面向具身智能体连续控制任务的可解释线性因果策略模型

平梦梦[1]，李策[1,2]，唐广旭[1]，肖利梅[1]，奉志强[1]，王宗顺[1]
（1. 兰州理工大学自动化与电气工程学院，730050，兰州；2. 兰州理工大学微电子现代产业学院，730050，兰州）

摘要：为解决具身智能体在高维状态与连续动作空间中的策略可解释性问题，提出了基于强化学习的状态－动作结构因果建模框架，利用结构因果模型显式刻画状态与动作依赖，提升策略的可解释性。该框架在仿真环境中训练并冻结策略网络，然后基于经验回放，构建离线状态－动作数据集；在此基础上，采用线性因果结构学习方法，对状态－动作中的因果结构进行显式建模；为提升因果结构学习的稳定性与泛化能力，引入策略敏感度驱动的加权稀疏先验与分组约束正则项，并结合稳定性筛选机制，通过抑制噪声与偶然相关性对结构推断的干扰，提升所学因果关系的性能。在9个MuJoCo动力学仿真平台具身连续控制任务上的实验结果表明：所提方法在平均奖励性能上比软演员－评论家提高了1.58%，中位数提高了0.38%；方法能够学习到稳定的因果结构，在与主流方法的对比中，多数任务表现出具有竞争力的性能。该研究在不增加额外交互成本的前提下实现了有效的因果策略建模，可为连续控制任务中强化学习策略的状态－动作因果分析提供参考。

关键词：深度强化学习；连续控制任务；结构因果模型；线性因果结构学习；策略可解释性
中图分类号：TP18/TP13 **文献标志码**：A
DOI：10.7652/xjtuxbXXXXXX001 **文章编号**：0253-987X(XXXX)XX-0001-12

An Interpretable Linear Causal Policy Model for Continuous Control Tasks of Embodied Agents

PING Mengmeng[1], LI Ce[1,2], TANG Guangxu[1], XIAO Limei[1], FENG Zhiqiang[1], WANG Zongshun[1]
(1. School of Automation and Electrical Engineering, Lanzhou University of Technology, Lanzhou 730050, China; 2. School of Microelectronics and Modern Industry, Lanzhou University of Technology, Lanzhou 730050, China)

Abstract: To address the problem of policy interpretability for embodied agents in high-dimensional state and continuous action spaces, this study proposes a state－action structural causal modeling framework for RL policies, which aims to improve decision interpretability by explicitly characterizing state－action dependencies with structural causal models. The framework first trains and freezes a policy network in simulation, then constructs an offline state－action dataset from an experience replay buffer. Based on this dataset, we adopt a linear causal structure learning method to explicitly model the causal structure between states and actions. To further improve the stability and generalization of causal structure learning, we introduce a policy-sensitivity-driven weighted sparsity

收稿日期：2026-04-22。
网络出版日期：XXXX-XX-XX 网络出版地址：

prior with group-regularized constraints, together with a stability selection mechanism. These components suppress the interference of noise and spurious correlations in structure inference, thereby enhancing the robustness of the learned causal relations. Experimental results on nine MuJoCo dynamics simulation platform embodied continuous control tasks show that the proposed method improves the average reward performance over soft actor-critic by 1.58% and the median performance by 0.38%, while learning stable causal structures. In comparison with mainstream methods, the proposed method demonstrates competitive performance across most tasks, indicating that it can maintain strong control capability while providing explicit causal interpretability. Therefore, this study achieves effective causal policy modeling without introducing additional interaction costs, providing methodological support for state - action causal analysis of reinforcement learning policies in continuous control tasks.

Keywords: deep reinforcement learning; continuous control tasks; structural causal model; linear causal structure learning; policy interpretability

在具身智能与真实机器人部署等需要明确可解释依据的应用场景中,决策模型不仅要能获得较高回报,还要能够说明状态信息如何驱动动作生成,否则将限制策略调试,并在分布外状态下放大不可控行为风险。近年来,深度强化学习[1-4]在连续控制任务中取得了显著进展。基于策略梯度与演员-评论家方法在机器人控制、运动规划以及复杂决策系统中表现出优异性能[5-10]。然而,这类方法通常依赖高维非线性函数逼近器,其策略决策过程以“黑箱”形式存在,难以明确刻画状态变量与动作输出之间的内在依赖关系[11-14]。因此,在尽量保持控制性能的前提下,对策略的决策结构进行可解释建模,已成为强化学习走向可信部署的关键。

面向具身智能体的连续控制任务,强化学习通常需要在高维观测和连续动作空间中学习稳定的闭环控制策略[15],以演员-评论家[16-20]为代表的方法通过随机策略探索与经验回放,提升了样本效率和训练稳定性,已成为机器人控制与运动规划等场景的主流范式。例如,XU等[18]提出动作解耦的混合动作空间强化学习方法,在无人机任务中实现了稳定、高效的控制性能。WU等[19]提出基于Transformer结构的直接信念状态预测方法,通过应对强化学习中的观测延迟问题,有效避免了递归状态预测带来的误差累积。CHENG等[20]提出基于注意力的专家混合多任务强化学习方法,通过为不同任务自适应地组合任务相关专家网络,缓解了参数共享带来的任务干扰。LIU等[21]利用可解释强化学习,识别并纠正智能体决策错误,通过双层优化联合奖励塑形与策略学习,显著提升了强化学习性能。上述方法在训练过程中依赖经验回放缓冲区存储的大量状态-动作交互样本,然而这些数据在传统流程中主要用于价值更新与策略优化,较少被用于揭示策略内部的决策结构。因此,在策略训练完成后,其仍难从结构层面回答,究竟哪些状态变量驱动了动作生成,以及这些依赖在不同采样与噪声扰动下是否稳定且可解释。

结构因果模型(Structural Causal Model,SCM)为刻画变量之间的有向依赖关系提供了明确语义的框架。研究人员将其引入强化学习策略分析,从离线状态-动作数据中显式恢复状态到动作的结构依赖,从而为决策提供解释依据。例如,PROSOERI等[22]指出,因果推断与反事实预测能够支持可信的医疗干预决策。NAKANISHI等[23]提出因果近似逆模型解释框架,将近似逆模型解释与Peter-Clark(PC)因果发现算法相结合,实现了兼具模型可解释性与数据驱动因果分析的全局特征解释,适用于医疗等高可信场景。在稳定性方面,LEE等[24]提出基于条件闭合集合的因果发现,能够在存在高相关性与缺失值的情况下避免不可靠的条件独立检验,并在美国老龄认知研究数据集做了验证。

综上,现有因果发现研究主要集中于环境动力学建模或静态观测数据分析,较少关注强化学习策略的因果结构建模[25-27]。然而,强化学习场景下的状态-动作变量在语义与尺度上高度异质,策略探索噪声与采样偏置易引入偶然相关[28]。在高维连续变量条件下,传统基于独立性检验的约束类方法[29-31]对假设较为敏感,结构学习结果往往不稳定。因此,如何在不引入额外环境交互的前提下,从离线数据中稳定学习可靠的状态-动作因果结构,已

成为亟需解决的问题。

为此，本文提出一种基于软演员-评论家的状态－动作结构因果建模（Soft Actor-Critic-Based State－Action Structural Causal Model，SAC-SA-SCM，SAC-SA-SCM），从策略层面显式建模状态到动作的因果依赖关系。该方法基于冻结的SAC策略与离线状态－动作数据，通过引入有向无环约束学习线性因果结构，并结合策略敏感度加权稀疏先验、分组约束正则化及稳定性筛选机制，实现对策略因果结构的建模。

1　相关工作

1.1　连续控制中的深度强化学习

在连续控制任务中，智能体需要在高维状态观测与连续动作空间下学习稳定的控制策略。与离散动作空间不同，连续控制问题对策略表达能力和优化稳定性提出了更高要求。近年来，基于深度强化学习的方法，在机器人控制、运动规划及复杂物理系统决策等领域均取得了显著进展。

典型方法如软演员-评论家（Soft Actor-Critic，SAC）[16]，通过引入最大熵优化目标，在提升探索能力的同时增强策略学习的稳定性与鲁棒性，已成为连续控制任务中的主流算法之一。此外，围绕策略表达能力和训练效率，研究者提出了多种改进方法，如基于动作解耦的混合动作空间建模、多任务强化学习中的专家混合机制，以及利用Transformer结构进行状态建模与长时依赖建模等。这些方法在复杂环境中均表现出良好的控制性能。

尽管上述方法在性能上已得到不断提升，但其策略通常依赖深度神经网络进行参数化，决策过程以隐式方式编码在高维参数空间中，难以直接刻画状态变量对动作生成的具体作用机制。这种“黑箱”特性在实际应用中限制了策略调试、可靠性分析及安全验证，尤其是在具身智能与真实机器人部署场景中。因此，对策略可解释性的需求日益突出。

1.2　结构因果建模与因果发现方法

SCM模型为刻画变量之间的有向依赖关系提供了具有明确语义的统一框架。有向依赖关系是指变量之间具有方向性的直接影响关系，在因果图中通常表示为由原因变量指向结果变量的有向边。通过引入结构方程与因果图表示，SCM模型能够从观测数据中揭示变量之间的直接影响关系，从而支持因果解释与反事实分析。在复杂系统建模与决策分析等领域，结构因果方法已被广泛应用。

传统因果发现方法主要基于条件独立性检验或评分搜索策略，例如PC算法及基于得分的结构搜索方法。然而，这类方法在高维连续变量场景中往往面临计算复杂度高、对假设敏感以及结构稳定性不足等问题，难以直接应用于强化学习中的连续状态－动作数据。近年来，基于连续优化的因果发现方法为解决上述问题提供了新的思路。以NOTEARS[32]为代表的方法，通过将有向无环图的结构约束转化为连续可微函数，使得因果结构学习问题能够在梯度优化框架下进行求解。该类方法在连续变量与高维数据场景中表现出良好的可扩展性与稳定性，为因果建模在复杂系统中的应用提供了重要工具。

尽管结构因果建模在多个领域取得了进展，但现有研究多集中于静态观测数据分析或系统动力学建模，对强化学习策略内部的决策结构关注较少。特别是在连续控制场景中，如何从离线状态－动作数据中表示具有明确语义的结构依赖关系，并在高维与噪声条件下保持结构稳定性，仍是一个有待深入研究的问题。

1.3　强化学习中的因果建模方法

近年来，因果建模方法逐渐被引入强化学习领域，以提升策略的泛化能力、稳定性和决策可解释性。现有研究主要从环境动力学建模、策略学习增强以及决策分析等角度展开。

在环境建模方面，一类工作通过引入因果结构刻画状态转移机制，从而缓解分布偏移带来的性能退化问题。例如，通过学习具有因果语义的环境模型，策略可在不同任务或环境变化下具备更好的泛化能力。在策略优化方面，也有研究利用因果推断方法进行奖励塑形或策略修正，以减少伪相关对策略学习的干扰。此外，因果推理还被用于支持反事实分析与决策解释，从而提高强化学习在高风险场景中的可信度。

尽管上述方法在不同层面推动了因果方法与强化学习的融合，但现有研究大多集中于环境层面的因果建模或策略学习过程中的辅助优化，对策略本身的结构依赖关系关注较少。特别是在连续控制任务中，策略通常由深度神经网络表示，其内部决策机制以隐式方式存在，难以直接从结构层面刻画状态变量对动作生成的作用路径。另一方面，在高维连续状态－动作空间中，策略数据往往具有强相关性与噪声干扰，传统因果发现方法在此场景下

容易受到采样偏置与偶然相关的影响，导致结构学习结果不稳定。因此，如何在不引入额外环境交互的前提下，从离线状态－动作数据中建模具有明确语义的策略因果结构，仍是强化学习可解释性研究中的关键问题。

2 基于SA-SCM的策略解释方法

2.1 SAC-SA-SCM总体框架

针对深度强化学习策略在连续控制任务中的“黑箱”决策问题，本文提出了SAC-SA-SCM方法，通过结构因果建模对策略的状态－动作映射关系进行显刻画，从而实现对策略决策机制的可解释分析。不同于直接分析策略网络参数，本文将已训练完成的SAC策略视为一个从状态到动作的生成机制，并基于离线交互数据，建模其状态-动作的结构关系。本文SAC-SA-SCM方法的整体框架如图1所示。

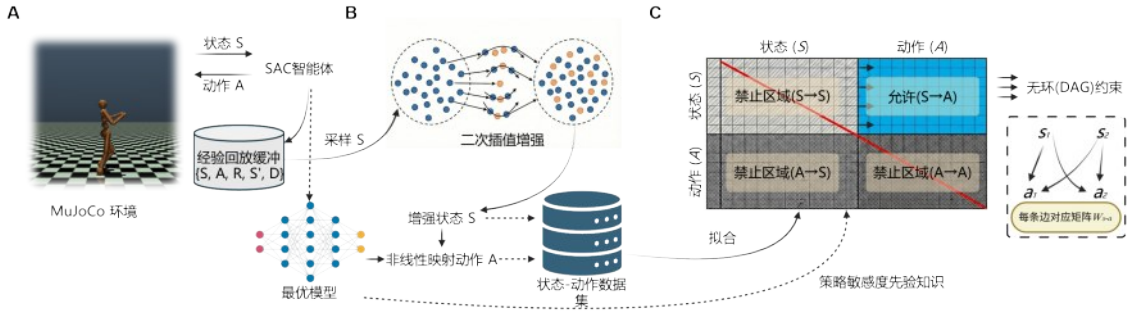


图1 本文SAC-SA-SCM方法整体框架

Fig. 1 Overview of the proposed SAC-SA-SCM framework

SAC-SA-SCM方法包含3个核心环节：首先，A阶段利用SAC与环境交互获得最优策略及经验回放数据；其次，B阶段构建状态－动作结构因果模型，并在结构学习过程中引入策略敏感度先验与分组约束；最后，C阶段通过稳定性筛选提取可靠的因果结构，用于策略解释。

为实现对强化学习策略决策机制的结构化解释，本文在已训练完成的SAC策略基础上，构建状态-动作结构因果模型，用于刻画状态变量对动作决策的直接因果影响关系。

在建模层面，设状态变量 $\boldsymbol{S}\in\mathrm{R}^{d_S}$，动作变量 $\boldsymbol{A}\in\mathrm{R}^{d_A}$，将其拼接为联合变量 $\boldsymbol{X}=(\boldsymbol{S},\boldsymbol{A})\in\mathrm{R}^{d}$，其中 $d=d_S+d_A$，d_s 为状态维度，d_A 为动作维度。基于结构因果模型，采用线性结构形式对变量间依赖关系建模如下

$$\boldsymbol{X}=\mathrm{XW}+\boldsymbol{E} \tag{1}$$

式中：$\boldsymbol{W}\in\mathrm{R}^{d\times d}$ 为结构权重矩阵；$\boldsymbol{E}$ 为未建模扰动项；矩阵元素 $W_{i,j}$ 为变量 x_i 对变量 x_j 的直接因果影响。

考虑到本文关注的是策略层面的决策结构，而非环境动力学过程，因此需要对因果方向进行显式约束。根据策略“状态决定动作”的因果语义，仅允许状态变量→动作变量（$S\rightarrow A$），同时禁止状态间影响关系（$S\rightarrow S$），动作变量作为任何变量的父节点（$A\rightarrow S$，$A\rightarrow A$），并排除自环结构，从而保证所学习因果结构与策略决策机制一致。在该约束下，结构权重矩阵可写为

$$\boldsymbol{W}=\begin{pmatrix}0 & \boldsymbol{W}_{\mathrm{SA}}\\ 0 & 0\end{pmatrix} \tag{2}$$

式中：$\boldsymbol{W}_{\mathrm{SA}}\in\mathrm{R}^{d_S\times d_A}$ 为状态-动作因果子结构，是后续策略解释的核心。

基于上述建模，策略结构的表达式如下

$$\boldsymbol{A}=\boldsymbol{S}\boldsymbol{W}_{\mathrm{SA}}+\boldsymbol{E}_A \tag{3}$$

在结构学习阶段，采用基于连续优化的结构因果建模方法进行参数估计。在状态－动作分块结构（S→A）设定下，因果方向由建模结构天然确定，不存在反馈环依赖，因此无需引入显式有向无环约束。

通过最小化状态－动作重构误差，并结合结构稀疏性约束进行因果结构学习，定义优化目标如下

$$\min_{\boldsymbol{W}}\frac{1}{2n}\|\boldsymbol{X}-\mathrm{XW}\|_F^2+\lambda\|\boldsymbol{W}\|_1 \tag{4}$$

式中：n 为样本数；$\|\cdot\|_F$ 为Frobenius范数；λ 为正则化系数。式(4)中，第1项为状态到动作的重构误差，用于衡量结构模型对动作生成过程的拟合程度，第2项为 ℓ_1 稀疏正则项，用于约束状态－动作依赖结构的稀疏性，避免冗余连接并提升结构可解释性。

模型拟合完成后，从结构矩阵中提取状态-动作子结构 $\boldsymbol{W}_{\mathrm{SA}}$，作为策略解释的核心结果。在推理阶段，给定状态输入 $\boldsymbol{S}$，首先进行标准化处理 $\bar{\boldsymbol{S}}=(\boldsymbol{S}-\mu_S)/\sigma_S$，$\mu_s$ 为状态变量 $\boldsymbol{S}$ 在样本维度上的均值，σ_s 为状态变量 $\boldsymbol{S}$ 在样本维度上的标准差。随后通过结构方程得到动作预测 $\bar{\boldsymbol{A}}=\bar{\boldsymbol{S}}\boldsymbol{W}_{\mathrm{SA}}$，最终经反标准化得到原空间动作输出。该过程实现了从状态到动作的显式因果推理，使得策略决策具备可解释性与可分析性。

2.2　基于策略敏感度的因果先验建模（sens机制）

在SA-SCM结构学习过程中，采用统一的 ℓ_1 正则项约束结构权重矩阵 $\boldsymbol{W}$ 的稀疏性。然而，在连续控制策略中，不同状态变量对动作决策的重要性存在显著差异。若对所有状态-动作连接施加相同的正则惩罚，容易抑制对策略决策具有实际贡献的关键因果依赖关系，从而降低结构解释的有效性。为此，本文引入基于策略敏感度的因果先验建模方法，通过利用已训练策略对输入状态的响应特性，对不同状态-动作连接施加自适应正则权重，从而引导模型优先保留对决策更为重要的因果关系。为便于描述，本文将该机制简称为sens机制。

设策略 π_θ 的动作输出均值函数为 $\mu_\theta(s)$，其中 $s\in\mathbf{R}^{d_S}$，$\mu_\theta(s)\in\mathbf{R}^{d_A}$。定义状态维度 s_i 对动作维度 a_j 的策略敏感度

$$\Gamma_{ij}=\mathbb{E}_{s\sim\mathcal{D}}\left[\left|\frac{\partial\mu_{\theta,j}(s)}{\partial s_i}\right|\right] \tag{5}$$

式中：$\mathcal{D}$ 为从经验回放中采样得到的状态分布；$\mathbb{E}_{s_i\sim\mathcal{D}}[\cdot]$ 操作为在经验回放缓冲区 $\mathcal{D}$ 中采样大量状态 s_i，并对括号内的量进行平均。式(5)计算得到的敏感度刻画了状态变量扰动对策略输出的平均影响强度。

基于上述敏感度，对结构学习中的 ℓ_1 正则项进行加权调整，构造自适应正则系数如下

$$\Lambda_{ij}=\lambda\left(\Gamma_{ij}+\varepsilon\right)^{-\alpha} \tag{6}$$

式中：ε 为数值稳定项；α 为敏感度缩放系数。该形式使得对策略输出影响较大的状态变量（对应较大的 Γ_{ij}）在结构学习中受到较弱的惩罚，从而更容易保留对应的因果连接。

在此基础上，将敏感度先验引入SA-SCM结构学习目标中，优化目标如下

$$\min_{\boldsymbol{W}}\frac{1}{2n}\|\boldsymbol{X}-\mathbf{X}\mathbf{W}\|_F^2+\sum_{i,j}\Lambda_{ij}|\boldsymbol{W}_{ij}| \tag{7}$$

上述敏感度约束仅作用于状态-动作子结构 $\boldsymbol{W}_{SA}$，在不改变SA-SCM因果方向假设的前提下，对不同状态变量的重要性进行差异化建模，从而提升因果结构对策略决策关键变量的刻画能力。

综上可知，sens机制的引入，使得状态-动作结构因果模型策略在保持整体结构稀疏性的同时，有效地保留了对策略行为具有实际贡献的因果依赖关系，从而提高了解释结果的有效性与可信度。

2.3　分组约束下的因果结构正则化（group机制）

在连续控制任务中，状态变量通常具有明确的语义结构，例如位置、速度以及其他辅助状态等。这些变量在物理意义或功能属性上往往呈现出分组特性。然而，在标准的结构学习过程中，各状态维度通常被独立处理，统一施加稀疏正则约束，容易导致同一语义组内的因果连接被不一致地保留或抑制，从而降低因果结构解释的稳定性与一致性。为此，本文引入分组约束机制，通过在结构学习阶段对状态变量施加组级正则化，引导模型在同一语义组内保持一致的因果表达。为便于描述，本文将该机制简称为group机制。

将状态变量集合按照语义或功能划分为若干互不重叠的组 $\{\mathcal{G}_k\}$，其中 $\mathcal{G}_k$ 表示第 k 个状态子集。在SA-SCM策略中，核心关注的是状态到动作的因果子结构 $\boldsymbol{W}_{\mathrm{SA}}$，因此分组约束作用于该子结构的行空间。

在此基础上，引入组级正则项，对每一状态分组对应的因果连接进行整体约束，形式为 $\sum_k\lambda_g^{(k)}\left\|\boldsymbol{W}_{\mathrm{SA}}[\mathcal{G}_k,:]\right\|_1$，其中 $\boldsymbol{W}_{\mathrm{SA}}[\mathcal{G}_k,:]$ 表示 $\mathcal{G}_k$ 对应的所有状态变量到各动作维度的因果连接，$\lambda_g^{(k)}$ 为对应分组 g 的正则系数。该正则形式在结构学习过程中对同一分组内的变量施加一致约束，从而鼓励模型在组内维度上形成一致的稀疏模式，即对于同一语义组中的状态变量，其因果连接倾向于被整体保留或整体抑制，而非出现零散、不稳定的边结构。

进一步地，不同分组可根据其语义重要性设置差异化的正则权重。例如，对于与控制决策密切相关的关键状态组，可设置较小的 $\lambda_g^{(k)}$ 以降低惩罚强度；而对于辅助或冗余状态组，则施加较大的正则约束，以增强结构稀疏性与解释简洁性。

将分组正则项与基本的结构学习目标相结合，优化目标如下

$$\begin{aligned}\min_{\boldsymbol{W}}&\frac{1}{2n}\|\boldsymbol{X}-\mathbf{X}\mathbf{W}\|_F^2+\\&\lambda\|\boldsymbol{W}\|_1+\sum_k\lambda_g^{(k)}\left\|\boldsymbol{W}_{\mathrm{SA}}[\mathcal{G}_k,:]\right\|_1\end{aligned} \tag{8}$$

group机制同样仅作用于状态-动作子结构$\boldsymbol{W}_{\mathrm{SA}}$,在保持SA-SCM因果方向约束的前提下,引导模型学习到更加结构化且具有语义一致性的因果依赖关系。通过引入group机制,SA-SCM能够在高维状态空间中缓解结构学习的不稳定问题,提升因果结构在语义层面的可解释性与一致性。

2.4 训练方法

在完成SA-SCM建模及其敏感度与分组约束设计后,本小节进一步介绍模型的整体训练方法。与上述结构学习的理论形式不同,实际训练过程中需将各项约束统一整合为可优化的损失函数,并结合策略数据进行离线学习。

首先,SAC策略通过与环境交互获得经验回放数据$\mathcal{D}=\{(s_t,a_t)\}$,其中$t$为时间步,即强化学习交互过程中的第$t$个时刻。并固定策略参数$\pi_\theta$。在此基础上,SA-SCM的训练完全基于离线数据进行。

考虑到线性结构因果模型在连续状态空间中缺乏插值泛化能力,本文在训练阶段引入状态插值增强策略。从状态集合中随机采样$s^{(1)},s^{(2)}\in\mathcal{D}$,构造插值状态$\tilde{s}=\dfrac{s^{(1)}+s^{(2)}}{2}$,并通过策略生成对应动作$\tilde{a}=\pi_\theta(\tilde{s})$。由此得到扩展数据集$\tilde{\mathcal{D}}$,用于增强模型对状态空间的覆盖能力。

在训练目标上,将数据重构项、稀疏正则项以及sens机制与group机制统一整合为整体损失函数,表示如下

$$\mathcal{L}(\boldsymbol{W})=\frac{1}{2n}\|\boldsymbol{X}-\boldsymbol{X}\mathbf{W}\|_F^2+\sum_{i,j}\Lambda_{ij}|\mathbf{W}_{ij}|+\sum_k\lambda_g^{(k)}\|\mathbf{W}_{\mathrm{SA}}[\mathcal{G}_k,:]\|_1 \tag{9}$$

式中:第1项为数据重构误差;第2项为基于策略敏感度的加权稀疏正则;第3项为分组结构约束,可在连续优化框架下进行高效求解。

进一步地,为提升因果结构学习的稳定性,本文引入基于重采样的稳定性筛选机制。在训练阶段,采用自助重采样法对扩展数据集进行多次有放回采样,分别训练得到结构矩阵$\{\boldsymbol{W}^{(k)}\}_{k=1}^K$,并统计因果边的出现频率。最终,仅保留在多次训练中稳定出现的因果连接,从而得到更加可靠的结构结果。

综上,SAC-SA-SCM方法的整体训练流程可概括为:首先,通过SAC获得策略与经验数据;其次,进行状态插值增强以扩展数据分布;接着,在此基础上通过融合sens与group约束的结构学习损失进行优化;最后,结合稳定性筛选提取可靠的因果结构。该流程实现了从策略学习到因果解释的统一建模。

3 实验与讨论

3.1 实验设置

为验证所提SA-SCM策略解释方法在不同具身连续控制任务中的性能,基于MuJoCo-v4的9种连续控制基准环境开展系统仿真实验。动作通常表示施加在机器人关节或执行机构上的力矩或驱动力,环境根据该动作更新系统状态并返回奖励。性能值为评估回合上的平均回合累计奖励,性能值越高表示任务完成程度越好。实验中统一采用表1所示的参数设置。

表1 本文SA-SCM算法参数设置

Table 1 Parameter settings of the proposed SA-SCM algorithm

参数	数值
状态采样数	20 000
敏感度计算批大小	4 096
分组正则系数	0.01
稳定性重采样数	20
因果边概率阈值	0.7
权重阈值	0.2
因果子图边数	10
每次评估回合数	50
动作一致性评估步数	5 000

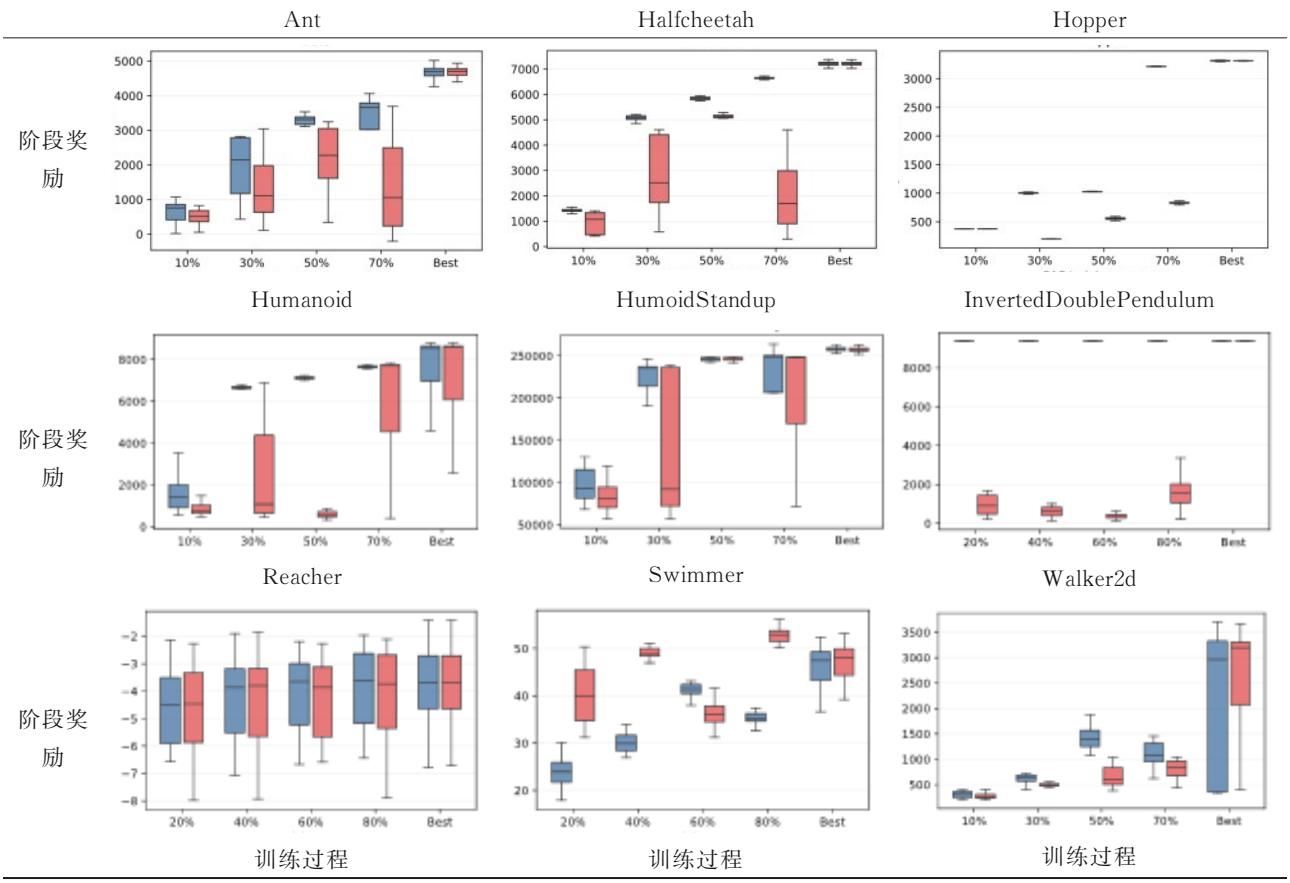
3.2 经验数据规模敏感性实验

为分析经验数据规模对状态-动作因果建模结果的影响,在不同SAC训练阶段下开展了仿真实验。强化学习中的经验数据来自智能体与环境的交互过程,交互步数越多,经验回放池中可用于因果建模的状态-动作样本越多,同时策略对应的状态分布也逐渐发生变化。因此,本文选取SAC训练进度的10%、30%、50%、70%以及充分训练阶段,分别构建状态-动作数据集,并比较SAC与SAC-SA-SCM的任务奖励,结果如图5所示其中蓝色代表SAC,红色代表本文所提方法。

从图5可以看出,在训练早期或经验数据较少时,SAC-SA-SCM的任务回报波动相对较大。在Ant、HalfCheetah、Humanoid和Inverted Double Pendulum等任务中,由于状态空间覆盖不足,所学习的状态-动作依赖关系容易受到局部采样分布

图 2　不同 SAC 训练阶段下 SAC 和 SAC-SA-SCM 的奖励性能对比

Fig. 2　Reward performance comparison of SAC and SAC-SA-SCM under different SAC training progress



影响，导致控制性能不稳定。随着训练进度增加，经验回放池中的状态－动作样本更加充分，策略行为逐渐稳定，SAC-SA-SCM 的任务回报整体趋于稳定，并在充分训练阶段与 SAC 保持接近表现。这说明，不同经验规模会影响显式因果模型的稳定性，但随着强化学习交互数据的不断积累，所学状态－动作因果结构能够较稳定地保留与任务完成相关的关键依赖关系。

3.3　基于关键状态扰动的策略干预效应验证

为验证所学习状态—动作结构是否能够反映原始策略的实际干预响应，本文在冻结策略条件下开展关键状态扰动实验，即保持其余状态维度不变，仅改变目标状态变量，并比较完整模型与原始 SAC 策略的动作输出变化。首先，开展平均因果效应（Average Causal Effect，ACE）一致性分析。在 9 个 MuJoCo 环境中分别选取效应最强的前 5 条状态—动作边，共 45 条。对每条边，在保持其余状态维度不变的条件下，将目标状态变量分别设置为均值加、减一个标准差，并计算对应动作维度的平均输出差，将其作为平均因果效应。随后，对比完整模型与 SAC 策略在相同干预下的效应方向和大小。

如图 3 所示，多数数据点接近一致性虚线，表明两种模型对关键状态扰动的响应总体一致。45 条边中有 37 条方向一致，一致率为 82.2%，且两者效应排序具有显著相关性，说明完整模型能够较好地保持原始 SAC 策略对关键状态扰动的主要响应关系。

为了检验动作响应是否随状态扰动幅度呈现连续且稳定的变化，本工作进一步做了剂量—效应分析。本文在 9 个环境中分别选取完整模型识别出的效应最强状态—动作边。在保持其余状态维度不变的条件下，逐步改变目标状态变量的扰动幅度，即由 逐步变化至 ，并分别计算完整模型与原始 SAC 策略相对于均值干预条件下的动作输出变化，从而得到关键状态变量的剂量—效应曲线。

如图 4 所示，蓝色表示 SAC 方法，红色表示本文方法。本文模型在多数环境中均表现出近似单调的剂量—效应关系，说明随着关键状态变量干预幅度的变化，对应动作输出会产生方向明确的连续

响应。SAC 策略与完整模型在大多数环境中的总体变化方向一致，如 Ant、Hopper 等环境，表明完整模型能够较好地提取原始策略中的主要状态—动作作用方向。与此同时，部分环境中两者的效应幅度存在差异，例如 HalfCheetah 和 HumanoidStandup 中 SAC 响应较弱，而完整模型呈现更明显的线性变化；Swimmer 环境则表现出较强的非线性突变。上述结果表明，完整模型能够刻画原始 SAC 策略的主要干预响应，但受线性结构假设限制，对部分非线性、饱和或弱响应关系仍存在一定近似误差。

3.4 对比实验结果与分析

为展示 SAC-SA-SCM 所提方法的有效性，将本文方法与现有的主流强化学习方法比较，包括 SAC[16]、近端策略优化（Proximal Policy Optimization，PPO）[33]和双延迟深度确定性策略梯度（Twin Delayed Deep Deterministic Policy Gradient，TD3）[34]。这些强化学习方法均采用相同环境交互步数、回合终止条件和评估回合数，并在相同随机种子设置下进行重复实验。此外，为了验证本文因果算法确实能够学习到环境交互中有效状态到动作的因果依赖，将近年来的因果算法与本文实验设置相同情况下的奖励结果进行比较。具体包括可微因果发现（Stable Differentiable Causal Discovery，SDCD）[35]、基于父分数的因果发现（Causal discovery via Parent Score，CaPS）[36]和贝叶斯优化有向无环图（DAG recovery via Bayesian Optimization，DrBO）[37]，通过将这 3 种因果算法采用与 SAC-SA-SCM 一致的状态标准化、线性动作映射、动作反标准化和范围裁剪方式构建可执行策略，从而在统一评估设置下计算累计奖励，其中

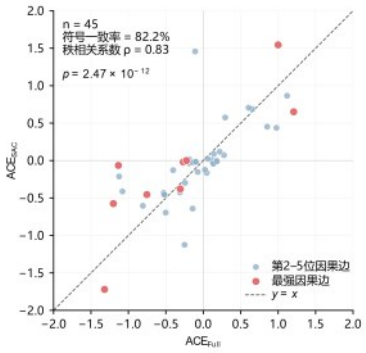
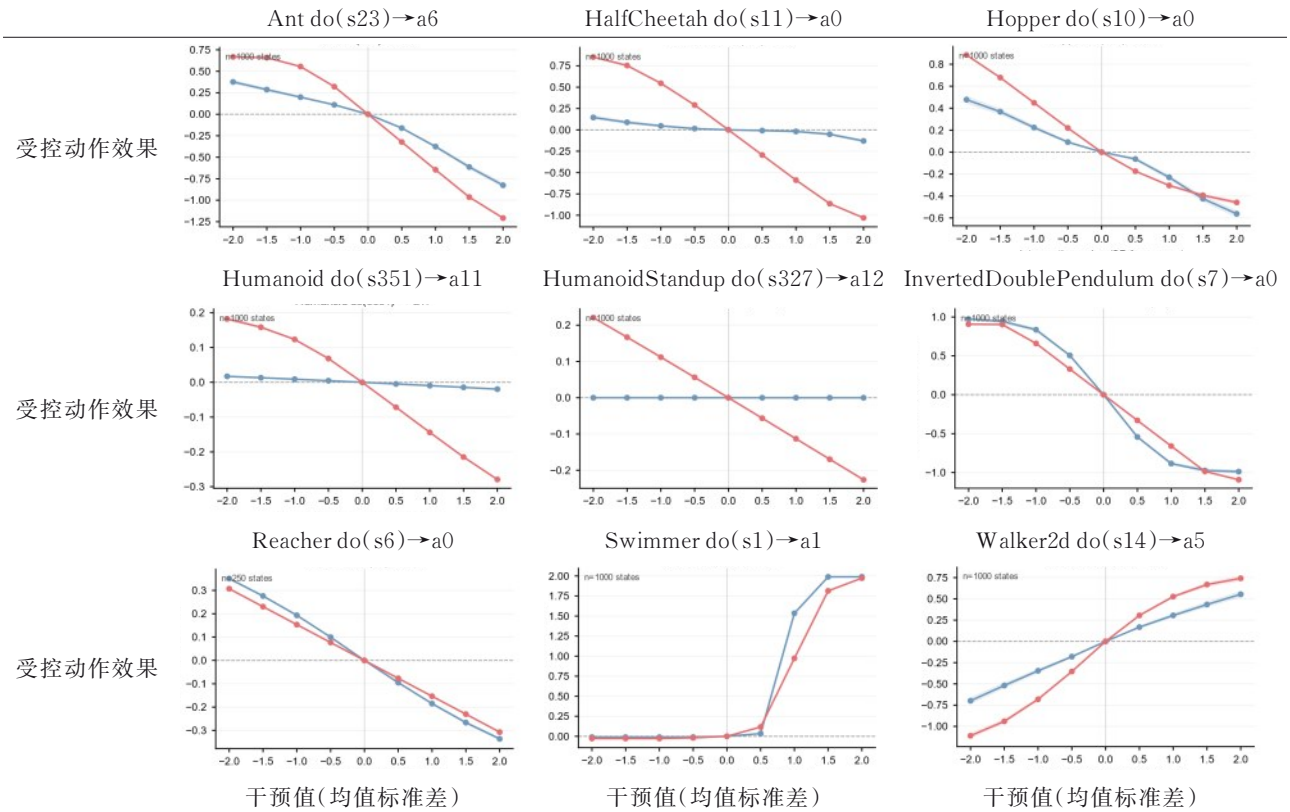


图 3 所提方法与 SAC 策略的平均因果效应一致性分析

Fig. 3 Consistency analysis of average causal effects between the full model and the SAC policy

图 4 关键状态变量的剂量—效应曲线

Fig. 4 Dose - response curves of key state variables



SAC-SA-SCM 在评估阶段不再调用原始 SAC 策略网络，而是由学习得到的线性 SA-SCM 根据当前状态生成动作。

由表 2 可见，在 Humanoid Standup 和 Inverted Double Pendulum 等 5 种环境下，相较于所有主流算法和因果算法，SAC-SA-SCM 的奖励值最大。在其余几种环境下，SAC-SA-SCM 与 SAC 算法的奖励值几乎处于同一水平。该结果表明，虽然 SAC-SA-SCM 将原始深度非线性策略转换为显式线性状态—动作模型，但其仍能够保留原策略中的主要控制关系。与因果算法相比，SAC-SA-SCM 在多数环境中的平均奖励更高，说明通用因果结构学习方法的优化目标并未直接面向强化学习策略中的动作生成过程。相比之下，本文通过限定状态指向动作的结构方向，并引入策略敏感度加权稀疏先验、分组约束和稳定性筛选，使模型能够较有效地保留与动作决策相关的关键状态依赖。

表 2　本文方法和其他方法的平均奖励性能对比

Table 2 Average reward performance comparison between the proposed method and other methods

环境	强化学习			因果算法			SAC－SA－SCM (full)
	SAC[16]	PPO[33]	TD3[34]	SDCD[35]	CaPS[36]	DrBO[37]	
Ant	**4 629　303**	2 139.706± 593.661	949.744± 24.586	－209.646± 143.104	232.662± 78.302	378.206± 230.596	4 312± 1 228
HalfCheetah	**7 214　77**	1 637.137± 27.410	5 511.829± 121.837	－298.268± 28.940	－2.323± 122.135	1 002.845± 425.613	7 201±109
Hopper	3 314±7	**3 393.100 2.965**	5.896±0.054	4.134±0.656	3.337±0.440	210.530± 46.780	3 314±3
Humanoid	**7 582 1 832**	3 202.735±1 488.742	586.092± 134.665	832.662± 578.038	156.636± 19.198	242.092± 13.673	7 365± 1 979
Humanoid Standup	256 256± 13 453	135 222.167± 17 013.598	55 482.249± 53.505	57499.624± 8716.653	74 337.644± 9 955.693	93 205.848± 37746.100	**257 266 3 021**
Inverted Double Pendulum	9 359.740± 0.184	9 048.669± 1 247.670	8 996.417± 1 775.736	164.469± 46.088	9 359.734± 0.225	733.164± 530.507	**9 359.760 0.211**
Reacher	－3.695 ± 1.269	－6.106± 1.706	－4.397± 1.371	－9.601± 3.998	－3.694± 1.274	－6.060± 1.460	**－3.681 1.271**
Swimmer	46.029 ± 4.534	46.598±3.698	13.052± 10.336	36.339± 3.364	46.997± 2.885	46.504±2.835	**47.135 3.973**
Walker2d	2 144± 1 355	699.642± 7.850	2 139.872± 8.505	356.952± 13.872	477.955± 26.315	357.831± 126.390	**2 592 1 020**

注：加粗数据为最优值。强化学习方法均采用相同的环境交互步数、随机种子进行独立训练与测试，因果算法均使用相同的 SAC 离线状态－动作数据进行结构学习。

3.5　消融实验结果与分析

为了验证所提 SAC-SA-SCM 模型是否具备一定的决策性能，表 3 展示了 9 种不同环境下，所提模型与 SAC、SCM 两种基准方法，以及分别与 group 和 sens 两种机制结合方法的平均奖励性能对比。其中，± 后为标准差；括号中数值表示相对于 SAC 基线的奖励变化百分比 Δ%。

由表可见，在各 MuJoCo 环境中，SAC 获得了稳定控制性能。未加入额外机制的 SCM 在部分高维或耦合环境中性能下降明显，而 group 和 sens 方法在多数环境中显著改善了性能，所提方法在平均奖励性能上提升了 1.58%，中位数 0.38%，多数环境中保持接近或略优于 SAC 的控制水平。

上述结果表明，单纯线性 SCM 虽然能够获得显式的状态－动作依赖结构，但在复杂连续控制任务中容易受到状态相关性、采样噪声和偶然相关的影响，导致关键控制依赖未能被稳定保留。分组约束能够利用状态变量的语义或功能分组信息，策略敏感度先验能够引导模型保留对动作输出影响较大的关键状态变量。3 种机制结合后，SAC-SA-

表3 不同机制下的平均奖励性能表现

Table 3 Average reward performance under different mechanisms

环境	SAC	SCM	group	sens	SAC－SA－SCM
Ant	**4 629 303**	4 124 ± 1 431 (－10.91%)	4 193 ± 1 426 (－9.42%)	4 016 ± 1 680 (－13.24%)	4 312 ± 1 228 (－6.85%)
HalfCheetah	**7 214 77**	6 716 ± 1 192 (－6.90%)	6 916 ± 862 (－4.13%)	7 083 ± 363 (－1.82%)	7 201 ± 109 (－0.18%)
Hopper	3 314 ± 7	2 047 ± 1 000 (－38.23%)	**3 319 6 (0.15%)**	2 954 ± 838 (－10.86%)	3 314 ± 3 (0)
Humanoid	7 582 ± 1 832	7 289 ± 2 073 (－3.86%)	**7 869 1 566 (3.79%)**	7 338 ± 2 033 (－3.22%)	7 365 ± 1 979 (－2.86%)
Humanoid Standup	256 256 ± 13 453	257 186 ± 4 140 (0.36%)	257 752 ± 2 938 (0.58%)	**258 682 4 022 (0.95%)**	257 266 ± 3 021 (0.39%)
Inverted Double Pendulum	9 359.740 ± 0.184	294 ± 113 (－96.86%)	**9 359.775 0.203 (0)**	9 359.767 ± 0.191 (0)	9 359.760 ± 0.211 (0)
Reacher	－3.695 ± 1.269	－3.690 ± 1.278 (0.14%)	－3.709 ± 1.281 (－0.38%)	－3.696 ± 1.270 (－0.03%)	**－3.681 1.271 (0.38%)**
Swimmer	46.029 ± 4.534	46.038 ± 4.286 (0.02%)	45.929 ± 4.002 (－0.22%)	45.919 ± 4.167 (－0.24%)	**47.135 3.973 (2.40%)**
Walker2d	2 144 ± 1 355	678 ± 948 (－68.38%)	2 358 ± 972 (9.98%)	2 358 ± 655 (9.98%)	**2 592 1 020 (20.90%)**
平均值		－25.62%	0.04%	－2.05%	**1.58%**
中位数		－6.90%	0.00%	－1.82%	**0.38%**

注：加粗数据为最优值。

SCM能够在保持接近SAC控制性能的同时，输出具有明确方向语义的状态－动作因果结构，从而在策略性能与可解释性之间取得较好的平衡。

4 结 论

针对具身智能体连续控制任务中深度强化学习策略可解释性不足的问题，本文提出SAC-SA-SCM算法，将状态－动作结构因果建模融入强化学习策略解释流程，实现了连续控制任务中策略决策依赖关系的显式表达。研究得到的主要结论如下。

(1)提出了面向强化学习连续控制任务的状态到动作显式因果建模框架。该框架基于训练完成的SAC策略和经验回放缓冲区中的状态－动作样本，学习状态变量到动作维度之间的结构化依赖关系，从而实现策略决策过程的因果解释，使模型能够回答“哪些状态变量驱动了动作生成”的问题。

(2)提出了策略敏感度驱动的加权稀疏先验、分组约束正则化和稳定性筛选机制。加权稀疏先验用于突出对动作输出更敏感的关键状态变量，分组约束正则化用于增强状态－动作依赖结构的一致性，稳定性筛选机制通过多次自助重采样统计因果边出现频率，抑制噪声扰动和偶然相关导致的不稳定因果连接。

(3)在9个MuJoCo连续控制环境中进行了实验验证。结果表明，SAC-SA-SCM在保持SAC原有控制能力的同时，平均性能相比SAC提高1.58%，中位数性能提高0.38%；同时，该方法能够输出人类可理解的状态－动作因果图结构，实现连续控制策略的结构化解释。

未来工作将进一步在此基础上进行扩展，如实现非线性的因果结构建模或者更多基于此的优化算法，在此基础上能够适应更加复杂的控制任务。

参考文献：

[1] SEWAK M. Deep Reinforcement Learning [M]. Springer, 2019.

[2] HENDERSON P, ISLAM R, BACHMAN P, et al. Deep Reinforcement Learning That Matters [C]. Proceedings of the AAAI Conference on Artificial

Intelligence, 2018, 32(1): 3207-3214.

[3] LE N, RATHOUR V S, YAMAZAKI K, et al. Deep Reinforcement Learning in Computer Vision: A Comprehensive Survey [J]. Artificial Intelligence Review, 2022, 55(4): 2733 - 819.

[4] 刘光远, 曹晶仪, 杜婕. 联合遗传算法和强化学习的虚拟网络功能映射与调度方法 [J]. 西安交通大学学报, 2024, 58(8): 175 - 84.
LIU G, CAO J, DU J. Function Mapping and Scheduling Method of Virtual Network Combining Genetic Algorithm and Reinforcement Learning [J]. Journal of Xi'an Jiaotong Univerisity, 2024, 58(8): 175 - 84.

[5] METZ Y, GEISZL A, BAUR R, et al. Reward Learning from Multiple Feedback Types [C] proceedings of the International Conference on Learning Representations, 2025: 1 - 51.

[6] DONG J, HSU H-L, PAJIC M, et al. Efficiently Robust in-Context Reinforcement Learning with Adversarial Generalization and Adaptation [C]. Proceedings of the Advances in Neural Information Processing Systems Workshop: Reliable ML from Unreliable Data, 2025: 1 - 37.

[7] CALLAGHAN A, MASON K, MANNION P. Moma-Ac: A Preference-Driven Actor-Critic Framework for Continuous Multi-Objective Multi-Agent Reinforcement Learning [J]. Neurocomputing, 2025: 132032.

[8] LAN G, HAN D-J, HASHEMI A, et al. Asynchronous Federated Reinforcement Learning with Policy Gradient Updates: Algorithm Design and Convergence Analysis [C]. Proceedings of the International Conference on LearningRepresentations, 2024: 1-31.

[9] WANG Z, LIU J, PAN L. Learning Intractable Multimodal Policies with Reparameterization and Diversity Regularization [J]. arXiv preprint 2025: arXiv:2511.01374.

[10] 马宇, 安豆, 林熙祥, 等. 面向飞行器智能协同控制的分层双时延策略梯度强化学习方法 [J]. 西安交通大学学报, 2025, 59(9): 88 - 98.
MA Y, AN D, LIN X, et al. Hierarchical TwinDelayed Policy Gradient Reinforcement Learning for Intelligent Cooperative Control of Aircraft [J]. Journal of Xi'an Jiaotong University, 2025, 59(9): 88-98.

[11] DELFOSSE Q, SHINDO H, DHAMI D, et al. Interpretable and Explainable Logical Policies Via Neurally Guided Symbolic Abstraction [J]. Advances in Neural Information Processing Systems, 2023, 36: 50838 - 58.

[12] 唐蕾, 牛园园, 王瑞杰, 等. 强化学习的可解释方法分类研究 [J]. 计算机应用研究, 2024, 41(6): 1601 - 1609.
TANG L, NIU Y, WANG R, et al. Classification study of interpretable methods for reinforcement learning [J]. Application Research of Computers, 2024, 41(6): 1601-1609.

[13] GLANOIS C, WENG P, ZIMMER M, et al. A Survey on Interpretable Reinforcement Learning [J]. Machine Learning, 2024, 113(8): 5847 - 90.

[14] MOTT A, ZORAN D, CHRZANOWSKI M, et al. Towards Interpretable Reinforcement Learning Using Attention Augmented Agents [C]. Proceedings of the Advances in Neural Information Processing Systems, 2019, 32: 1 - 10.

[15] TODOROV E, EREZ T, TASSA Y. Mujoco: A Physics Engine for Model-Based Control [C] Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, 2012: 5026 - 33.

[16] HAARNOJA T, ZHOU A, ABBEEL P, et al. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor [C]. Proceedings of the International conference on machine learning, 2018: 1861 - 70.

[17] ADAMCZYK J, MAKARENKO V, TIOMKIN S, et al. Average-Reward Soft Actor-Critic [C]. Proceedings of the Reinforcement Learning Conference, 2025: 1 - 19.

[18] XU Y, WEI Y, JIANG K, et al. Action Decoupled Sac Reinforcement Learning with Discrete-Continuous Hybrid Action Spaces [J]. Neurocomputing, 2023, 537: 141 - 51.

[19] WU Q, WANG Y, ZHAN S S, et al. Directly Forecasting Belief for Reinforcement Learning with Delays [C]. Proceedings of the International Conference on Machine Learning, 2025.

[20] CHENG G, DONG L, CAI W, et al. Multi-Task Reinforcement Learning with Attention-Based Mixture of Experts [J]. IEEE Robotics and Automation Letters, 2023, 8(6): 3812 - 3819.

[21] LIU S, ZHU M. Utility: Utilizing Explainable Reinforcement Learning to Improve Reinforcement Learning [C]. Proceedings of the International Conference on Learning Representations, 2025: 1 - 30.

[22] PROSPERI M, GUO Y, SPERRIN M, et al. Causal

Inference and Counterfactual Prediction in Machine Learning for Actionable Healthcare [J]. Nature Machine Intelligence, 2020, 2(7): 369 - 75.

[23] NAKANISHI T. Causalaime: Leveraging Peter-Clark Algorithms and Inverse Modeling for Unified Global Feature Explanation in Healthcare [C] Proceedings of the World Conference on Explainable Artificial Intelligence, 2025: 332 - 56.

[24] LEE K, RIBEIRO B, KOCAOGLU M. Constraint-Based Causal Discovery from a Collection of Conditioning Sets [C]. Proceedings of the Conference on Uncertainty in Artificial Intelligence, 2025: 1 - 31.

[25] CHEN Z, TIAN Z, ZHU J, et al. C-Cam: Causal Cam for Weakly Supervised Semantic Segmentation on Medical Image [C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 11676 - 85.

[26] SUN Z, HAN Q, YANG H, et al. Invariant Deep Uplift Modeling for Incentive Assignment in Online Marketing Via Probability of Necessity and Sufficiency [C]. Proceedings of the International Conference on Machine Learning, 2025: 1 - 19.

[27] SUN Z, HAN Q, YANG H, et al. Invariant Deep Uplift Modeling for Incentive Assignment in Online Marketing Via Probability of Necessity and Sufficiency [C]. Proceedings of the International Conference on Machine Learning, 2025.

[28] SPIRTES P, GLYMOUR C. An Algorithm for Fast Recovery of Sparse Causal Graphs [J]. Social Science Computer Review, 1991, 9(1): 62 - 72.

[29] YI H, HE Y, CHEN D, et al. The Robustness of Differentiable Causal Discovery in Misspecified Scenarios [C]. Proceedings of the International Conference on Learning Representations, 2025: 1 - 39.

[30] 孙悦雯, 柳文章, 孙长银. 基于因果建模的强化学习控制: 现状及展望 [J]. 自动化学报, 2023, 49(3): 661 - 77.
SUN Y-W, LIU W-Z, SUN C-Y. Causalityin Reinforcement Learning Control: The State of the Art and Prospects. Acta Automatica Sinica, 2023, 49(3): 661—677.

[31] KARLSSON R K A, KRIJTHE J H. Falsification of Unconfoundedness by Testing Independence of Causal Mechanisms [C]. Proceedings of the International Conference on Machine Learning, 2025: 1-20.

[32] SUN X, SCHULTE O, LIU G, et al. Nts-Notears: Learning Nonparametric Dbns with Prior Knowledge [C]. Proceedings of the International Conference on Artificial Intelligence and Statistics, 2023: 1-23.

[33] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal Policy Optimization Algorithms [J]. arXiv preprint arXiv:1707.06347, 2017.

[34] FUJIMOTO S, HOOF H, MEGER D. Addressing Function Approximation Error in Actor-Critic Methods [C]. Proceedings of the International Conference on Machine Learning. PMLR, 2018: 1587-1596.

[35] NAZARET A, HONG J, AZIZI E, et al. Stable Differentiable Causal Discovery [J]. arXiv preprint arXiv:2311.10263, 2023.

[36] XU Z, LI Y, LIU C, et al. Ordering-based Causal Discovery for Linear and Nonlinear Relations [C]. Proceedings of the Advances in Neural Information Processing Systems, 2024, 37: 4315-4340.

[37] DUONG B, GUPTA S, NGUYEN T. Causal Discovery via Bayesian Optimization[J]. arXiv preprint arXiv:2501.14997, 2025.