摘要

随着深度学习技术的持续发展，策略学习算法，包括强化学习与模仿学习，已在棋类博弈、街机游戏等任务中取得显著突破。然而，在面对真实世界中的复杂环境时，策略学习仍面临诸多挑战，例如高维状态与动作空间、多模态感知输入、稀疏或延迟的奖励信号、任务目标的多样性以及环境动态的复杂性等。这些因素导致传统方法在样本效率、泛化能力与适应性方面表现受限，难以直接应用于实际场景。

为提升策略学习在复杂环境中的效率与实用性，本文围绕基于环境理解的引导式策略学习这一主题开展系统研究。引导式策略学习在强化学习与模仿学习的基本框架上引入额外的奖励塑造或表征迁移等手段，作为引导信号以提升策略学习效率；而环境理解则要求这些引导信号能够充分反映环境的语义结构与动态特征，从而增强策略在真实世界的复杂环境中的适应能力。本文分别从奖励塑造与表征迁移两种引导方式，探索了感知对齐、价值评估与行为抽象等环境理解方式在不同任务场景下的实现路径与引导效果，主要研究贡献包括以下四个方面：

第一，提出一种基于环境状态新颖性估计与影响建模的奖励塑造引导方法，用以解决策略学习在去中心化多智能体系统中的协同探索问题。该方法通过近似全局新颖性估计，并引入基于后见认知的多智能体影响建模，设计了两类内在奖励信号，实现多智能体系统中的价值评估。这些奖励信号能够有效引导智能体识别并理解环境中具备潜在价值的状态与行为，从而显著提升探索效率。实验结果表明，该方法在多个去中心化任务中均显著提升了探索效率与任务完成率，验证了所提出内在奖励机制在多智能体环境中对策略学习的引导能力。

第二，提出一种基于视觉理解与定位的表征迁移与奖励塑造引导方法，用于解决开放世界场景中的语言指令遵循策略的学习问题。该方法利用预训练多模态模型实现感知对齐与价值评估。具体而言，其利用多模态模型的视觉理解能力，构建可替代自然语言指令的任务表征，并结合辅助奖励信号引导策略学习。实验结果表明，该方法在开放世界中的多项基础技能学习任务中表现优异。同时，得益于多模态模型所具备的开放词汇泛化能力，该方法在面对训练阶段未出现过的自然语言指令时仍能实现零样本泛化，验证了预训练多模态模型在复杂环境下对策略学习的引导作用。

第三，提出一种基于任务完成度感知建模的奖励塑造引导方法，通过构建强化学习友好的奖励模型，深化对环境的价值评估能力。该方法首先构建高质量的视频-文本数据集，并在此基础上训练多模态模型，为下游强化学习任务提供辅助奖励信号。为增强奖励的引导能力，方法在模型训练阶段引入与任务完成度相关的正样本扰动机制，使模型输出的奖励信号能够反映任务的完成进度。实验结果表明，该方法相较于传统

的多模态模型作奖励信号的方式，在多个强化学习任务中显著提升了策略学习的效率与性能，验证了任务完成度的感知在复杂环境中增强策略学习引导能力的关键作用。

第四，提出一种基于隐变量建模与语言对齐的表征迁移引导方法，用于解决策略学习中的语义技能学习与组合泛化问题。该方法通过行为抽象，构建了一个基于离散隐变量的层级策略框架：底层技能编码通过语义正则化机制对齐至自然语言描述中包含的原子指令集合，上层策略则以自然语言为输入条件，预测技能序列中离散隐变量的分布，从而捕捉其潜在的多峰分布特性。实验结果表明，该方法在单一技能学习与多技能组合泛化任务中均表现优异，验证了自然语言引导机制在促进技能学习与泛化能力方面的有效性。

总体而言，本文围绕复杂环境中的策略学习问题，提出了多种基于环境理解的引导式学习方法。所提出方法在引导机制上涵盖奖励塑造与表征迁移，在环境理解的实现上兼顾感知对齐、价值评估和行为抽象等路径，共同构建了一个统一的引导式策略学习框架。实验证明，本文方法在多种复杂环境中均能显著提升策略学习的效率与性能，充分体现了基于环境理解的引导式策略学习的有效性与广阔应用前景。

关键词：策略学习，强化学习，模仿学习，奖励塑造，表征迁移，环境理解