的机器人比赛和相关学术活动：

机器人竞赛经历：

1. VEX Robotics Competition —Spin Up 赛季（2022 年-2023 年）

主要负责机器人程序开发和结构设计。随队晋级 VEX World Championship，在 Arts Division 进入四强。

2. FIRST Tech Challenge —DECODE 赛季（2025 年-2026 年）

同时承担机器人程序与结构方面的主要工作。随队参加中国锦标赛，并获得“毕昇分区”Finalist Alliance。

2026 Shanghai High School Science Award
仅用于2026年成绩申请学术公示

3. FIRST Robotics Competition —REBUILT 赛季（2025 年-2026 年）

主要负责机器人程序开发。随队参加上海 Regional，并获得 Engineering Inspiration Award（工程启发奖）。

4. 2026 年海淀区中学生机器人足球比赛

主要负责机器人足球程序的参数调整与现场调试。此次赛事是本论文研究问题的主要来源。

5. 2026 世界人形机器人运动会

担任队长，负责代码调试、赛前测试、现场安排以及队伍整体协调，并参与从赛事报名、机器人准备到正式比赛的大部分工作。此次经历进一步加深了我对多机器人协作与自主决策问题的理解。

6. 目前，我是北京市学生金鹏社团-人大附中 FIRST机器人社团的社长。

相关学术活动：

1. 北京大学斯坦福中心人工智能创客计划：

入选相关人工智能项目组，并围绕 Embodied AI 与 Reinforcement Learning 方向进行学习与探索。相关学习经历帮助我进一步接触强化学习及智能体决策问题，也为本研究中对 MAPPO、多智能体强化学习和角色学习的理解提供了基础。

2026 S.T.Yau High School Science Award
仅用于2026丘成桐中学科学奖公示

附录：理论推导、实现细节与评价定义

A. 符号与切空间侧抑制的几何说明

本节在正文式（3）—（5）的基础上统一符号，并说明 TLI 与球面余弦平方能量之间的关系。推导只描述单次前向表示更新的局部性质，不构成对 PPO 训练收敛或任务性能单调改善的保证。

A.1 主要符号

表 A1　正文主要符号及其含义。

符号	含义	维度或取值
N	同一环境中参与角色竞争的受控智能体数量	正整数
o_i^t	智能体 i 在时刻 t 的局部观测	3v3 中为 100 维
h_i^t	共享编码器与 GRU 产生的循环隐藏状态	64 维
z_i^t	角色编码器的未归一化输出	16 维
u_i^t	归一化角色表示	单位球面 S^{15}
g_i^t	由同周期队友表示产生的切空间竞争方向	16 维
$\hat{u}_i^t$	TLI 更新并回缩后的角色表示	单位球面 S^{15}
p_i^t	K 个软角色通道上的概率向量	$K = 3$，概率和为 1
C_k	第 k 个可学习角色原型	16 维
e_i^t	由软角色概率混合得到的角色嵌入	16 维
m_i^t	环境提供的合法动作掩码	19 维二值向量
s^t	训练 Critic 使用的全局状态	3v3 中为 136 维

A.2 球面余弦平方能量与黎曼梯度

记 $U = (u_1, \dots, u_N)$，且每个 u_i 均为单位向量。考虑对所有非自配对进行平均的余弦平方能量：

$$E(U) = \frac{1}{2(N-1)} \sum_{i=1}^{N} \sum_{j \neq i} \left(u_i^\top u_j\right)^2 \tag{A1}$$

令 $c_{ij} = u_i^\top u_j$。对 u_i 求欧氏梯度，再通过 $I - u_i u_i^\top$ 投影到单位球面的切空间，可得：

$$\nabla_{u_i} E = \frac{2}{N-1} \sum_{j \neq i} c_{ij} u_j, \qquad \operatorname{grad}_{u_i} E = (I - u_i u_i^\top)\, \nabla_{u_i} E = 2 g_i \tag{A2}$$

正文式（4）中的 g_i 与该黎曼梯度方向一致，常数因子可以吸收到步长 α 中。因而，TLI 可以理解为沿球面余弦平方能量的下降方向进行一次前向切向更新，再通过归一化完成回缩。“下降方向” 只针对当前表示与无穷小步长的局部分析；网络参数更新、动作采样和 PPO 优化会共同改变后续动力学。

A.3 切向性质、作用强度与退化点

由切空间投影可知，更新方向与当前角色向量正交，因此更新前的向量范数满足：

$$u_i^\top g_i = 0, \qquad \|u_i - \alpha g_i\|_2^2 = 1 + \alpha^2 \|g_i\|_2^2 \tag{A3}$$

式（A3）说明未归一化更新不会直接沿当前角色向量产生一阶径向收缩；归一化操作将更新后的向量重新映射到单位球面。对单个队友 j，其作用项的模长为：

$$\left\| c_{ij} \left(u_j - c_{ij} u_i \right) \right\|_2 = |c_{ij}| \sqrt{1 - c_{ij}^2} \tag{A4}$$

该模长在 $c_{ij} = 0$ 和 $c_{ij} = \pm 1$ 时均为零，在中等相关区域取得较大值。因此，TLI 不是“相似度越高，排斥必然越强”的单调机制；精确同向或反向仍属于退化情形。若只考虑两个向量的对称无穷小更新，余弦平方的一阶变化为：

$$\left. \frac{\mathrm{d}\, c_{ij}^2}{\mathrm{d}\alpha} \right|_{\alpha = 0} = -4 c_{ij}^2 \left(1 - c_{ij}^2 \right) \leq 0 \tag{A5}$$

除 $c_{ij} = 0, \pm 1$ 的退化点外，式（A5）为非正，说明局部更新倾向于降低该对向量的余弦平方。有限步长、多智能体平均以及策略网络的同步学习不由该式完全刻画。

A.4 与普通 cosine² 正则的区别

普通余弦平方对照使用下式作为附加训练损失：

$$L_{\cos^2} = \frac{2}{N(N-1)} \sum_{i<j} \left(u_i^\top u_j \right)^2 \tag{A6}$$

该正则通过反向传播更新网络参数，不在每次动作前向中直接替换当前角色表示；TLI 则在当前同周期角色向量上执行显式切向竞争。由于角色编码器输出已经归一化，cosine² 正则的梯度同样会受到球面约束影响，因此两者的差异主要在作用位置、时序与控制方式，而不是简单的“有几何”与“无几何”。

A.5 平均聚合与执行信息边界

正文式（4）采用 $1/(N-1)$ 的队友平均，而不是直接求和。其目的在于避免受控智能体数量增加时，单个智能体接收的总竞争幅度仅因队友数量而线性放大。该平均并不假设不同团队规模具有相同表示分布，也不保证不同规模下使用同一 α 会产生完全一致的学习动力学。

TLI 需要在同一环境、同一时刻收集队友角色表示 u_j，因此 Actor 执行除了自身局部历史外，还依赖这一表示交换。该信息路径与集中训练时输入 Critic 的全局状态相互独立：全局状态不进入动作生成，而同周期角色表示交换参与 TLI 前向。

B. 网络结构与训练实现

B.1 核心网络与环境配置

表 B1　网络、环境与优化的核心配置。

2026 S.T. Yau High School Science Award
2026 年 S.T. Yau 高中科学奖 参赛作品
论文公示 2026 年成都市青少年科学实践活动

项目	当前配置	说明
3v3 观测 / 全局状态	100 / 136 维	观测含 97 维环境信息与 3 维身份编码；全局状态仅供 Critic 训练
共享循环特征	隐藏维度 64，GRU	所有智能体共享参数，但分别维护循环状态
角色编码器	64→64→16，ReLU	输出经数值保护归一化到单位球面
角色概率与原型	$K = 3$，$\tau = 1$，$C \in R^{3\times16}$	使用完整软概率混合，不绑定固定足球职位
角色适配器	16→64，初始化 gain=0.01	以残差形式调节隐藏状态，不改变合法动作集合
切空间侧抑制	$\alpha = 0.1$，peer-mean	仅在同环境、同时间的非自身配对间计算
动作输出	19 维 logits	在概率归一化前应用环境合法动作掩码
回合长度	最多 300 环境步	进球、球权丢失、出界或超时可终止回合
Actor / Critic 基础学习率	5×10^{-4}	部分追加训练使用参数组缩放
PPO 与循环采样	$\gamma = 0.99$；GAE $\lambda = 0.95$；PPO epoch=15；mini-batch=1；chunk=10	具体运行沿用对应冻结配置

B.2 前向计算与动作掩码

在每个决策周期，局部观测先由共享编码器和 GRU 转换为 h_i；角色编码器随后生成 z_i 并归一化为 u_i。TLI 只在同一环境、同一时刻的角色向量之间计算，得到 $\hat{u}_i$；角色 logits、软概率 p_i 和原型混合 e_i 均基于 $\hat{u}_i$ 产生。角色嵌入经适配器形成隐藏状态残差，再由共享动作头输出 19 维 logits。

环境掩码 m_i 在 Softmax 归一化前屏蔽不可用动作。角色结构只改变合法动作之间的偏好，不新增或删除环境动作。若某个状态出现全零掩码，应作为环境异常单独记录，而不计入正常策略评价。集中式 Critic 仅在训练时接收 136 维全局状态；执行时动作策略不使用该全局状态。

B.3 附加角色动作通路与训练稳定化设置

部分 3v3 实验在正文式（8）的隐藏角色适配之外，增加轻量角色条件动作残差。对每个角色通道设置一个 64→19 的线性残差头 R_k，并按照软角色概率加权：

$$l_i^{\text{final}} = b_\theta(\tilde{h}_i) + \sum_{k=1}^{K} p_{ik} R_k(\tilde{h}_i) \tag{B1}$$

各残差头在加入时采用零初始化，使初始动作 logits 与未增加该通路时一致。软混合保留了低硬占用角色作为插值通道的可能性，因此不使用 argmax 角色选择单一动作头。该动作残差属于增强角色条件传

2026 S.T. Yau High School Science Award
仅用于2026年丘成桐中学科学奖参赛公示

递的实现设置，不改变正文 TLI 的定义。

部分追加训练还采用轻量的角色行为差异辅助项和冻结参考策略约束。前 50k 微调步的权重分别为 0.05 和 0.02，随后线性衰减至 0.01 和 0.005。

B.4 参数冻结、学习率与训练来源

为保留已经形成的角色坐标系，追加训练中冻结 RoleEncoder、W_r 与角色原型 C；Base / GRU、角色适配器 / 动作头以及 Critic 的学习率分别按基础学习率的 0.03、0.10 和 0.25 倍缩放。训练回报包含环境提供的 EPV 与终局事件信号。

B.5 约 2M 持续训练稳定性检查

长程稳定性实验从完整配置 step 199800 的检查点继续训练，累计训练至 step 1999800。续训过程中不重新搜索 TLI、角色数量或辅助损失系数；250200、500100、750000、1000200、1250100、1500000 和 1750200 检查点分别进行了 50 局阶段性评估。对应进球率为 0.42、0.42、0.48、0.48、0.36、0.44 和 0.44。

表 B2　约 2M 最终检查点的独立评估结果。

批次	回合数	进球率	平均回报	平均回合长度	球权丢失率
A	100	0.440	1.113	91.85	0.49
B	100	0.400	0.900	88.62	0.55
合并	200	0.420	1.006	90.23	0.52

两批最终评估合并得到 84/200=0.420，Wilson 95% 置信区间为 [0.354, 0.489]。该长程曲线属于完整配置自身从约 200k 检查点继续训练的单条轨迹，与作为匹配消融初始化来源的 5M 前驱不是同一条性能曲线。

C. 评价指标与统计口径

C.1 任务指标

进球率沿用正文式（10），即评估回合中以己方进球终止的回合比例。平均回合长度以环境步计，训练回报只在奖励定义一致的配置之间解释。进球、球权丢失、出界和超时应依据环境终止标签计数，而不由回报数值反推事件类型。

C.2 角色占用与有效角色数

设共收集 M 条有效智能体决策记录，p_{nk} 为第 n 条记录在角色 k 上的软概率，硬角色由最大概率通道确定。软占用与硬占用分别为：

$$\bar{p}_k = \frac{1}{M}\sum_{n=1}^{M} p_{nk}, \qquad \bar{q}_k = \frac{1}{M}\sum_{n=1}^{M} \mathbf{1}\left[\arg\max_j p_{nj} = k\right] \tag{C1}$$

角色有效数统一采用占用分布熵的指数形式：

2026 S.T. Yau High School Science Award
仅用于2026年成桐高中科学奖论文公示

$$N_{\text{eff}}^{\text{soft}} = \exp\left[-\sum_k \bar{p}_k \log(\bar{p}_k + \epsilon)\right], \qquad N_{\text{eff}}^{\text{hard}} = \exp\left[-\sum_k \bar{q}_k \log(\bar{q}_k + \epsilon)\right] \tag{C2}$$

C.3 几何指标

对每个同周期非自智能体对计算 $c_{ij} = u_i^\top u_j$，并分别汇总有符号余弦、绝对余弦和余弦平方：

$$\bar{c} = \frac{1}{M_p}\sum c_{ij}, \qquad \overline{|c|} = \frac{1}{M_p}\sum |c_{ij}|, \qquad \overline{c^2} = \frac{1}{M_p}\sum c_{ij}^2 \tag{C3}$$

C.4 角色切换与驻留

硬角色切换只沿同一回合、同一智能体的连续时间边统计。若 $r_{e,i,t} \neq r_{e,i,t-1}$，则记一次原始切换；原始切换率以所有有效连续边为分母。3 步和 5 步过滤要求新角色至少连续保持相应步数才确认一次切换，用于减少概率边界附近的短暂 argmax 抖动。

C.5 角色与实际动作的关联

实际 soft 加权动作画像将每条真实动作按照当时的角色概率分摊到各角色通道：

$$P_k^{\text{soft}}(a) = \frac{\sum_{n=1}^{M} p_{nk}\mathbf{1}[a_n = a]}{\sum_{n=1}^{M} p_{nk}} \tag{C4}$$

经验角色—动作互信息由硬角色与真实动作的联合频数估计：

$$I(R;A) = \sum_{r,a} P(r,a)\log\frac{P(r,a)}{P(r)P(a)} \tag{C5}$$

C.6 随机性与结果汇总

单个固定检查点上的多回合统计主要反映评估抽样变化；不同训练种子的均值和标准差描述训练过程的不确定性。正文对 2v1 报告三个训练种子的均值与标准差，对 3v3 匹配微调则报告三个共享前驱微调种子的结果。

2026 S.T.Yau High School Science Award
仅用于2026丘成桐中学科学奖论文公示