深度强化学习中策略表征研究简述

陈真[2,3]，吴卓屹[2,3]，张霖[1,2,3*]

(1. 北京航空航天大学 杭州国际创新研究院，浙江 杭州 311115；2. 北京航空航天大学 自动化科学与电气工程学院，北京 100191；3. 复杂产品智能制造系统技术全国重点实验室，北京 100854)

摘要：深度强化学习(deep reinforcement learning，DRL)在多个领域取得了显著成功，但DRL的策略网络在泛化性、多任务适应性和样本效率等方面仍面临巨大挑战。策略表征作为提升DRL能力的重要研究方向，通过构建更高效、更泛化的策略表达形式，提升智能体对环境变化及新任务的适应能力。概述了策略表征领域的关键研究进展，介绍了从传统的基于多层感知机(multi-layer perceptron，MLP)策略到基于指针网络、序列生成模型、扩散模型、超网络、模块化设计以及专家混合模型以及基于序列化Token的跨模态策略等多样化策略架构，还从策略输入和中间表达的语义如何编码和优化等策略表征方法层面归纳分析前沿研究。总结并对未来可能的发展趋势进行了展望。

关键词：策略表征；深度强化学习；泛化能力；多任务学习

中图分类号：TP391.9 文献标志码：A 文章编号：1004-731X(2025)07-1753-17

DOI: 10.16182/j.issn1004731x.joss.25-0533

引用格式: 陈真, 吴卓屹, 张霖. 深度强化学习中策略表征研究简述[J]. 系统仿真学报, 2025, 37(7): 1753-1769.

Reference format: Chen Zhen, Wu Zhuoyi, Zhang Lin. Research on Policy Representation in Deep Reinforcement Learning[J]. Journal of System Simulation, 2025, 37(7): 1753-1769.

Research on Policy Representation in Deep Reinforcement Learning

Chen Zhen[2,3], *Wu Zhuoyi*[2,3], *Zhang Lin*[1,2,3*]

(1. Hangzhou International Innovation Institute, Beihang University, Beijing 311115, China; 2. School of Automation Science and Electrical Engineering, Beihang University, Beijing 100191, China; 3. State Key Laboratory of Intelligent Manufacturing Systems Technology, Beijing 100854, China)

Abstract: Deep reinforcement learning (DRL) has achieved remarkable success in various domains. Nevertheless, existing policy networks in DRL still face significant challenges in areas such as generalizability, multi-task adaptability, and sample efficiency. Policy representation, as a crucial research direction for enhancing DRL capabilities, aims to improve an agent's adaptability to environmental changes and novel tasks by constructing more efficient and generalizable forms of policy expression. This paper provided a concise overview of key research advances in the field of policy representation. It introduced diverse policy architectures, ranging from traditional multi-layer perceptron (MLP) -based policies to those based on pointer networks, sequence generation models, diffusion models, hypernetworks, modular designs, mixture of experts models, and cross-modal policies based on serialized tokens. The paper sorted out cutting-edge research concerning policy representation methods, specifically addressing how semantic information within policy inputs and intermediate representations is encoded and optimized. It concluded with a summary and discussed prospects for future development.

Keywords: policy representation; deep reinforcement learning; generalizability; multi-task learn

收稿日期：2025-06-09 修回日期：2025-06-23

基金项目：国家自然科学基金(62373026)

第一作者：陈真(1998-)，男，博士，研究方向为基于深度强化学习的组合优化。

通信作者：张霖(1966-)，男，博士，教授，研究方向为复杂系统建模仿真。

0 引言

深度强化学习(deep reinforcement learning，DRL)通过融合强化学习(RL)的决策优化能力与深度学习的强大特征表征能力[1]，在许多领域取得了显著的突破，例如机器人控制(如双足机器人运动规划与控制[2])、多自由度机械臂操作[3]、游戏决策(如AlphaGo[4]、StarCraft[5])、大模型后训练(如ChatGPT类模型[6]的后续策略优化)等。然而，当前主流DRL方法通常针对特定任务或环境进行设计，难以在任务或环境发生变化时快速适应，泛化能力普遍不足[7]。这一定程度上限制了RL技术在真实复杂环境中的应用部署。

泛化能力的提升对RL在真实世界中的成功落地至关重要。与特定任务定制的RL策略不同，具备良好泛化能力的策略能快速迁移到不同任务或未见过的环境中，有效降低训练代价，提升样本效率，显著拓展RL的实际应用范围[8]。策略表征是近年来提出的重要思路，意在通过改进策略网络的表达形式，提升智能体的泛化能力与多任务适应性[9]。

策略表征是指策略网络如何将观测状态或任务信息有效地编码、组织并映射为智能体的行动策略[10]。早期的策略网络主要以多层感知机(multi-layer perceptron，MLP)为主，这种结构简单且易于实现，但难以处理高维或结构复杂的输入信息，也难以实现跨任务的有效泛化。近年来，针对各类应用需求，研究者引入了更为复杂但有效的策略表征架构，包括序列建模、扩散模型、超网络等先进结构。这些新型策略表征架构展现出更强的泛化性和迁移能力，在复杂任务环境中取得了显著效果。

鉴于策略表征对RL泛化的关键作用，本文系统梳理DRL中策略表征的最新进展与方法。通过分类总结与横向对比，期望为读者勾勒出策略表征研究的技术谱系，并进一步探讨当前存在的主要挑战及未来值得关注的研究方向，为从事DRL算法研究与实际应用的相关学者和工程人员提供参考与启示。

1 背景

1.1 深度强化学习概述

1.1.1 基本概念

强化学习(RL)是机器学习中的一个重要分支，研究智能体如何通过与环境交互学习获得最优的行为策略，以最大化长期累积回报。RL通常基于马尔可夫决策过程(Markov decision process，MDP)进行形式化建模。一个标准的MDP可以定义为元组：

$$M=(S,A,P,r,\gamma) \tag{1}$$

式中：S为状态空间，包含智能体可感知的所有状态；A为动作空间，定义智能体可执行的所有动作；P：$S\times A\times S\to[0,1]$为状态转移概率，表示在当前状态$s$下执行动作$a$后进入新状态$s'$的概率，记为$P(s'|s,a)$；$r\to\mathbb{R}$, 为奖励函数，表示在状态$s$下执行动作$a$后获得的即时奖励，记为$r(s,a)$；$\gamma\in[0,1]$为折扣因子(Discount Factor)，用以平衡当前奖励与未来奖励之间的权衡关系。RL的目标是寻找一个最优策略π^*，以最大化预期折扣累积奖励，即：

$$\pi^*=\arg\max_{\pi} E_{\tau\sim\pi}[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)] \tag{2}$$

式中：$\tau=(s_0,a_0,s_1,a_1,\cdots)$为智能体与环境交互所产生的一条轨迹。通常在策略学习过程中，定义状态值函数和动作值函数。对于基于策略梯度的方法，策略通常用参数化函数$\pi_\theta(a|s)$表示，通过优化策略梯度来寻找最优策略。

DRL是RL与深度学习融合发展的产物，即在传统RL范式中引入深度神经网络作为策略或价值函数的函数逼近器，从而显著提升了RL对高维、复杂问题的处理能力。

1.1.2 典型应用中的建模方式

DRL 在实际应用中通常需结合特定问题的领域知识进行建模，以有效描述状态、动作及奖励等关键元素。本文以机器人控制与大模型后训练2个典型领域的建模方式为例说明。

(1) 机器人控制

机器人控制领域的 RL 建模通常包括以下要素：

状态空间 S：一般由机器人自身的状态(如关节位置、速度)和外部环境状态(如目标位置、障碍物信息)组成。

动作空间 A：常为连续的关节控制指令(例如关节的角速度或力矩)。

奖励函数 $r(s,a)$：通常设计为任务达成度的奖励与动作成本的惩罚，如：

$$r(s,a)=r_{\text{goal}}(s)-\alpha\|a\|^2 \tag{3}$$

代表性的机器人控制 RL 任务包括机械臂的精准抓取任务[11]、双足机器人的步态规划[12]以及四足机器人的运动控制[13]等。

(2) 大模型后训练

大模型的后训练阶段通常包括监督微调和基于人类反馈的强化学习(RL from human feedback，RLHF)等方法。其中，RLHF 是提升模型对齐能力的核心技术，通过引入人类偏好指导策略优化，使生成结果更符合人类预期[14]。其 RL 建模方式为

状态空间 S：大模型生成的序列或其上下文表示。

动作空间 A：大模型生成的下一个词(Token)或响应内容。

奖励函数 $r(s,a)$：通常由人类反馈或模型自身评价指标(如输出文本的连贯性、准确性、符合用户意图等)构成。

典型的应用场景包括 OpenAI 提出的 ChatGPT[15]、Anthropic 的 Claude 系列模型[16]等，它们通过 RL 进一步优化初始预训练模型，显著提高模型输出的质量与安全性。

1.2 DRL 策略表征

在 DRL 中，策略表征指的是如何通过神经网络结构与表征机制，将环境状态或任务上下文映射为合理的动作分布，从而支持策略的优化与泛化过程[10]。形式上，一个参数化策略可记为

$$\pi_\theta(a|s)=f_\theta(s) \tag{4}$$

式中：θ 为策略网络的参数；f_θ 为策略的函数近似器，其输出通常为一个动作的概率分布(离散动作空间)或均值与方差参数分布(连续动作空间)。

在具备任务上下文或目标信息的场景下，如多任务学习[17]、元强化学习(Meta-RL)[18]、目标条件策略[19]等，策略表征通常需引入附加输入 c，构成条件策略形式为

$$\pi_\theta(a|s,c) \tag{5}$$

式中：c 为任务嵌入、历史轨迹、提示向量(Prompt)或语言描述等，旨在为策略提供环境外的辅助信息。

策略表征的好坏直接影响模型的泛化能力(策略能否在未见过的状态、任务或环境变化下保持有效性)、样本效率(策略在有限交互下是否能快速获得高质量的策略映射)、适应性与迁移性(是否能够快速适应新任务或实现零样本迁移)。因此，策略表征不仅是策略网络设计的核心问题之一，也是当前 DRL 性能及泛化性研究的关键切入点。

在早期的 RL 研究中，策略表征主要依赖简单的 MLP 结构。该结构直接将状态输入映射为动作概率分布或期望动作输出。这种方式在低维、规则性强的任务中具备良好效果，但在处理高维感知输入(如图像、点云)、结构化任务(如图结构、多智能体系统)或任务分布多样的场景(如元学习、多任务学习)时存在显著局限性。为解决上述问题，研究者在策略表征方面提出了多种创新架构与方法。本文将从架构与方法视角进行更加详细的介绍。

2 策略表征架构

在DRL中，策略的表达能力直接影响智能体对环境状态的理解、动作选择的准确性以及策略在复杂任务中的泛化表现。在本文中，策略表征架构主要指的是用于建模策略函数的神经网络结构形式及其内部组成单元，即策略网络“如何建”的问题。不同架构在表达能力、参数共享机制和泛化能力等方面差异显著，属于模型设计的核心层面。

本章将从经典的MLP策略出发，介绍近年来在策略表征结构上涌现的一系列重要进展，包括指针网络、序列模型、扩散模型、超网络、模块化设计模型、混合专家模型、基于序列化Token的模型等，展现出结构与任务之间日益增强的适配性与灵活性。

2.1 传统基于MLP的策略

MLP结构是DRL策略最为基本和主要的表征方式，即通过若干全连接层将状态向量$s\in\mathbb{R}^d$映射为动作分布或动作值函数。该策略在离散动作空间上可形式化为

$$\pi_\theta(a|s)=\mathrm{softmax}(f_\theta(s)) \tag{6}$$

在连续动作空间上可形式化为

$$\pi_\theta(a|s)=N(\mu_\theta(s),\mathrm{diag}(\exp(2\ln\sigma_\theta(s)))) \tag{7}$$

式中：f_θ为MLP网络；θ为网络参数，若为离散动作空间，则输出为动作概率分布(经softmax归一化)；若为连续动作空间，则策略通过MLP输出动作分布的参数，通常为高斯分布的均值$\mu_\theta(s)$与对角协方差矩阵的对数方差$\ln\sigma_\theta^2(s)$，从而构成动作的条件分布$N(\mu,\sigma^2)$。

MLP策略由于结构简单、参数少，常被用于如Atari游戏、低维控制任务中，且易于训练。然而，该类表征存在以下局限：

(1) 结构无感知性：MLP无法感知输入中蕴含的结构信息(如图结构、序列依赖等)，对高维或非欧几里得数据的建模能力有限；

(2) 泛化性弱：针对特定任务学习的参数难以迁移至其他任务，限制了其在多任务学习或现实环境中的适应性；

(3) 表达能力受限：难以捕捉长程依赖或策略多样性，导致策略输出趋于单模态或受噪声干扰大。

尽管如此，MLP策略仍是许多RL方法的基准设计，如DQN[20]、DDPG[21]、PPO[22]、SAC[23]等主流算法皆基于此进行扩展，是后续复杂表征设计的重要起点。

2.2 基于指针网络的策略

指针网络最早由Vinyals等[24]提出，旨在处理输出空间大小随输入变化而动态扩展的序列建模问题。最早将指针网络与DRL结合的工作包括文献[25]，通过策略梯度方法训练指针网络解决旅行商问题(traveling salesman problem，TSP)与容量限制的车辆路径问题(capacitated vehicle routing problem，CVRP)等。在此类组合优化问题中，策略需在若干个可能的离散目标之间做出排序或选择，输出并不是固定维度的动作，而是输入元素集合上的“索引指针”。MLP难以直接处理这种动态长度和位置相关的动作空间，而指针网络则通过显式的注意力机制解决了这一问题。指针网络通常由2个核心模块组成：

(1) 编码器：将输入序列$\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$(例如待排序任务、城市坐标等)表示为特征向量序列$\{\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_n\}$，其中$\boldsymbol{h}_i\in\mathbb{R}^d$，为第$i$个输入元素的向量表示。常使用循环神经网络(recurrent neural network，RNN)或者Transformer编码器表示。

$$\boldsymbol{h}_i=E(\boldsymbol{x}_i),\boldsymbol{x}_i\in\mathbb{R}^d \tag{8}$$

式中：E为编码器Encoder。

(2) 解码器与注意力机制：使用RNN或者Transformer解码器逐步生成动作，即指向某个输入元素的索引，每一步通过注意力机制计算“指向概率”：

$$u_i^t=\boldsymbol{v}^{\mathrm{T}}\tanh(W_1\boldsymbol{h}_i+W_2d_t),\alpha^t=\mathrm{softmax}(u^t) \tag{9}$$

式中：$\boldsymbol{v}^{\mathrm{T}}$为矩阵$\boldsymbol{v}$的转置；$d_t$为解码器第$t$步的状态；$\alpha_i^t$为第$t$步骤选择第$i$个输入的概率。最终策略定义为

$$\pi(a_t=i|s_t)=\alpha_i^t \tag{10}$$

即通过注意力机制对输入点进行“指针式”的选择。

此后该范式在多种组合优化任务中得到广泛拓展，例如，采用Transformer架构的指针网络，基于RL实现大规模CVRP的高效求解[26]，将指针网络应用于容量约束路径问题，引入多头注意力和贪婪采样[27]，结合演员–评论者模型和图结构信息，提升多任务调度泛化性能[28]，将指针策略迁移至图着色问题，引入图嵌入以支持约束决策[29]。

指针网络作为策略架构，具备多项独特优势。其最显著的特点是能够灵活适应动态动作空间，特别适用于输出维度随输入变化而调整的任务，如路径规划或排序决策等典型组合优化问题。同时，指针机制天然支持在离散动作空间中进行精确的选择与排序操作，为RL中的选择型策略提供了结构优势。此外，将指针网络与Transformer结构结合，不仅可以提升模型的表达能力，还能显著增强其在多任务场景中的泛化性能。尽管指针网络最初源于监督学习，但其结构与RL任务中的结构性决策高度契合，已成为组合优化类DRL研究中的核心表征架构之一。

2.3 基于序列建模的策略

新兴思路将RL中的策略学习过程转化为条件序列建模任务，即将状态–动作–奖励等决策相关变量表示为一个序列，通过自回归模型预测最优动作。这一思路打破了传统RL中“值函数+策略梯度”的范式，引入了自然语言处理中的Transformer架构，从而形成了以基于决策的Decision Transformer[30]和基于轨迹的Trajectory Transformer[31]为代表的新型策略表征方法。

Decision Transformer是首个将序列建模方法引入RL的工作。该方法将一条轨迹中的期望回报$\hat{R}_t$、状态S_t、动作a_t拼接为一个三元序列，构建形式如下：

$$\tau=(\hat{R}_1,s_1,a_1,\hat{R}_2,s_2,a_2,\cdots,\hat{R}_T,s_T) \tag{11}$$

式中：T为轨迹的终止时间步。策略通过自回归因果注意力网络建模，在每一个时间步预测动作为

$$a_t\sim\pi(a_t|\hat{R}_{\leqslant t},s_{\leqslant t},a_{\leqslant t}) \tag{12}$$

训练目标为最小化预测动作与真实动作之间的重建损失：

$$L_{\mathrm{DT}}=\sum_{t=1}^{T}\left\|a_t-\hat{a}_t\right\| \tag{13}$$

该方法无需显式学习价值函数或使用策略梯度，天然适配大规模离线数据，并支持以目标回报为条件的行为调节。

Trajectory Transformer进一步发展了序列建模在RL中的表达方式。不同于Decision Transformer以回报为条件的行为预测，Trajectory Transformer直接建模整条轨迹的联合概率分布，并将状态、动作、奖励序列离散化处理，适用于基于轨迹规划的方法。Trajectory Transformer的策略学习过程不再生成单步动作，而是在轨迹空间中通过自回归方式生成完整轨迹，在决策精度和可解释性方面具有一定优势，适用于规划控制类任务。Prompt-DT[32]是对Decision Transformer的多任务扩展，借助“任务提示(Prompt)”向量，引导模型在多任务环境下进行条件行为生成。该提示向量可由任务标签、历史轨迹或上下文聚合器提取，用于嵌入序列建模过程，从而提升策略对任务分布变化的泛化能力。

这类基于序列建模的策略结构以统一的建模范式将RL任务抽象为序列预测问题，从而简化了传统方法中对值函数的依赖与设计难度。该类方法充分借鉴自然语言处理中的序列建模经验，天然适配大规模离线数据训练，具有较强的数据利用效率。此外，通过引入目标回报或Prompt等条件控制变量，策略具备良好的任务调节与迁移能

力。其与Transformer架构的高度兼容性也带来了结构上的灵活性与扩展性。然而，尽管该类方法在策略泛化与表达能力方面展现出广阔前景，其训练过程对数据质量较为敏感，稳定性仍有待进一步提升。

2.4 基于扩散模型的策略

扩散模型在图像生成、自然语言建模等领域取得显著进展，其强大的分布建模能力也逐渐被引入到DRL中，用以构建更灵活、更具表现力的策略表征。与传统策略结构常依赖的简单分布(如高斯分布、Softmax输出)不同，扩散策略通过模拟一个逐步“去噪”过程，可以有效建模多模态、复杂形状的动作分布，特别适用于对策略表达能力要求较高且有大量训练数据的离线RL场景。

扩散策略的基本思想是：首先将动作变量 $a \in \mathbb{R}^d$ 扩散为高斯噪声 $\tilde{a}_T \sim N(0,1)$，然后通过一个参数化反扩散过程逐步还原出原始动作。

$$\tilde{a}_T \xrightarrow{\text{denoise}} \tilde{a}_{T-1} \xrightarrow{\cdots} \tilde{a}_0 \approx a \tag{14}$$

该过程通常定义为条件扩散建模，策略形式为

$$\pi_\theta(a|s) := D_\theta(a;s) \tag{15}$$

这种方式不再直接预测动作，而是模拟从高斯噪声到动作的逐步“净化”路径，从而可灵活表达复杂策略分布，尤其在高奖励区域分布稀疏、轨迹多样性高的离线任务中表现优异。

文献[33]将扩散模型引入Q学习框架。该方法以状态为条件训练扩散模型，学习目标由Q函数引导，使生成的动作倾向于高价值区域，有效解决了多模态动作空间下的策略退化问题。文献[34-35]提出在扩散训练中引入优势函数或Q函数作为加权目标，以实现偏向高回报动作的训练。由于扩散采样步数较多，部分研究探索了近似采样机制以加速推理过程。例如文献[36]提出的动作近似方法可在训练期间用少量DDIM步骤替代完整扩散链，有效提升采样速度。近期研究尝试将扩散建模拓展至离散动作空间。文献[37]提出基于梯度的离散扩散过程，用于提升离线策略学习效率。文献[38]进一步构建层次化框架，将离散扩散与离线Q-learning结合，实现有限样本条件下的鲁棒策略学习。

扩散策略能够建模多峰、多模态以及非对称等复杂的动作分布，在策略空间中提供了更丰富的行为多样性。同时，该类方法在高维状态空间、稀疏奖励环境以及典型的离线RL任务中表现出良好的鲁棒性，有效缓解了传统策略在此类场景下易退化的问题。此外，扩散策略具备良好的兼容性，能够灵活地与Q-learning、行为克隆、能量建模等训练机制结合，构建多种灵活的策略优化范式。然而，其主要局限在于训练与采样成本相对较高，尤其依赖较长的扩散步数和对反扩散过程的精确建模。未来，如何在策略泛化能力与训练效率之间实现更优权衡，将是扩散策略表征继续演进的关键方向。

2.5 超网络驱动的策略

超网络作为一种“网络生成网络”，由文献[39]提出，近年来在图像识别、迁移学习等监督学习领域取得广泛应用[40]。尽管其在强化学习中的应用相对较少，但已有研究表明，超网络结构在提升策略表达灵活性和跨任务泛化能力方面具有显著潜力[41]。

超网络主要被用于根据任务特征或上下文信息动态生成策略网络或价值函数的参数，从而在多任务、元学习和零样本迁移等场景中提升适应能力。例如，文献[42]在多智能体强化学习中使用超网络根据智能体属性生成个性化的价值函数。文献[43]进一步拓展了超网络在协作型多智能体强化学习中的策略生成机制。文献[44]使用超网络估计DRL的架构权重，展现了超网络在资源受限情况下的策略生成优势。文献[45]则将超网络应用于连续模型更新任务中，以建模环境动态的演化过程，展示了其在持续学习中的应用潜力。

近期的研究聚焦于超网络驱动的策略迁移与

任务泛化能力。文献[46]利用超网络学习目标条件策略，实现了命令空间上的行为生成，而不仅仅是针对固定奖励目标的策略优化。与之类似，文献[47]提出一种用于零样本任务泛化的超网络方法，通过将一组任务元信息映射为策略权重，直接在未见任务上实现合理行为，无需微调。文献[48]则在多任务学习中引入超网络结构，利用任务嵌入向量生成参数共享的策略网络实例，从而提升任务间知识迁移效率。

超网络在DRL中的策略表征中具备显著优势，尤其体现在可根据任务或上下文动态生成策略参数，提升了多任务适应性与零样本泛化能力，同时减少了参数冗余，适用于多智能体与元学习等场景。然而，其生成过程仍存在训练不稳定、结构不透明等问题，且对任务先验的依赖较强。未来可探索引入结构归纳偏置、构建模块化或层次化超网络，并结合Prompt或扩散机制，推动超网络在通用策略表征中的实用化落地。

2.6 模块化策略

模块化策略表征是一种在DRL中日益受到关注的策略设计方法，其核心思想是将策略网络划分为多个功能模块，每个模块负责处理特定的子任务或控制特定的智能体部分。一般按照按任务结构、机器人结构或智能体功能预定义模块来进行模块划分。这种方法特别适用于处理具有复杂结构的智能体和多智能体系统，能够提高策略的泛化能力和适应性。

在机器人控制领域，模块化策略早已有探索。如文献[49-50]提出将策略条件于机器人的结构编码，以适配可重构或定制的机器人平台。文献[51-52]进一步引入图结构描述智能体形态，使策略可迁移至不同的形态结构中，并通过图上的消息传递机制进行动作决策，实现了策略在结构维度上的零样本泛化。文献[53]提出的共享模块化策略架构，利用相同的模块化神经网络控制不同形态的智能体，实现了在未见过的智能体结构上的零样本泛化能力。此外，文献[54]提出了一种基于模块化图结构的策略学习方法，通过将机器人的结构表示为图，并在图上应用策略学习，实现了对多种机器人设计的控制。

尽管模块化策略结构具备良好的参数复用性、结构适配性，但仍面临图结构设计依赖强、训练开销高、推理效率受限等问题[55]。未来的工作可从自动化神经结构搜索、模块可组合性增强、跨形态策略重用机制设计等角度进一步优化模块化表征策略的表达能力与部署效率。

2.7 混合专家策略

混合专家模型近年来逐渐成为DRL中重要的策略表征手段，尤其在面对多模态策略分布、状态空间非平稳性及多任务控制等复杂情境时，展现出强大的建模能力。MoE策略架构通过引入多个功能专家子网络，由路由机制或门控网络根据当前状态或上下文信息动态分配专家权重，实现多策略融合或策略选择，增强了策略的适应性与表达力[56]。

在早期研究中，MoE多用于将状态空间分区，每个专家学习特定子区域的策略，以应对高维状态下的局部最优控制问题，如多模型强化学习模型[57]和模块化奖励框架[58]。随后的一些研究进一步提出了基于专家结构的值分解机制[59]，支持以模块化方式独立训练子策略并合成全局策略。更复杂的策略结构如文献[60-61]，在actor-critic框架中构建多个专家对，分别处理不同任务或状态区域，显著提升了训练效率与动作质量。

MoE策略，正进一步发展为可组合、多激活、上下文驱动的结构形式。例如，使用高斯混合模型表示多峰策略(PMOE)，有效提升了非单峰环境下的策略稳定性与表达力[62]；而双技能学习和稀疏混合专家系统等方法则通过上下文条件激活专家，使策略行为随环境变化而灵活调整[63-64]。在迁移学习场景中，MoE也被用于构建多任务策略转移框架，如通过状态依赖的狄利克雷先验实现源

任务之间的知识组合[62]，通过引入课程式的强化学习框架获取多样技能并提升策略适应性[65]，或采用多专家联合激活机制融合多种行为模式[66]。此外，MoE 在策略表征中还被广泛结合 Transformer 编码器[67]与贝叶斯推断等技术，推动了对复杂行为的上下文建模与不确定性表达，并已经在多种任务中展现出通用策略能力，例如在四足机器人控制中结合 MoE 和 Transformer 实现通用模型架构[68]。

尽管 MoE 显著增强了策略网络的结构表达能力，但在实际部署中仍面临训练不稳定、专家退化、激活偏差及推理开销较大等挑战[63,68]。未来的研究可进一步探索任务结构感知的门控机制、专家结构归纳偏置、以及高效的前 k(Top-k)激活或稀疏路由策略，提升 MoE 策略在大规模、多模态任务下的泛化能力和训练效率。总的来看，MoE 策略表征机制正在促使强化学习策略从“统一型策略映射”向“模块化行为系统”不断演化，尤其适用于面对多样任务、复杂状态动态与策略多解性的实际决策场景[56,62,65]。

2.8 基于序列化 Token 的跨模态策略

随着多模态学习与序列建模方法的发展，基于序列化 Token 的跨模态统一策略架构逐渐成为 RL 策略表征领域的新兴研究方向。这类架构的核心思想是将视觉图像、语言文本、机器人控制信号等异构模态信息统一离散化为标准化的 Token 序列，通过共享的序列模型实现统一编码与策略学习，从而在不同任务与环境间实现策略的高效泛化与适配能力。

在统一序列建模方面，Gato 模型[69]将图像、文本与控制信号离散化为 Token，通过 Transformer 进行统一建模，开启了“以序列为中心”的通用智能体范式，后续如 PaLM-E[70]和 RT-2[71]等进一步借助大型视觉语言模型实现了从网络知识到现实控制的迁移能力。围绕世界模型对齐，GenRL[72]提出通过语言与视觉提示驱动的生成式潜在世界建模，使得智能体能在想象中学习和泛化复杂任务。针对多步推理与任务结构建模，InfiGUIAgent[73]引入视觉语言感知与层级反思机制，有效应对图形界面中长程依赖与多层指令的问题。在数据效率与快速适应方面，RoboCat[74]利用自我收集与少样本更新机制，在不同平台和任务中实现了高效迁移。VIMA[75]则通过统一 prompt 模态展现了强泛化能力。在多感官融合方向，FuSe[76]利用语言作为中介桥梁，将视觉策略迁移到触觉与听觉模态，有效提升了机器人在异构传感条件下的适应性与成功率。最后，在轻量化部署与快速语境适应方面，REGENT[77]引入检索增强机制，在推理过程中动态引入语境提示，使通用策略网络在陌生环境中无需微调即可执行复杂任务。上述 6 条路径共同推动了通用策略模型在架构统一性、模态融合能力、数据效率与泛化能力等方面的协同突破，为构建多模态、高适应性的智能体系统提供了坚实的基础。

通用策略大模型在结构统一性、模态融合、多任务泛化方面取得了一系列突破，为强化学习策略表征提供了新的建模范式与研究视角。

上述各类策略表征架构在输入结构、动作输出形式及使用场景上各具特点，表 1 对本章介绍的主要策略表征架构及特性进行了系统梳理和对比。

表 1　不同策略表征架构的对比
Table1　Comparison of Different Policy Representation Architectures

策略架构类型	输入结构	动作输出形式	典型适用场景
基于 MLP 的策略架构	状态向量(可拼接任务标签)	Softmax 概率/高斯分布函数	简单任务、连续/离散控制问题
基于指针网络的策略架构	状态特征+动态元素序列(如城市、任务点)	指向输入元素的注意力索引	排序、路径规划、组合优化问题(TSP/ CVRP)
基于序列建模的策略架构	回报+状态+动作组成的轨迹序列(可附加 Prompt)	自回归生成下一个动作	离线 RL、条件策略建模、多任务控制
基于扩散过程的策略架构	状态特征+随机噪声向量+(可选)目标/奖励引导	反扩散采样生成的高维动作	多模态策略生成、Offline RL、复杂控制
超网络驱动的策略架构	状态嵌入+任务/上下文向量	由超网络生成的策略网络参数	多任务迁移、协作控制、零样本泛化
基于模块化结构的策略架构	局部状态+图结构连接信息(节点/边)	各模块局部策略输出的联合动作	多智能体系统、结构可变机器人
基于混合专家的策略架构	状态向量+上下文向量(可为任务或环境提示)	多专家动作的门控融合输出	多模态任务、非平稳策略集成、长期控制
基于序列化 Token 的策略架构	多模态信息统一为 Token 序列	序列模型自回归生成动作	跨模态任务、多任务泛化、统一策略学习与部署

3　策略表征方法简述

区别于上一章聚焦于策略网络的结构设计形式，本节所讨论的策略表征方法，更侧重于策略输入与中间表示的语义编码方式，即研究“策略如何理解状态、任务与上下文”的问题。策略表征方法关注的核心在于：策略在决策过程中所依赖的关键信息(例如任务嵌入、对比特征、目标变量或结构归纳)应如何在训练阶段被合理抽象、组织与优化，从而提升策略的泛化能力、样本效率与稳定性。这类方法往往用于元学习、离线 RL、多任务控制等复杂场景，作为提升策略能力的重要中间机制。

需要指出的是，本章所介绍的策略表征方法，并非 RL 中普遍适用或必不可少的组成部分。在一些单一任务、低维状态或结构简单的环境中，依赖奖励信号驱动、配合简单策略网络的常规方法，往往已足以完成决策任务。然而，随着 RL 在现实系统中的应用不断拓展，面对任务分布广泛、结构多变、数据稀疏等挑战，如何使策略具备更强的跨任务迁移能力、自适应行为生成能力以及对复杂语义结构的理解能力，已成为当前研究的核心目标。本文所综述的各类策略表征方法，正是围绕这一趋势提出的重要探索路径，代表了 RL 从“单任务对策”走向“结构感知与语义泛化”的关键演进方向。

本章选取了 4 类主要和典型的策略表征方法，分别反映了近年来研究者在任务建模、自监督表征、结构建图与行为抽象等维度上的探索。需要特别指出的是，本章所归纳的几类表征方式包括上下文建模、语义对比、结构归纳与技能/目标嵌入并非彼此对立的孤立方法。在实践中，它们常常可以相互结合、协同增益。因此，本节对 4 类方法的划分，旨在体现其各自的研究出发点与建模重点，而非形成严格的分类边界。

3.1　任务上下文建模

任务上下文建模，旨在通过显式或隐式方式将任务相关信息编码为低维潜在向量，辅助策略网络生成更具适应性的行为输出，从而提升在多任务、分布外环境中的泛化能力。这类方法已广泛应用于 Meta-RL 与情境强化学习(contextual RL，

cRL)2个子方向。

在Meta-RL中，策略学习的目标不再是直接拟合某个任务的最优行为，而是学习一种“如何快速适应新任务”的能力。早期研究如RL^2[78]将策略与递归结构(如RNN)结合，将任务内的历史轨迹编码为隐含状态，用于条件化当前策略，避免测试时梯度更新。MAML[79]则提出通过显式梯度更新优化模型初始参数，提升少步适应能力。基于上下文的无模型方法如PEARL[80]利用变分推断模块从历史轨迹中推断任务嵌入向量，并作为条件输入传递给策略网络。其后续研究进一步融合价值函数[81]、模型动力学[82]和能量建模[83]等结构，提升任务泛化的表征能力[84]。

与之相对，情境强化学习假设环境变化可以由某种可观测的上下文变量显式编码，例如物理参数(风速、地形、摩擦系数)或任务语义标签。早期研究包括文献[85-86]，后续如文献[87]基于文献[88]提出的评估协议，系统性地验证了不同RL算法在物理属性变化下的泛化能力。这些方法一般将上下文信息拼接进状态向量中，通过训练统一策略网络来适应不同环境的设定。进一步的研究也探索使用超网络根据上下文动态生成策略参数，或在不可观测场景下采用元学习结构间接推断上下文[89]。

尽管Meta-RL与cRL在建模方式上存在差异，前者多依赖历史轨迹推理，后者多采用显式输入拼接，但两者本质上都在构建任务或环境的语义嵌入，以作为策略条件化的控制信号。这类策略表征机制在当前RL中扮演着连接任务推理与策略生成之间的关键中间层，已成为提升RL泛化能力的重要研究方向。未来，如何设计具备更强任务建模能力、能够统一处理显式与隐式上下文信息的策略表征方法，仍是未来值得深入探索的研究方向。

3.2 自监督表征学习

自监督表征学习已成为强化学习中的重要研究趋势。相比于传统方法仅依赖环境奖励来学习状态表示，这些方法在策略模型训练中引入额外的对比或预测损失来提取更加判别的特征。通过这种方式，网络能够学习到聚焦任务关键因素的中间表示，从而加快策略学习并提高泛化能力。

例如，文献[90]提出的CURL方法使用实例级对比学习训练视觉编码器，确保同一观测的不同增强视图在特征空间中匹配。文献[91]的SPR方法通过多步预测未来状态的潜在表示并结合数据增强来实现时序一致性，这种自我预测与视图一致性目标帮助学习到更鲁棒的状态特征。文献[92]通过对比不同任务的上下文嵌入的自监督学习来训练紧凑的任务表示，并设计基于信息增益的探索策略以收集具有判别力的轨迹。也有工作将对比学习与时序或目标结构相结合，如文献[93]的KSL方法通过动作条件化的潜表示预测任务强化了表示的时间平滑性，文献[94]的TCL方法通过对比同一任务轨迹下的状态窗口为Meta-RL中的上下文编码器提供强监督。

自监督表征学习方法在强化学习中发挥了重要作用。通过各种对比或预测目标来塑造状态表示，这些方法普遍提高了策略学习的样本效率和稳定性，为视觉控制、元学习等多种RL范式带来了性能增益[95]。

3.3 结构关系归纳

基于结构关系归纳的策略表征方法指的是在策略建模过程中显式引入任务或系统中已有的结构先验，例如对象之间的交互、模块间的拓扑、智能体之间的通信等，并以此为基础构建具有结构归纳偏置的策略网络。这些方法通常不将结构信息仅作为输入特征或感知模块(例如图结构信息作为状态特征)，而是将其直接融入策略网络的架构中，例如利用图神经网络在策略网络内传播信息、按功能/部件划分策略模块并通过消息传递或门控进行交互，以及构建可组合的子策略以支持策略的迁移和复用等[53,96]。通过这些结构化的策略

表征，模型可以有效地共享参数、显式捕获对称性，进而提升策略在未见结构或新任务中的泛化和零样本迁移能力。整体而言，这类方法普遍适用于多智能体协作、机器人运动控制、物理系统和组合决策等多种场景。

典型工作如共享模块化策略[53]，将策略网络划分为若干相同结构的子模块，每个模块仅感知局部状态，通过模块间的消息传递实现集中式或分布式的行为协调，从而支持策略在不同形态机器人间的迁移与复用。类似思想在 NerveNet 等[52]图神经网络策略中也得到体现，该方法将机器人的部件建模为图中的节点，通过图卷积学习节点间的交互与协同动作，从而实现对可变结构和损坏情形下的稳健控制。在多智能体场景中，MAGEC 等方法[97]利用图结构对智能体集群建模，结合策略梯度算法实现了对复杂通信关系下的全局协调策略学习。除了模块划分与图建模外，模块可组合性也是结构归纳的重要方向，例如某些元学习方法可自动提取多任务环境中的基础策略模块，并在新任务中进行组合重构，支持组合式迁移与结构泛化[96]。在面向对象的任务建模中，文献[98]引入了对象-关系图表示，将多个实体之间的交互建模为可学习的结构，进一步借助消息传递机制强化策略对场景关系的感知与推理能力。

总体来看，这类方法通过在策略表征中显式引入结构归纳偏置，使得策略能够在多样化的系统结构、任务目标与协同场景中实现更强的泛化与可复用性，正在成为构建可扩展通用智能体的重要方向。未来，随着图神经网络和可组合架构的发展，这些方法有望在更复杂的多模态和物理交互场景中发挥更大作用，为强化学习策略的可解释性、可迁移性和鲁棒性提供坚实基础。

3.4 技能/目标嵌入

技能/目标嵌入通常在分层强化学习和目标条件强化学习中使用，其核心思想是在策略中引入一个表示技能或目标的潜在向量，使策略能够通过调整该向量生成不同的行为模式。技能和目标嵌入策略已成为提升策略泛化性、样本效率和可解释性的关键研究方向。这些方法通过引入技能或目标的嵌入表示，使策略能够在多任务和复杂环境中更有效地学习和适应。

在强化学习中，技能(Skills)通常指的是可以重复调用的行为子策略或动作片段，用于执行某种“亚目标”或“子任务”。技能嵌入指用低维连续向量或离散符号对技能进行编码，使策略网络能够通过该嵌入控制或选择技能执行行为。一类方法侧重于从离线数据中提取可重用的技能集合[99]。例如，文献[100]提出的可解释分层策略模型，利用语言描述关联技能模块，实现技能的可组合性与可解释性。另一类则通过手动构建技能集[101]，如文献[102]提出的 Plan4MC 框架，将复杂任务分解为查找、操作、合成等基础技能，并构建技能图以支持计划与重用。但上述这些技能通常以高维连续向量(如高斯嵌入)表示，尽管表达力强，但存在可解释性差与训练不稳定的问题。因此最新的研究文献[103]提出 SkillTree 框架，试图通过将复杂的连续动作空间映射为离散的技能空间，并结合可微分的决策树结构，实现了策略的可解释性和高效性。

目标嵌入是将任务目标转化为潜在向量表示，作为策略或值函数的条件输入，从而实现对不同目标任务的快速适应与行为调节。早期研究主要采用目标状态拼接的方式，如文献[104]通过将目标状态与当前状态合并作为输入，引导策略逼近任意目标点。然而，这类方法只能处理简单的导航式目标，缺乏对复杂任务结构的表达能力。随着自然语言处理技术的发展，研究者尝试将语言指令作为目标输入，并通过循环网络或预训练模型(如 GRU、BERT)进行编码，实现了目标的语义嵌入[105]。这类方法拓展了目标的表达空间，但在泛化性与语义一致性方面仍面临挑战。近年来，研究开始聚焦于将目标表示为可控潜在变量，即从离线数据或交互中自动学习一组抽象技能或目

标表示，用于调节策略行为。这类方法不再依赖显式目标状态，而是构建目标驱动的行为调控空间，具有更强的行为多样性与可组合性。典型方法包括DR-GRL[106]、ASE[107]等，它们分别从视觉目标控制、物理角色控制等不同角度提升了策略在多任务环境下的适应能力。在此基础上，最新的研究引入了逻辑与结构化目标建模，如将目标表示为线性时序逻辑(LTL)[108]、组合确定性有限自动机(cDFA)[109]或时间逻辑因果图(TL-CD)[110]，并结合图神经网络对其进行结构编码。这类方法不仅具备更强的表达能力，还能支持复杂目标之间的推理、组合与泛化，推动了目标嵌入从低维条件向结构语义建模的演进。

总之，技能/目标嵌入方法将策略与技能或目标解耦，通过潜在变量为复杂任务提供了灵活可控的表示，在提升探索效率和任务可组合性方面发挥了重要作用。未来可将技能/目标嵌入与多模态感知(语言、视觉、符号)进一步深度融合，使嵌入表示既具抽象能力又具交互能力，推动其在多任务迁移、指令执行和长期规划中的广泛落地。

4 结论

随着DRL在机器人控制、复杂系统优化及人机交互等场景中的应用日益深化，策略表征作为连接“感知–推理–决策”的关键中间层，其结构设计与语义表达能力正逐步成为制约强化学习性能上限的核心因素。本文系统梳理了近年来DRL中策略表征领域的关键研究进展，全面总结了策略表征方法中的关键类型，如任务建模、语义对比、结构建图与行为抽象等，并分析了当前策略表征研究所面临的主要挑战，提出了未来可能的发展方向，期望为从事DRL算法研究与实际应用的相关学者和工程人员提供参考与启示。

从建模仿真视角来看，深度强化学习中的策略表征技术可以视为一种面向复杂系统的“智能控制策略生成器”，其核心目标是通过状态—动作映射建模智能体行为，从而支撑面向任务目标的模型驱动仿真。无论是在多智能体交互建模、柔性制造仿真，还是自主系统仿真中，策略表征方法均可用于提升模型行为的泛化能力与适应能力。因而，本研究对于推动智能建模范式的发展、提升建模仿真系统的智能决策能力具有积极意义。

同时，在相应架构与方法的后面针对单个技术点进行了技术展望。除此以外，当前策略表征研究依然存在“碎片化建模、依赖任务工程、缺乏统一表征框架”等问题，制约了策略的通用性与跨场景适配能力。因此，未来策略表征的发展亟需构建系统性的策略表征方法论与标准化表征体系。未来可探索基于语义对齐的潜在表示空间，融合上下文、目标与环境状态，构建跨任务一致的表示结构，从而增强泛化性和策略调控一致性。

参考文献：

[1] 刘朝阳, 穆朝絜, 孙长银. 深度强化学习算法与应用研究现状综述[J]. 智能科学与技术学报, 2020, 2(4): 312-326.
Liu Zhaoyang, Mu Zhaoxu, Sun Changyin. An Overview on Algorithms and Applications of Deep Reinforcement Learning[J]. Chinese Journal of Intelligent Science and Technology, 2020, 2(4): 312-326.

[2] 李静, 丁佳文, 沈南燕, 等. 基于深度强化学习的双足机器人行走策略研究[J]. 机器人技术与应用, 2025(3): 44-49.

[3] Li Minne, Wu Lisheng, Wang Jun, et al. Multi-view Reinforcement Learning[C]//Advances in Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2019: 1-12.

[4] Silver D, Hubert T, Schrittwieser J, et al. A General Reinforcement Learning Algorithm That Masters Chess, Shogi, and Go Through Self-play[J]. Science, 2018, 362 (6419): 1140-1144.

[5] Vinyals O, Babuschkin I, Czarnecki W M, et al. Grandmaster Level in StarCraft II Using Multi-agent Reinforcement Learning[J]. Nature, 2019, 575(7782): 350-354.

[6] Stiennon N, Ouyang Long, Wu J, et al. Learning to Summarize from Human Feedback[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2020: 3008-3021.

[7] Cobbe K, Klimov O, Hesse C, et al. Quantifying

Generalization in Reinforcement Learning[C]// Proceedings of the 36th International Conference on Machine Learning. Chia Laguna Resort: PMLR, 2019: 1282-1289

[8] Mazoure B, Doan T, Li Tianyu, et al. Low-rank Representation of Reinforcement Learning Policies[J]. Journal of Artificial Intelligence Research, 2022, 75: 597-636.

[9] Ofir Nabati, Guy Tennenholtz, Shie Mannor. Representation-driven Reinforcement Learning[C]// Proceedings of the 40th International Conference on Machine Learning. Chia Laguna Resort: PMLR, 2023: 25588-25603.

[10] Yang Long, Huang Zhixiong, Lei Fenghao, et al. Policy Representation via Diffusion Probability Model for Reinforcement Learning[EB/OL]. (2023-05-22) [2025-06-01]. https://arxiv.org/abs/2305.13122.

[11] Levine S, Pastor P, Krizhevsky A, et al. Learning Hand-eye Coordination for Robotic Grasping with Deep Learning and Large-scale Data Collection[J]. The International Journal of Robotics Research, 2018, 37(4/5): 421-436.

[12] Jared Di Carlo, Wensing P M, Katz B, et al. Dynamic Locomotion in the MIT Cheetah 3 Through Convex Model-predictive Control[C]//2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Piscataway: IEEE, 2018: 1-9.

[13] Mock J W, Muknahallipatna S S. Sim-to-real: A Performance Comparison of PPO, TD3, and SAC Reinforcement Learning Algorithms for Quadruped Walking Gait Generation[J]. Journal of Intelligent Learning Systems and Applications, 2024, 16(2): 23-43.

[14] Kaufmann T, Weng P, Bengs V, et al. A Survey of Reinforcement Learning from Human Feedback[EB/OL]. (2024-04-30) [2025-06-01]. https://arxiv. org/abs/2312.14925.

[15] Welsby P, Cheung B M Y. ChatGPT[J]. Postgraduate Medical Journal, 2023, 99(1176): 1047-1048.

[16] Bai Yuntao, Kadavath S, Kundu S, et al. Constitutional AI: Harmlessness from AI Feedback[EB/OL]. (2022-12-15) [2025-06-01]. https://arxiv.org/abs/2212.08073.

[17] Hessel M, Soyer H, Espeholt L, et al. Multi-task Deep Reinforcement Learning with PopArt[C]//Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence. Palo Alto: AAAI Press, 2019: 3796-3803.

[18] Fakoor R, Chaudhari P, Soatto S, et al. Meta-Q-Learning [EB/OL]. (2020-04-04) [2025-06-01]. https://arxiv. org/abs/1910.00125.

[19] Liu Jinxin, Wang Donglin, Tian Qiangxing, et al. Learn Goal-conditioned Policy with Intrinsic Motivation for Deep Reinforcement Learning[C]//Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence and Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence and the Twelveth Symposium on Educational Advances in Artificial Intelligence. Palo Alto, CA, USA: AAAI Press, 2022: 7558-7566.

[20] Mnih V, Kavukcuoglu K, Silver D, et al. Human-level Control Through Deep Reinforcement Learning[J]. Nature, 2015, 518(7540): 529-533.

[21] Lillicrap P T, Hunt J J, Pritzel A, et al. Continuous Control with Deep Reinforcement Learning[EB/OL]. (2019-07-05) [2025-06-01]. https://arxiv. org/abs/1509.02971.

[22] Schulman J, Wolski F, Dhariwal P, et al. Proximal Policy Optimization Algorithms[EB/OL]. (2017-08-28) [2025-06-01]. https://arxiv.org/abs/1707.06347.

[23] 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 35th International Conference on Machine Learning. Chia Laguna Resort: PMLR, 2018: 1861-1870.

[24] Vinyals O, Fortunato M, Jaitly N. Pointer Networks[C]// Advances in Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2015: 1-9.

[25] Bello I, Pham H, Le Q V, et al. Neural Combinatorial Optimization with Reinforcement Learning[EB/OL]. (2017-01-12) [2025-06-01]. https://arxiv. org/abs/1611.09940.

[26] Kool W, Herke van Hoof, Welling M. Attention, Learn to Solve Routing Problems! [EB/OL]. (2019-02-07) [2025-06-01]. https://arxiv.org/abs/1803.08475.

[27] Nazari M, Oroojlooy A, Snyder L, et al. Reinforcement Learning for Solving the Vehicle Routing Problem[C]// Advances in Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2018: 1-11.

[28] Sudhakar R V, Dastagiraiah C, Pattem S, et al. Multi-objective Reinforcement Learning Based Algorithm for Dynamic Workflow Scheduling in Cloud Computing[J]. Indonesian Journal of Electrical Engineering and Informatics, 2024, 12(3): 640-649.

[29] Li Wei, Li Ruxuan, Ma Yuzhe, et al. Rethinking Graph Neural Networks for the Graph Coloring Problem[EB/OL]. (2022-08-19) [2025-06-01]. https://arxiv. org/abs/2208.06975.

[30] Chen Lili, Lu K, Rajeswaran A, et al. Decision Transformer: Reinforcement Learning via Sequence Modeling[C]//Advances in Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2021: 15084-15097.

[31] Janner M, Li Qiyang, Levine S. Offline Reinforcement Learning as One Big Sequence Modeling Problem[C]// Advances in Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2021: 1273-1286.

[32] Xu Mengdi, Shen Yikang, Zhang Shun, et al. Prompting Decision Transformer for Few-shot Policy Generalization [C]//Proceedings of the 39th International Conference on Machine Learning. Chia Laguna Resort: PMLR, 2022: 24631-24645.

[33] Wang Zhendong, Hunt J J, Zhou Mingyuan. Diffusion Policies as an Expressive Policy Class for Offline Reinforcement Learning[EB/OL]. (2023-08-25) [2025-06-01]. https://arxiv.org/abs/2208.06193.

[34] Chen Huayu, Lu Cheng, Ying Chengyang, et al. Offline Reinforcement Learning via High-fidelity Generative Behavior Modeling[EB/OL]. (2023-02-28) [2025-06-01]. https://arxiv.org/abs/2209.14548.

[35] Lu Cheng, Chen Huayu, Chen Jianfei, et al. Contrastive Energy Prediction for Exact Energy-guided Diffusion Sampling in Offline Reinforcement Learning[C]// Proceedings of the 40th International Conference on Machine Learning. Chia Laguna Resort: PMLR, 2023: 22825-22855.

[36] Kang Bingyi, Ma Xiao, Du Chao, et al. Efficient Diffusion Policies for Offline Reinforcement Learning [C]//Proceedings of the 37th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2023: 67195-67212.

[37] Coleman M, Russakovsky O, Allen-Blanchette C, et al. Discrete Diffusion Reward Guidance Methods for Offline Reinforcement Learning[C]//ICML 2023 Workshop: Sampling and Optimization in Discrete Space. San Diego: ICML, 2023: 1-9.

[38] Qiao Ruixi, Cheng Jie, Dai Xingyuan, et al. Offline Reinforcement Learning with Discrete Diffusion Skills [EB/OL]. (2025-03-26) [2025-06-01]. https://arxiv. org/abs/2503.20176.

[39] Ha D, Dai A, Le Q V. Hypernetworks[EB/OL]. (2016-12-01) [2025-06-01]. https://arxiv.org/abs/1609.09106.

[40] Johannes von Oswald, Henning C, Grewe B F, et al. Continual Learning with Hypernetworks[EB/OL]. (2022-04-11) [2025-06-01]. https://arxiv.org/abs/1906.00695.

[41] Zhao D, Kobayashi S, João Sacramento, et al. Meta-learning Via Hypernetworks[C]//4th Workshop on Meta-Learning at NeurIPS 2020 (MetaLearn 2020). Piscataway: IEEE, 2020: 1-8.

[42] Rashid T, Samvelyan M, Christian Schroeder De Witt, et al. Monotonic Value Function Factorisation for Deep Multi-agent Reinforcement Learning[J]. The Journal of Machine Learning Research, 2020, 21(1): 7234-7284.

[43] Iqbal S, Christian A Schroeder De Witt, Peng Bei, et al. Randomized Entity-wise Factorization for Multi-agent Reinforcement Learning[C]//Proceedings of the 38th International Conference on Machine Learning. Chia Laguna Resort: PMLR, 2021: 4596-4606.

[44] Hegde S, Huang Zhehui, Sukhatme G S. HyperPPO: A Scalable Method for Finding Small Policies for Robotic Control[C]//2024 IEEE International Conference on Robotics and Automation (ICRA). Piscataway: IEEE, 2024: 10821-10828.

[45] Huang Yizhou, Xie K, Bharadhwaj H, et al. Continual Model-Based Reinforcement Learning with Hypernetworks[C]//2021 IEEE International Conference on Robotics and Automation (ICRA). Piscataway: IEEE, 2021: 799-805.

[46] Francesco Faccio, Vincent Herrmann, Aditya Ramesh, et al. Goal-conditioned Generators of Deep Policies[C]// Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence. Palo Alto: AAAI Press, 2023: 7503-7511.

[47] Sahand Rezaei-Shoshtari, Charlotte Morissette, Francois R Hogan, et al. Hypernetworks for Zero-shot Transfer in Reinforcement Learning[C]//Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence. Palo Alto: AAAI Press, 2023: 9579-9587.

[48] Graffeuille O, Koh Y S, Jörg Wicker, et al. Multi-task Learning with Hypernetworks and Task Metadata[C]// ICLR 2024 Conference. New York: ICLR, 2024: 1-18.

[49] Chen Tao, Murali A, Gupta A. Hardware Conditioned Policies for Multi-robot Transfer Learning[C]// Proceedings of the 32nd International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2018: 9355-9366.

[50] Schaff C, Yunis D, Chakrabarti A, et al. Jointly Learning to Construct and Control Agents Using Deep Reinforcement Learning[C]//2019 International Conference on Robotics and Automation (ICRA).

Piscataway: IEEE, 2019: 9798-9805.

[51] Pathak D, Lu C, Darrell T, et al. Learning to Control Self-assembling Morphologies: A Study of Generalization Via Modularity[C]//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2019: 2295-2305.

[52] Wang Tingwu, Liao Renjie, Ba J, et al. NerveNet: Learning Structured Policy with Graph Neural Networks [C]//ICLR 2018 Conference. New York: ICLR, 2018: 1-26.

[53] Huang Wenlong, Mordatch I, Pathak D. One Policy to Control Them All: Shared Modular Policies for Agent-agnostic Control[C]//Proceedings of the 37th International Conference on Machine Learning. Chia Laguna Resort: PMLR, 2020: 4455-4464.

[54] Whitman J, Travers M, Choset H. Learning Modular Robot Control Policies[J]. IEEE Transactions on Robotics, 2023, 39(5): 4095-4113.

[55] Nurbek G. Exploring Graph Neural Networks in Reinforcement Learning: A Comparative Study on Architectures for Locomotion Tasks[D]. Edinburg: The University of Texas Rio Grande Valley, 2024.

[56] Ren Jie, Li Yewen, Ding Zihan, et al. Probabilistic Mixture-of-experts for Efficient Deep Reinforcement Learning[EB/OL]. (2021-04-19) [2025-06-01]. https://arxiv.org/abs/2104.09122.

[57] Kenji Doya, Kazuyuki Samejima, Ken-ichi Katagiri, et al. Multiple Model-based Reinforcement Learning[J]. Neural Computation, 2002, 14(6): 1347-1369.

[58] Kazuyuki Samejima, Kenji Doya, Mitsuo Kawato. Inter-module Credit Assignment in Modular Reinforcement Learning[J]. Neural Networks, 2003, 16(7): 985-994.

[59] Harm van Seijen, Bram Bakker, Leon Kester. Switching Between Different State Representations in Reinforcement Learning[C]//Proceedings of the 26th IASTED International Conference on Artificial Intelligence and Applications. USA: ACTA Press, 2008: 226-231.

[60] Peng Xuecin, Berseth G, Michiel van de Panne. Terrain-adaptive Locomotion Skills Using Deep Reinforcement Learning[J]. ACM Transactions on Graphics, 2016, 35 (4): 81.

[61] Paolo Tommasino, Daniele Caligiore, Marco Mirolli, et al. A Reinforcement Learning Architecture That Transfers Knowledge Between Skills When Solving Multiple Tasks[J]. IEEE Transactions on Cognitive and Developmental Systems, 2019, 11(2): 292-317.

[62] Gimelfarb M, Sanner S, Lee C G. Contextual Policy Transfer in Reinforcement Learning Domains via Deep Mixtures-of-experts[C]//Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence. Chia Laguna Resort: PMLR, 2021: 1787-1797.

[63] Willi T, Obando-Ceron J, Foerster J, et al. Mixture of Experts in a Mixture of RL Settings[EB/OL]. (2024-06-26) [2025-06-01]. https://arxiv.org/abs/2406.18420.

[64] Mátyás Vincze, Laura Ferrarotti, Leonardo Lucio Custode, et al. SMoSE: Sparse Mixture of Shallow Experts for Interpretable Reinforcement Learning in Continuous Control Tasks[C]//Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence. Palo Alto, CA, USA: AAAI Press, 2025: 20982-20990.

[65] Celik O, Taranovic A, Neumann G. Acquiring Diverse Skills Using Curriculum Reinforcement Learning with Mixture of Experts[EB/OL]. (2024-06-10) [2025-06-01]. https://arxiv.org/abs/2403.06966.

[66] Peng Xuebin, Chang M, Zhang G, et al. MCP: Learning Composable Hierarchical Control with Multiplicative Compositional Policies[C]//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2019: 3686-3697.

[67] Obando-Ceron J, Sokar G, Willi T, et al. Mixtures of Experts Unlock Parameter Scaling for Deep RL[EB/OL]. (2024-06-26) [2025-06-01]. https://arxiv. org/abs/2402.08609.

[68] Song Wenxuan, Zhao Han, Ding Pengxiang, et al. GeRM: A Generalist Robotic Model with Mixture-of-experts for Quadruped Robot[C]//2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Piscataway: IEEE, 2024: 11879-11886.

[69] Reed S, Zolna K, Parisotto E, et al. A Generalist Agent [EB/OL]. (2022-11-11) [2025-06-01]. https://arxiv. org/abs/2205.06175.

[70] Driess D, Xia Fei, Sajjadi Mehdi S M, et al. PaLM-E: An Embodied Multimodal Language Model[EB/OL]. (2023-03-06) [2025-06-01]. https://arxiv.org/abs/2303.03378.

[71] Brohan A, Brown N, Carbajal J, et al. RT-2: Vision-language-action Models Transfer Web Knowledge to Robotic Control[EB/OL]. (2023-07-28) [2025-06-01]. https://arxiv.org/abs/2307.15818.

[72] Mazzaglia P, Verbelen T, Dhoedt B, et al. GenRL: Multimodal-foundation World Models for Generalization in Embodied Agents[EB/OL]. (2024-10-30) [2025-06-01]. https://arxiv.org/abs/2406.18043.

[73] Liu Yuhang, Li Pengxiang, Wei Zishu, et al. InfiGUIAgent: A Multimodal Generalist GUI Agent with Native Reasoning and Reflection[EB/OL]. (2025-01-08) [2025-06-01]. https://arxiv.org/abs/2501.04575.

[74] Bousmalis K, Vezzani G, Rao D, et al. RoboCat: A Self-improving Generalist Agent for Robotic Manipulation [EB/OL]. (2023-12-22) [2025-06-01]. https://arxiv. org/abs/2306.11706.

[75] Jiang Yunfan, Gupta A, Zhang Zichen, et al. VIMA: General Robot Manipulation with Multimodal Prompts [EB/OL]. (2023-05-28) [2025-06-01]. https://arxiv. org/abs/2210.03094.

[76] Jones J, Mees O, Sferrazza C, et al. Beyond Sight: Finetuning Generalist Robot Policies with Heterogeneous Sensors via Language Grounding[EB/OL]. (2025-01-14) [2025-06-01]. https://arxiv.org/abs/2501.04693.

[77] Sridhar S, Dutta S, Jayaraman D, et al. REGENT: A Retrieval-augmented Generalist Agent That Can Act In-context in New Environments[EB/OL]. (2025-02-24) [2025-06-01]. https://arxiv.org/abs/2412.04759.

[78] Duan Yan, Schulman J, Chen Xi, et al. RL^2: Fast Reinforcement Learning via Slow Reinforcement Learning[EB/OL]. (2016-11-10) [2025-06-01]. https://arxiv.org/abs/1611.02779.

[79] Finn C, Abbeel P, Levine S. Model-agnostic Meta-learning for Fast Adaptation of Deep Networks[C]// Proceedings of the 34th International Conference on Machine Learning. Chia Laguna Resort: PMLR, 2017: 1126-1135.

[80] Rakelly K, Zhou A, Finn C, et al. Efficient Off-policy Meta-reinforcement Learning via Probabilistic Context Variables[C]//Proceedings of the 36th International Conference on Machine Learning. Chia Laguna Resort: PMLR, 2019: 5331-5340.

[81] Lee K, Seo Y, Lee S, et al. Context-aware Dynamics Model for Generalization in Model-based Reinforcement Learning[C]//Proceedings of the 37th International Conference on Machine Learning. Chia Laguna Resort: PMLR, 2020: 5757-5766.

[82] Sodhani S, Zhang A, Pineau J. Multi-task Reinforcement Learning with Context-based Representations[C]// Proceedings of the 38th International Conference on Machine Learning. Chia Laguna Resort: PMLR, 2021: 9767-9779.

[83] Wang J , Zhang J , Jiang H , et al. Offline Meta Reinforcement Learning with In-distribution Online Adaptation[EB/OL]. (2023-06-01). [2025-06-01]. https://arxiv.org/abs/2305.19529

[84] Beck J, Vuorio R, Liu Zheran, et al. A Survey of Meta-reinforcement Learning[EB/OL]. (2023-01-19) [2025-06-01]. https://arxiv.org/abs/2301.08028v1.

[85] Hallak A, Dotan Di Castro, Mannor S. Contextual Markov Decision Processes[EB/OL]. (2015-02-08) [2025-06-01]. https://arxiv.org/abs/1502.02259.

[86] Choi J, Guo Y, Moczulski M, et al. Contingency-Aware Exploration in Reinforcement Learning[EB/OL]. (2019-05-04) [2025-06-01]. https://arxiv.org/abs/1811.01483.

[87] Lagos J, Urho Lempiö, Rahtu E. Evaluating Generalization in Contextual Reinforcement Learning [EB/OL]. (2023-04-03) [2025-06-01]. https://arxiv. org/abs/2304.00793.

[88] Lanz D, Jürgen Seiler, Jaskolka K, et al. Compression of Dynamic Medical CT Data Using Motion Compensated Wavelet Lifting with Denoised Update[EB/OL]. (2023-02-02) [2025-06-01]. https://arxiv.org/abs/2302.01014.

[89] Krishna K M. Continuous Deutsch Uncertainty Principle and Continuous Kraus Conjecture[EB/OL]. (2023-10-02) [2025-06-01]. https://arxiv.org/abs/2310.01450.

[90] Laskin M, Srinivas A, Abbeel P. CURL: Contrastive Unsupervised Representations for Reinforcement Learning[C]//Proceedings of the 37th International Conference on Machine Learning. Chia Laguna Resort: PMLR, 2020: 5639-5650.

[91] Schwarzer M, Anand A, Goel R, et al. Data-efficient Reinforcement Learning with Self-predictive Representations[EB/OL]. (2021-05-20) [2025-06-01]. https://arxiv.org/abs/2007.05929.

[92] Fu Haotian, Tang Hongyao, Hao Jianye, et al. Towards Effective Context for Meta-reinforcement Learning: An Approach Based on Contrastive Learning[C]// Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence and the Thirty-Third Conference on Innovative Applications of Artificial Intelligence and the Eleventh Symposium on Educational Advances in Artificial Intelligence. Palo Alto, CA, USA: AAAI Press, 2021: 7457-7465.

[93] McInroe T, Lukas Schäfer, Albrecht S V. Learning Temporally-consistent Representations for Data-efficient Reinforcement Learning[EB/OL]. (2021-10-11) [2025-06-01]. https://arxiv.org/abs/2110.04935.

[94] Wang B, Xu S, Keutzer K, et al. Improving Context-based Meta-reinforcement Learning with Self-supervised Trajectory Contrastive Learning[EB/OL]. (2021-03-10) [2025-06-01]. https://arxiv.org/abs/2103.06386.

[95] Eysenbach B, Zhang Tianjun, Levine S, et al. Contrastive Learning as Goal-conditioned Reinforcement Learning [C]//Advances in Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2022:

35603-35620.

[96] Schug S, Kobayashi S, Akram Y, et al. Discovering Modular Solutions that Generalize Compositionally[EB/OL]. (2024-03-25) [2025-06-01]. https://arxiv. org/abs/2312.15001.

[97] Goeckner A, Sui Yueyuan, Martinet N, et al. Graph Neural Network-based Multi-agent Reinforcement Learning for Resilient Distributed Coordination of Multi-robot Systems[C]//2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). [S.l. : IEEE, 2024: 5732-5739.

[98] Zambaldi V, Raposo D, Santoro A, et al. Relational Deep Reinforcement learning[EB/OL]. (2018-06-28) [2025-06-01]. https://arxiv.org/abs/1806.01830.

[99] Shiarlis K, Wulfmeier M, Salter S, et al. TACO: Learning Task Decomposition via Temporal Alignment for Control [C]//Proceedings of the 35th International Conference on Machine Learning. Chia Laguna Resort: PMLR, 2018: 4654-4663.

[100] Shu Tianmin, Xiong Caiming, Socher R. Hierarchical and Interpretable Skill Acquisition in Multi-task Reinforcement Learning[EB/OL]. (2017-12-20) [2025-06-01]. https://arxiv.org/abs/1712.07294.

[101] Lee Y, Yang Jingyun, Lim J J. Learning to Coordinate Manipulation Skills via Skill Behavior Diversification [EB/OL]. (2019-12-20). [2025-06-01]. https://openreview.net/forum?id=ryxB2lBtvH.

[102] Yuan Haoqi, Zhang Chi, Wang Hongcheng, et al. Skill Reinforcement Learning and Planning for Open-world Long-horizon Tasks[EB/OL]. (2023-12-04) [2025-06-01]. https://arxiv.org/abs/2303.16563.

[103] Wen Yongyan, Li Siyuan, Zuo Rongchang, et al. SkillTree: Explainable Skill-based Deep Reinforcement Learning for Long-horizon Control Tasks[C]// Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence. Palo Alto: AAAI Press, 2025: 21491-21500.

[104] Schaul T, Horgan D, Gregor K, et al. Universal Value Function Approximators[C]//Proceedings of the 32nd International Conference on Machine Learning. Chia Laguna Resort: PMLR, 2015: 1312-1320.

[105] Narasimhan K, Barzilay R, Jaakkola T. Grounding Language for Transfer in Deep Reinforcement Learning [J]. Journal of Artificial Intelligence Research, 2018, 63 (1): 849-874.

[106] Qian Zhifeng, You Mingyu, Zhou Hongjun, et al. Weakly Supervised Disentangled Representation for Goal-conditioned Reinforcement Learning[J]. IEEE Robotics and Automation Letters, 2022, 7(2): 2202-2209.

[107] Peng Xuebin, Guo Yunrong, Halper L, et al. ASE: Large-scale Reusable Adversarial Skill Embeddings for Physically Simulated Characters[J]. ACM Transactions on Graphics, 2022, 41(4): 94.

[108] Jackermeier M, Abate A. DeepLTL: Learning to Efficiently Satisfy Complex LTL Specifications for Multi-task RL[EB/OL]. (2025-03-29) [2025-06-01]. https://arxiv.org/abs/2410.04631.

[109] Yalcinkaya B, Lauffer N, Vazquez-Chanlatte M, et al. Compositional Automata Embeddings for Goal-conditioned Reinforcement Learning[C]//Advances in Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2024: 72933-72963.

[110] Paliwal Y, Rajarshi Roy, Jean-Raphaël Gaglione, et al. Reinforcement Learning with Temporal-logic-based Causal Diagrams[C]//International Cross-Domain Conference for Machine Learning and Knowledge Extraction. Cham: Springer Nature Switzerland, 2023: 123-140.