分类号：TP391　　学校代码：10363
密　级：公开　　学　　号：2220920110

安徽工程大学

Anhui Polytechnic University

硕士学位论文

题目　离线环境下的强化学习方法研究

论 文 作 者　杨小涵
指 导 教 师　李钧（副教授）
学科（专业）　软件工程
研 究 方 向　领域软件工程

论文提交日期：2025 年 5 月 9 日

分类号：TP391　　单位代码：10363
密 级：公开　　学　号：2220920110

题　目　离线环境下的强化学习方法研究

英文并列
题　目　**Research on reinforcement learning methods in offline environment**

学生姓名：杨小涵
指导教师：李钧（副教授）
专　　业：软件工程
研究方向：领域软件工程

论文答辩日期：2025 年 5 月 10 日

安徽工程大学学位论文原创性声明

本人郑重声明：我恪守学术道德，崇尚严谨学风。所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已明确注明和引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品及成果的内容。论文为本人亲自撰写，我对所写的内容负责，并完全意识到本声明的法律后果由本人承担。

学位论文作者签名：杨小强

日期：2025 年 5 月 9 日

安徽工程大学学位论文版权使用授权书

学位论文作者完全了解学校有关保留、使用学位论文的规定，同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅或借阅。本人授权安徽工程大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

保密 □，在___年解密后适用本版权书。

本学位论文属于

不保密 ☑。

学位论文作者签名：杨小强　　指导教师签名：

日期：2025 年 5 月 9 日　　日期：2025 年 5 月 9 日

离线环境下的强化学习方法研究

摘 要

强化学习是一种使用智能体通过不断试错学习策略的机器学习范式。在该过程中，智能体以环境反馈中的奖励为导向，无需设定标签判断决策对错。强化学习为解决复杂决策问题提供了不可或缺的助力，如机器人控制、自动驾驶、个性化推荐等领域均有亮眼表现。然而随着强化学习算法应用场景和需求的日益复杂，强化学习中的环境交互风险大、实时采样耗时长等问题阻碍了强化学习的进一步发展。

离线环境下的强化学习算法因其无需实时采样花销小、使用离线数据集训练样本利用率高等特性受到广泛关注。离线环境下的强化学习聚焦于如何仅使用离线数据集训练智能体获得最优策略，其本质为离线强化学习。但是离线强化学习的决策优劣与离线数据集的质量息息相关。若训练所用的离线数据集分布与部署时的实际数据分布不同，则会导致分布偏移问题，造成智能体对未知动作做出过高评价的恶劣后果；若离线数据集良莠不齐、忽略环境不确定性带来的安全问题，则可能智能体无法兼顾高奖励和部署时的安全要求，造成智能体偏好高风险决策的不良结果。针对离线强化学习上述问题，提出以下两种方法：

（1）针对离线数据集分布与实际数据分布不同造成的分布偏移问题，提出了一种基于数据增强的双层离线强化学习框架（Diffusio

n-DA）。Diffusion-DA 的核心思想是通过使用信息熵不同的数据集对智能体进行训练，帮助智能体在学习原始数据分布的同时，捕捉优质轨迹。该框架分为低熵数据收敛层和高熵扩散探索层，智能体首先在低熵稳定的原始数据集中进行采样以习得稳定策略，然后使用高熵扩散数据集进行训练与调整以增强策略的鲁棒性。其中扩散模型进行数据增强生成扩散数据集，在贴近原始数据分布的同时，增加样本的多样性。在 D4RL 基准下进行的大量实验表明，Diffusion-DA 框架下各算法智能体表现优异。

（2）针对离线数据集良莠不齐、忽略环境不确定性所带来的安全问题，提出了一种基于双 Q 引导的离线强化学习安全策略增强算法（SPI）。SPI 的核心思想是使用扩散模型拟合相似数据分布的同时，引导生成满足安全约束的高奖励动作。首先，该算法重新映射离线数据集以增强成本属性。其次，将扩散模型、动作 Q 值函数和安全 Q 值函数耦合在一起，促使生成的动作收敛到符合安全约束的分布。在 DSRL 基准下进行的大量实验表明，SPI 在大多数任务中的表现都优于相关基线。

关键词：离线环境，强化学习，扩散模型，分布偏移问题，安全问题

RESEARCH ON REINFORCEMENT LEARNING METHODS IN OFFLINE ENVIRONMENT

ABSTRACT

Reinforcement learning is a paradigm of machine learning where an agent learns policies through trial and error. Through continuous attempts and learning, the agent eventually makes optimal decisions. In this process, the agent is guided by the rewards from the environment's feedback, without the need for preset labels to judge the correctness of decisions. Reinforcement learning provides indispensable assistance in solving complex decision-making problems, such as in robot control, autonomous driving, and personalized recommendation, where it has shown remarkable performance. However, as the application scenarios and demands for reinforcement learning algorithms become increasingly complex, issues such as the high risk of environmental interaction and the long time required for real-time sampling hinder the further development of reinforcement learning.

Reinforcement learning algorithms in offline environment have garnered widespread attention due to their low cost without real-time sampling and high utilization of training samples from offline datasets. Reinforcement learning in offline environment focuses on how to train

agents to obtain optimal policies using only offline datasets, which is essentially offline reinforcement learning. However, the quality of decisions in offline reinforcement learning is closely related to the quality of the offline datasets. If the distribution of the offline dataset used for training differs from the actual data distribution during deployment, it can lead to distributional shift problems, causing the agent to overestimate unknown actions with adverse consequences. If the offline dataset is of mixed quality and ignores the safety issues arising from environmental uncertainties, the agent may not be able to balance high rewards with safety requirements during deployment, leading to a preference for high-risk decisions. To address the aforementioned issues in offline reinforcement learning, the following two methods are proposed:

(1)To solve the problem of distribution shift caused by the difference between offline dataset distribution and actual data distribution, a double-layer offline reinforcement learning framework (Diffusion-DA) based on data augmentation is proposed. The core idea of Diffusion-DA is to train the agent using datasets with different information entropies, helping the agent to learn the original data distribution while capturing high-quality trajectories. The agent first samples the original dataset with low entropy stability to learn the stability policy, and then uses the high entropy diffusion dataset for training and adjustment to enhance the robustness of the policy. The diffusion model carries out data augmentation to generate

diffusion dataset, which is close to the original data distribution and increases the diversity of samples. A large number of experiments conducted under the D4RL benchmark indicate that the algorithms' agents under the Diffusion-DA framework perform excellently.

(2) To address the safety issues arising from the mixed quality of offline datasets and the neglect of environmental uncertainties, a safety policy improvement algorithm for offline reinforcement learning based on double Q guidance (SPI) is proposed. The core idea of SPI is to use a diffusion model to fit similar data distributions while guiding the generation of high-reward actions that satisfy safety constraints. First, the algorithm remaps the offline dataset to enhance cost attributes. Then, it couples the diffusion model, action Q-value function, and safety Q-value function to encourage the generated actions to converge to a distribution that complies with safety constraints. Extensive experiments under the DSRL benchmark show that SPI outperforms relevant baselines in most tasks.

KEY WORDS: offline environment, reinforcement learning, diffusion model, distributional shift problem, safety issues

目录

图录

表录

第1章　绪论

1.1 研究背景及意义

随着人工智能的浪潮席卷而来，众多优异的机器学习方法变得为人所熟知。其中强化学习凭借其强大的自主学习和决策能力在复杂任务中的展示出了突破性表现，在多个领域捷报频出。众所周知，在游戏领域[1-4]， 2016 年的围棋人机大战中，将深度学习应用于强化学习搜索算法的 AlphaGo[2]最终以 4 比 1 击败当时的世界冠军李世石，并于次年以 3 比 1 战胜当时的世界第一柯洁，而迭代更新后的通用化版本 AlphaZero[3]通过自学围棋、象棋和将棋可以达到远超人类的水平。无独有偶，在科学研究领域[5,6]，DeepMind 与瑞士洛桑联邦理工学院的合作成果——基于强化学习的磁控制器[5]，将使用人工智能控制核聚变的思想转化为现实。

与其他机器学习方法不同，强化学习与人类的学习方式更为相似。强化学习依靠智能体与环境交互来完成策略的学习。在交互过程中，智能体始终由奖励驱动进行试错，根据不断地试错对策略做出调整。无需标签数据，独特的优化目标以及坚实的数学理论支撑赋予了强化学习解决复杂问题的能力，然而强化学习的发展在某些领域遇到了一些阻碍。例如，机器人控制[7,8]、医疗健康、自动驾驶等领域，智能体控制的实体与真实环境进行交互时，试错成本大、采样耗时长。如果采取仿真环境替代真实环境，模拟数据的可用性尚未可知，还需权衡模拟器保真度与搭建成本、搭建难度之间的关系。上述问题不能简单地归因于算法效率或安全性的不足，其根本原因在于在这些领域中，智能体与环境的交互行为本身受到严格限制。

离线环境下的强化学习方法研究应运而生，离线环境下的强化学习是一种仅使用静态数据集作为智能体训练数据来源的强化学习算法[9]。它支持在不与环境进行实时交互的设定下，以静态数据集中的批量数据训练智能体获得优质策略，故又名离线强化学习或批量强化学习。前者强调智能体具有离线不交互的特性，后者强调从数据集中批量采样训练智能体。相较于需要在线交互式的

强化学习算法，离线强化学习优点众多，在于个性化推荐[10,11]、医学治疗[12]、量化交易[13]等领域发挥巨大应用价值。（1）样本效率高。离线强化学习专注于如何挖掘静态数据集中的潜力以训练得到具有优质决策能力的可用智能体。基于数据驱动的离线强化学习为历史数据赋予了新价值，通过对数据集进行反复采样，提高数据利用效率。（2）节约成本。由于无需与环境进行交互，离线强化学习天然地规避了实时采样过程中资源与时间成本的消耗。此外，离线强化学习能够利用与目标策略不同的行为策略所采集的数据集进行训练。这一特性显著拓展了可用数据集的范围，并有效降低了数据集的获取成本。（3）训练安全性强。不同于在线强化学习要在探索和利用之间做出权衡，离线强化学习仅需考虑如何充分利用已有经验（即数据集中的数据）学习高质量策略。在离线训练阶段，由于不涉及与环境的直接交互，离线强化学习能够有效避免因探索行为可能引发的碰撞、损坏等不安全事件，从而显著降低潜在风险。

离线强化学习为众多领域的发展提供了有力支撑，然而在实际应用中仍面临一些挑战与限制。（1）分布偏移引起的众多问题。在离线强化学习中，理想的训练数据集应涵盖完整的动作与状态空间，然而在实际应用中，此类数据集往往难以获取。当数据集中数据分布不同于部署数据的分布时，模型对于超出认知范围之外的数据会做出错误的预测或判断，即分布偏移问题会引起外推误差[14]。在离线环境下，智能体在训练时只能依靠数据集中存储的以往交互经验对 Q 值进行拟合。因此实际部署时，智能体无法对未知动作的 Q 值进行有效估计，并且倾向于对状态动作对做出乐观评价，即外推误差会进一步加剧 Q 值高估问题[14]。（2）离线数据集良莠不齐、忽略环境不确定性带来的安全问题。离线强化学习允许使用任意策略收集的训练数据集对智能体进行训练。然而，由于不同策略的偏好各异，很难确保所收集的轨迹在追求高奖励动作的同时，能够充分关注到安全需求。在离线环境下训练时，智能体无法通过交互过程中的恶性事件（如危险状态）获取安全警告信号。因此，训练过程中可能会忽略实际部署时必须满足的安全约束条件，从而导致训练所得智能体在部署时产生重大安全决策失误。

综上所述，开展离线环境下的强化学习方法研究至关重要，本课题的研究具有重要研究价值和理论意义。

1.2 国内外研究现状

分布偏移问题是离线强化学习领域中最核心的挑战之一。然而，除了这一关键问题外，针对安全策略、最优策略和最优轨迹分布方向的研究也在迅速发展[15,16]。由于离线强化学习领域的研究问题不断涌现且尚未形成统一框架，故采用其他角度对现有研究进行分类。根据环境建模方式，离线强化学习算法主要分为：基于无模型的离线强化学习算法、基于有模型的离线强化学习算法、基于序列模型的离线强化学习算法[17]。

1.2.1 基于无模型的离线强化学习算法

基于无模型的离线强化学习算法无需对环境进行建模，要求智能体直接从离线数据集中学习到策略或价值函数。在这类算法中，模型的核心任务在于准确拟合策略网络或价值函数网络，并且不会显式构建奖励函数或状态转移矩阵。

1.2.1.1 基于策略约束的离线强化学习算法

基于策略约束的离线强化学习算法通过约束目标策略偏离行为策略的程度，使得算法具有良好效果。

2018 年，Fujimoto 等人[18]通过对比同策略智能体和异策略智能体在使用相同算法和相同数据集训练后的表现差异，发现智能体会对未知的状态动作对进行错误估值，并将这一现象命名为外推误差。为解决此问题，提出 BCQ（Batch-Constrained deep Q-learning）。该算法使用状态条件生成模型与 Q 网络结合，产生已知动作的同时，选择与训练数据集相似度最高的动作进行更新。

2020 年，Wang 等人[19]提出 CRR（Critic Regularized Regression），将离线策略优化简化为一种 Q 值函数过滤，并利用 KL 散度对策略添加隐式约束。Nadjahi 等人[20]提出 SPIBB（Safe Policy Improvement with Baseline Bootstrapping），该算法不再将状态动作对分类为不确定组和安全训练组两组，而是采用一种更软的策略，利用局部模型不确定性约束策略变化来控制价值估计中的误差。该方法可以在不确定的操作上承担更大的风险，同时保持可证明的安全性。

2022 年，Lyu 等人[21]认为只要分布外动作的估计值不影响最优策略的学习，就可以赋予较高的值，故提出 MCQ（Mildly Conservative Q-learning），通过给分布外动作赋予合适的伪 Q 值来进行训练，保证保守的同时提升泛化性。Wang

等人[22]提出 Diffuison-ql（Diffusion-Q-Learning），利用条件扩散模型来表示策略。将扩散模型应用于策略空间，形成以状态为条件的条件扩散模型。使用一个行为克隆项和一个策略改进项，鼓励扩散模型在与训练集相同的分布中采样动作的同时，试图采样高值动作（根据学习到的 q 值）。Janner 等人[23]提出了 diffuser，该算法使用 DDPM（Denoising Diffusion Probabilistic Models）通过迭代去噪轨迹来进行规划，以生成轨迹的准确性而非单步转移的准确性为目标，因此它不会受到推演误差会逐步累积放大的影响。Ajay 等人[24]提出 DD（Decision Diffuser），使用 Guided Diffusion 在离线数据集上生成奖励较高或满足某些约束或者学会某些技能的数据，再从中提取出策略。这个过程不涉及到强化学习中的 Bellman 更新，避免了外推误差。

2023 年，Lin 等人[25]提出 TREBI（Trajectory-based REal-time Budget Inference），从轨迹分布的角度来研究有约束的策略优化问题。这不仅可以推导出给定预算和离线数据集的最优轨迹分布，而且可以在轨迹水平上实现严格的约束满足（即理论上概率为 1），而不像以往大多数研究中的粗糙约束满足（即对轨迹的期望）。

2024 年，Jin 等人[26]提出了一种带有悲观价值的迭代算法 PEVI（Pessimistic Variant of the Value Iteration），通过引入不确定性量化器作为惩罚函数，这种惩罚函数简单地翻转了在线强化学习中奖励函数的符号，以促进探索，并且该算法无需假设数据集的充分覆盖。Zhu 等人[27]提出 MADiff（Multi-Agent Learning with Diffusion Models），该算法是首个将扩散模型引入多智能体学习场景的算法。MADiff 采用基于注意力的扩散模型来捕捉多智能体之间的协调行为，主要用于模拟多智能体之间多变的交互行为。在分散式执行时，MADiff 在处理自我智能体信息的同时，将队友智能体行为进行建模。在统一决策时，MADiff 可作为集中控制器收集所有智能体的状态、动作、环境信息，做出基于全局信息的综合最优决策。

当行为策略分布“容易”建模时，策略约束通常有很好的表现，但往往更保守。这类需要对行为策略做出评估，如果错误地估计了行为策略（例如，将单模态策略拟合到多模态数据时），策略约束方法可能会出现极差的表现。

1.2.1.2 基于正则化值函数的离线强化学习算法

基于正则化值函数的离线强化学习算法是约束值函数，将低值赋给分布外行为。

2020 年，Kumar 等人[14]提出了 CQL（Conservative Q-Learning for Offline Reinforcement Learning），通过在标准贝尔曼公式的基础上增加一个模块，学习一个保守的值函数来缓解值函数高估问题，学习的值函数是真实值函数的下界。

2022 年，Kostrikov 等人[28]提出 IQL（Implicit Q-Learning），该算法不学习分布外动作之外的动作，而是用已知的状态动作对进行学习，通过使用 SARSA 的方式重构策略和值函数，在策略的抽取方面采用了 AWR（Advantage Weighted Regression）方式抽取，直接确定值函数如何随着不同的动作而变化，而不是确定值函数如何随着不同的未来结果而变化。

2023 年，Wang 等人[29]提出 UWCQL（Uncertainty Weighted Conservative Q-Learning），在 CQL 算法的基础上额外计算不确定性，将不确定性作为权重赋予 CQL 算法中值函数的正则化项来更精确地对价值函数进行估计。通过不确定性判断状态动作对的分布，赋予不确定性较高的状态动作对更低的价值函数，反之对不确定性较低的状态动作对则不进行保守估计，从而解决 CQL 算法过于保守的问题，使智能体能够学习到最优策略价值函数进行估计。

2024 年，Huang 等人[30]提出了一个简单高效的 ORL-RC （Offline Reinforcement Learning with Relaxed Conservatism） 框架，提出了一个新的混合 Bellman 算子。这个混合算子通过一个超参数 β 来控制行为策略与目标策略之间权重，从而使得算法能够在学习过程中同时利用目标策略和行为策略的信息，最终学习到接近真实 Q 函数的 Q 函数。

基于正则化值函数的离线强化学习算法对行为策略没有限制，因此会更不保守。通常正则项会额外使用基于模型或者策略约束的方法来增加算法的保守性。

1.2.2 基于模型的离线强化学习算法

基于模型的离线强化学习算法从收集的数据集中学习动态模型，并根据动态模型学习值函数和策略，不需要依赖于收集的数据集来估计值函数。

2020 年，Kidambi 等人[31]提出了 MOReL（Model-Based Offline Reinforcement Learning），使用离线数据集学习悲观 MDP 模型，能够在 P-MDP

中学到一个接近最优的策略，其中 P-MDP 是真实 MDP 策略的一个下界，这同时能够减轻模型进行探索的影响。Yu 等人[32]提出了 MOPO（Model-Based Offline Policy Optimization），通过对不满足条件的状态动作对的奖励函数进行惩罚来限制策略的学习，将动力学模型不确定性转变为对奖励函数的惩罚，这种方式相当于在泛化性和风险之间做权衡。

2021 年，Yu 等人[33]提出了 COMBO（Conservative Offline Model-Based Policy Optimization），针对由模型生成的行为策略数据外的区域上的动作状态对，对其值函数进行规范化，该方法优化了真实策略的下界。Lu 等人[34]针对基于模型的离线强化学习方法的限制因素和超参数进行研究，最终得出以下结论：更大的惩罚可以改进现有的方法；使用规范形式的不确定性估计的惩罚和分布外动作度量具有更好的相关性。与分布漂移相比，不确定性与动态误差的相关性更大。

2022 年，Rigter 等人[35]提出 RAMBO（Robust Adversarial Model-Based Offline Reinforcement Learning），引入了 RAVL（Robust Adversarial Reinforcement Learning）的思想，不需要考虑复杂的不确定度计算，而是用对抗的思路保证算法的悲观主义。将环境的概率转移模型视为对抗智能体，目标是要让原来的智能体的收益尽量小。

2023 年，Kim 等人[36]提出 Count-MORL（Count-Based and Model-Based Offline Reinforcement Learning），通过集成计数保守主义来量化模型估计误差，提高了学习策略的最优性能保证。算法利用状态-动作对的计数估计来量化模型估计误差，通过降低估计误差来提高所学策略的保守性。

2024 年，Hong 等人[37]提出 model-based RL method for confounded POMDPs.（Model-Based Reinforcement Learning Method for Confounded Partially Observable Markov Decision Processes），用于处理受混淆因素影响的局部可观测马尔可夫决策过程。该算法构造了基于模型的识别模块，以支持智能体学习在混淆的局部可观测马尔科夫决策过程中，动作对奖励和未来状态转移的影响。并且设计了一种估计程序用于进行异策略的评估。

基于模型的离线强化学习通常适用于数据分布覆盖率相对较高且易于学习模型的情况。但是这类算法中的动态模型可能会过度拟合有限的数据集，并在

行为策略未访问的区域遭受外推错误，从而导致部署时学习策略的不稳定性。此外，在模型推理中，即使一步预测误差很小，复合误差（即仿真轨迹与现实之间累积的预测误差）也会很大。

1.2.3 基于序列模型的离线强化学习算法

基于序列预测的离线强化学习算法把每一条轨迹视为一个样本点，从而将研究目标转变为建模整条轨迹的分布，而将轨迹的最优性作为其中的条件变量来从分布中采样轨迹。

2021 年，Chen 等人[38]提出了 DT（Decision Transformer），该算法不再使用 TD learning 等传统强化学习方法训练策略，而是将强化学习重构成一个适合 Transformer 的序列问题。基于收集的经验使用序列建模目标来训练 Transformer 模型，直接进行决策。并且在建模时，将轨迹建模成 reward-to-go，状态，动作的序列，模型的输出为给定 reward-to-go 对应的动作。Janner 等人[39]提出 TT（Trajectory Transformer），采用 TT 拟合状态、动作和奖励的分布，并使用集束搜索来对解码出的候选轨迹（策略）进行搜索和优化。

2022 年，Wang 等人[40]提出 BooT（Bootstrapped Transformer），结合了自举的思想，利用学习到的模型自行生成更多的离线数据，进一步推动序列模型的训练。BooT 分成两个主要部分，一部分用学习的序列模型生成轨迹，另一部分使用生成的轨迹来扩充原始离线数据集，用于引导序列模型学习。Xu 等人[41]提出 Prompt-DT（Prompt-Based Decision Transformer），利用 Transformer 架构的顺序建模能力和提示框架，通过设计轨迹提示，编码特定任务信息，指导策略生成。

2023 年，Yamagata 等人[42]提出 QDT（Q-learning Decision Transformer），在 DT 的基础上做出改进。QDT 通过使用 Q-learning 的动态规划结果重标记训练数据中的 return-to-go，进而利用这些重标记的数据来训练 Decision Transformer。Zhu 等人[43]提出 DTRD（Decision Transformer in Long-Delayed Reward），用以解决解决奖励信号延迟的问题。在长延迟设置下，奖励信号稀少、分布不均。并且在某些极端情况下，智能体在整条轨迹结束之后，才会获得信号奖励。DTRD 通过设计奖励再分配机制，优化奖励与长序列中动作状态对的依赖关系，以缓解奖励信号延迟所带来的累积偏差。

2024 年，Huang 等人[44]提出 IDT（In-context Decision Transformer），模仿人类决策的高效层次结构，将序列重构为高层决策而非与环境交互的低级行动，从而避免过长的序列并更高效地解决复杂任务。

基于序列预测的离线强化学习算法基于某种模型假设，例如马尔可夫假设或时序模型。然而，真实环境可能不符合这些假设，导致模型与真实环境之间的误匹配。这种误匹配可能导致预测的不准确性，进而影响策略的学习和优化过程。

1.3 主要研究内容

近年来，强化学习在解决复杂决策问题方面取得了巨大成功，然而在线训练智能体的昂贵花销和高风险仍然是阻碍强化学习智能体在现实世界更广泛应用的主要挑战，这促进了对离线环境下强化学习方法的研究[45]。通过对离线环境下的强化学习算法现状进行分析，针对分布偏移、部署时策略安全性不足等现有问题进行研究，提出对应解决方案。论文研究思路框架如图 1-1 所示，主要研究内容如下：

（1）离线环境下的数据集增强框架研究。针对离线数据集分布与实际数据分布不同造成的分布偏移问题，提出了一种基于数据增强的双层离线强化学习框架 Diffusion-DA。该框架通过低熵数据收敛层和高熵扩散探索层两个层级，鼓励智能体接触优质策略的同时，兼具一定的泛化能力。低熵数据收敛层所用数据集直接采用未处理的原始数据集，支持智能体通过原始数据快速收敛至稳定。高熵扩散探索层中所用数据集则经过数据增强处理，对原有静态数据集进行增强，保证生成的数据分布与原有数据分布相似的同时具有良好的质量。

（2）离线环境下安全动作增强研究。针对离线数据集良莠不齐、忽略环境不确定性所带来的安全问题，提出了一种基于双 Q 引导的离线强化学习安全策略增强算法 SPI。该算法在去噪过程中，使用安全判别结果和奖励判别结果对扩散模型的生成结果进行微调，确保智能体最终决策结果具有安全和高奖励的保障。此外，为了确保安全判别器能够拟合安全数据分布做出正确判断，SPI 对原始数据中的 cost 属性进行重新映射，为智能体做出安全决策打下坚实基础。

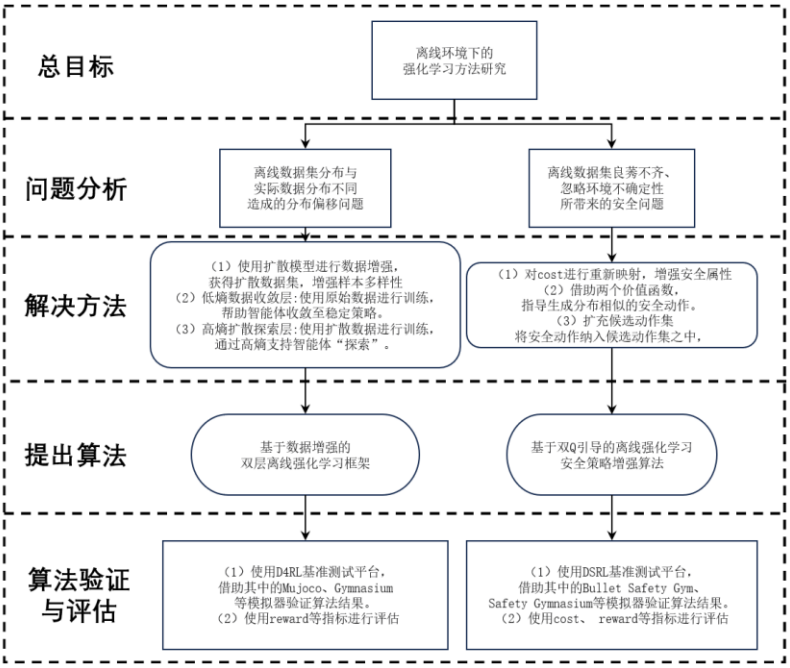

图 1-1 论文研究思路框架图

Figure 1-1 Framework diagram of the research idea of the dissertation

1.4 论文组织结构

第一章绪论。该章对离线环境下的强化学习算法的研究背景和意义进行描述，使用文字语言介绍离线强化学习相关知识和基本定义。然后介绍离线强化学习目前国内外研究现状，阐述离线强化学习算法的分类以及不同分类下算法的更新迭代过程。此外，总结全文主要研究内容，并展示了论文研究思路框架。最后说明全文组织结构，简要总结各章节内容。

第二章相关技术概述。该章以数学公式语言为主要媒介进行介绍。首先介绍强化学习，通过公式定义和推导展示强化学习完善的数学理论支持。其次根据是否对智能体施加安全约束，引入安全强化学习。然后根据智能体是否与环境发生实时交互，将强化学习细化至在线强化学习和离线强化学习两个子领域并分别进行分析。此外，介绍热门生成式模型的核心思路和创新。最后引出本章小结，对该章重点内容进行重点提炼和总结。

第三章基于数据增强的双层离线强化学习框架。该章在引言部分表对该框架的问题分析与定义、研究动机等基本信息进行表述，主要从离线强化学习、外推误差、扩散模型等方面进行介绍。其次，阐述了该框架的方案设计，介绍方案设计思路和具体架构，如数据增强、低熵数据收敛层与高熵扩散探索层等。

然后在实验评估中，展示该框架的实验设计、评估指标，如 reward 指标等。借助 D4RL 中的不同实验环境，对框架的有效性进行评估和论证，分析现有实验结果进行归因。最后通过文章小结对框架内容进行回顾。

第四章基于双 Q 引导的离线强化学习安全策略增强算法。该章首先在引言中介绍了安全离线强化学习、受限马尔科夫链、约束下求解最优动作等基本信息。其次，根据对研究问题的分析确立了该算法的方案设计，主要分为增强安全特征，安全分布建模，提取安全策略三个部分。然后在实验评估中，利用 DSRL 中的安全离线任务衡量算法安全性、参数敏感性等各项性能，并借助消融实验论证算法设计的合理性。最后在文章小结中，对算法内容进行回顾。

第五章为总结与展望，通过总结和展望两个部分，完成全文内容的梳理概括，总结所作工作的未来改进方向。

第2章　相关理论概述

2.1 强化学习

简言之，强化学习是一种通过交互，令学习者懂得如何完成任务的范式。学习者作为被操控对象可被部署在系统中控制机器人、推荐系统等实体进行行动，而在强化学习理论中学习者的被统称为 agent，译为智能体或代理。智能体的训练过程与婴儿学步相似，通过不断地试错最终习得技巧。但在训练时，智能体不会收到任何向左或向右等类似的明确行动指令。智能体在不同环境状态下尝试不同的动作，获得环境对此的反馈，最终理解状态、动作与反馈之间的因果关系。过程如图 2-1 所示。

图 2-1 智能体与环境交互

Figure 2-1 Agent interacting with the environment

图中一次循环被称为一个时间步 t，从第一个时间步到最后一个时间步被称为一幕。例如，在围棋竞技中，每落一子即是一个时间步，而从棋盘为空到胜负可分即是一幕。一幕中的完整轨迹可以表示为：

$$\tau = S_0, A_0, R_0, S_1, A_1, R_1, S_2, A_2, R_2, \ldots S_t, A_t, R_t \tag{2-1}$$

奖励 R 是一个用于评价当前智能体所选动作好坏的标量，故环境反馈的奖励值可为正值或负值。考虑到长期序列决策的需求，强化学习基于奖励定义了另一个评价标准——回报 G_t^{r} 。

$$G_t^{\mathrm{r}} \doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \tag{2-2}$$

其中$\gamma(\gamma \in [0,1])$表示折扣率，用于调节智能体对奖励的“视野”。当γ趋近于 0 时，智能体会更加“急功近利”，追求眼前的即时奖励；当γ趋近于 1 时，智能体会更加“高瞻远瞩”，注重未来的潜在奖励。回报是从当前时间步到此幕结束时所获得的累积奖励。强化学习的优化目标是最大化期望回报或最大化期望累积奖励，这是由两个特性所决定。（1）智能体所采用的动作不仅作用于当前状态，还会影响到后续状态序列。（2）智能体采取动作后所得奖励，多数为延迟奖励，少数为即时奖励。强化学习的最优策略需满足$\pi^* = \arg\max \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} G_t^{\mathrm{r}} \mid s_0 = s\right]$，即寻求一个可以获得最大期望回报的策略。

环境是一个宽泛的概念，并不完全等同于人类习惯认知中的环境。例如，智能体操控智能车自适应巡航时，道路即是与其进行交互的环境。而如果智能体仅用于智能车自适应循环时的电量控制时，此时环境是智能车内除电量监控装置外其余部件和道路。

状态S是用于反映环境或系统当前的情况的标量。当状态信息符合马尔可夫性质时，状态转移概率之间有以下特性：

$$\mathbb{P}[S_{t+1} \mid S_t] = \mathbb{P}[S_{t+1} \mid S_1, ..., S_t] \tag{2-3}$$

即下一个状态S_{t+1}仅与当前状态S_t有关，与其余状态无关。在不同应用领域中，根据智能体是否能观测到全部的环境状态信息，会选择使用普通的马尔可夫决策过程或特殊的部分可观察马尔可夫决策过程进行建模。例如，在围棋竞技中，智能体所接受的全部信息（棋盘全部棋子摆放位置）此时等同于环境信息（棋盘全部棋子摆放位置），智能体会使用马尔可夫决策过程进行建模；在扑克竞技中，智能体所接受的全部信息（场上所打出的牌面信息和自我牌面信息）此时不等同于环境信息（场上所打出的牌面信息和全部玩家的牌面信息），智能体会使用部分可观察马尔可夫决策过程进行建模。

2.1.1 强化学习智能体的主要构成部分

强化学习中有重要的三个构成部分，分别是策略、值函数和模型。

策略指的是智能体内部的行为机制，是从状态到动作的映射，即智能体在

当前状态下会选择怎样的动作输出。强化学习中有两类代表性策略表达：确定性策略和随机策略。

确定型策略适用于指令明确、环境不确定度低的领域，智能体在接收到环境反馈的状态后，会直接输出指定的动作。

$$a = \pi(s) \tag{2-4}$$

随机策略适用于环境复杂、动态变化大的领域，智能体在接受到环境反馈的状态后，会通过动作的概率分布确定输出的动作。

$$\pi(a \mid s) = \mathbb{P}[A_t = a \mid S_t = s] \tag{2-5}$$

值函数类似于奖励，用于评价智能体所选动作。奖励仅考虑当前状态下所选择动作的好坏，是一个与状态动作相关的直接反馈。值函数考虑从当前状态直到这一幕结束时，智能体可获得的所有累积奖励。值函数是一种更为远视的评价标准，但他依赖于奖励。智能体借助值函数从长序列的角度，将当前和未来的状态动作对或状态均纳入考虑范围之内，如图 2-2 所示。

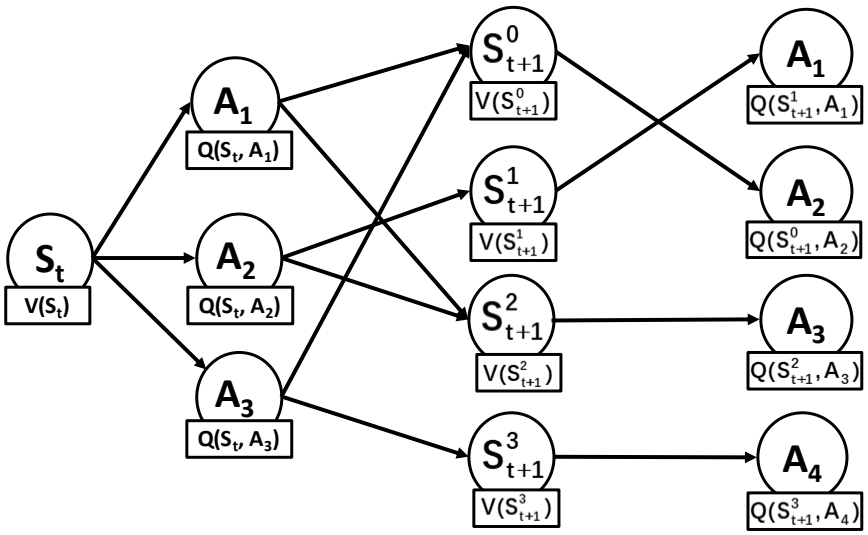

图 2-2 蒙特卡洛树

Figure 2-2 Monte Carlo trees

强化学习中，值函数分为动作价值函数和状态价值函数。

动作价值函数 $Q_\pi(s,a)$ 用于表示当前状态 s 下选择动作 a，且在此幕结束之前的所有状态下始终执行策略 π 所获得的全部期望回报。动作价值函数从长序列的角度，评价当前状态动作对的优劣。

$$\begin{aligned} Q_\pi(s,a) &\doteq \mathbb{E}_\pi[G_t^{\mathrm{r}} \mid S_t = s, A_t = a] \\ &= \mathbb{E}_\pi\left[\sum_{\infty}^{k=0} \gamma^k R_{t+k+1} \mid S_t = s, A_t = a\right], \text{for all } s \in \mathcal{S} \end{aligned} \tag{2-6}$$

状态价值函数$v_\pi(s)$用于表示在此幕结束之前的所有状态下始终执行策略π所获得的全部期望回报。状态价值函数从长序列的角度，评价当前状态的优劣。

$$\begin{aligned} V_\pi(s) &\doteq \mathbb{E}_\pi[G_t^{\mathrm{r}} \mid S_t = s] \\ &= \mathbb{E}_\pi\left[\sum_{\infty}^{k=0} \gamma^k R_{t+k+1} \mid S_t = s\right], \text{ for all } s \in \mathcal{S} \end{aligned} \tag{2-7}$$

通过动作价值　函数（Q 函数）与状态价值函数（V 函数）的相互代入与推导，可以建立两者之间的关系。

$$V_\pi(s) = \sum_{a \in A} \pi(a \mid s) Q_\pi(s,a) \tag{2-8}$$

$$Q_\pi(s,a) = \sum_{s',r} p(s', r \mid s, a)(r + \gamma V_\pi(s')) \tag{2-9}$$

再通过简单的推导之后，可以得到贝尔曼期望方程。贝尔曼期望方程将期望回报根据时间步进行划分，将状态价值函数和动作价值函数调整为由即时奖励和未来奖励的加权和的形式。

$$\begin{aligned} V^\pi(s) &= \mathbb{E}_\pi[R_t + \gamma V^\pi(S_{t+1}) \mid S_t = s] \\ &= \sum_{a \in A} \pi(a \mid s)\left(r(s,a) + \gamma \sum_{s' \in S} p(s' \mid s, a) V^\pi(s')\right) \end{aligned} \tag{2-10}$$

$$\begin{aligned} Q^\pi(s,a) &= \mathbb{E}_\pi[R_t + \gamma Q^\pi(S_{t+1}, A_{t+1}) \mid S_t = s, A_t = a] \\ &= r(s,a) + \gamma \sum_{s' \in S} p(s' \mid s, a) \sum_{a' \in A} \pi(a' \mid s') Q^\pi(s', a') \end{aligned} \tag{2-11}$$

模型是指环境模型，是在智能体内部用于模拟环境反馈的模型。模型在接收到当前状态和智能体执行的动作后，会产生该状态动作对相关联的奖励或转移状态。例如$P_{ss'}^a = \mathbb{P}[S_{t+1} = s' \mid S_t = s, A_t = a]$用于预测在状态$s$下执行动作$a$后，状态的跳转；而$R_s^a = \mathbb{E}[R_{t+1} \mid S_t = s, A_t = a]$用于预测在状态$s$下执行动作$a$后，奖励的产生。

根据模型输出信息的详细程度，强化学习将模型分为分布模型和样本模型。当模型对状态动作对预测出完整的状态转移概率和奖励概率时，智能体可以根据模型信息模拟出状态动作对的每一个后继状态从而进行准确的规划，此类模

型为分布模型。分布模型可以对环境进行详细的表征，利于智能体做出准确无误的细致规划，但需要大量计算资源且建模难度与环境复杂度呈正相关趋势；当模型对状态动作对仅预测出一种可能性并进行采样时，智能体可以根据输出信息模拟出当前状态动作对的一种可能后继状态从而进行快速的规划，此类模型为样本模型。样本模型可以对模型进行简单建模，并且易于实现，但表征能力很难达到分布模型的水平。

根据强化学习算法中是否包含策略、值函数和模型三大重要组成部分，可将其分为多类。如图 2-3 所示。

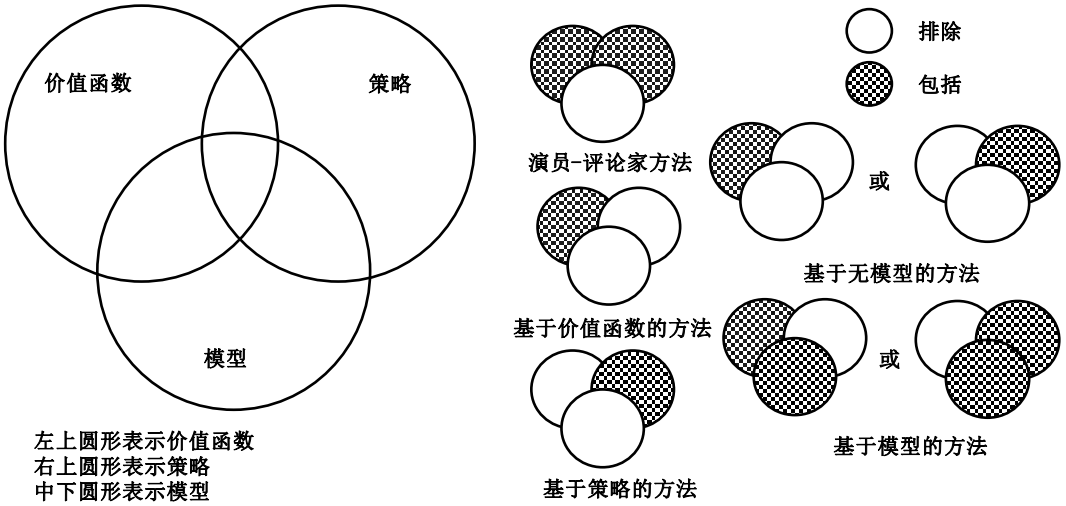

图 2-3 强化学习算法分类

Figure 2-3 Classification of reinforcement learning algorithms

演员-评论家算法是强化学习领域中一类不容忽视的重要算法[9]。该类算法核心是依托演员和评论家两个模块的相互优化，共同鼓励智能体收敛至最优策略。演员是由某种策略构成，如随机策略或贪婪策略，用于根据当前状态产生动作。评论家通常由值函数构成，用于对演员选择的动作进行评价。演员根据评论家的反馈对策略进行更新，评论家根据演员选择的动作对值函数进行更新。

2.1.2 强化学习中的重要问题

强化学习发展如火如荼，但仍存在一些重要问题：利用与探索、预测与控制。

利用与探索。利用与探索困境发生在智能体还未对状态和动作进行准确建模之时。在决策时，智能体需要权衡选择已知动作中收益最高者还是选择尝试

新的未知动作。若一味选择已知动作中收益最高者，会错失未知的高收益动作。而选择勇敢尝试未知动作，却要承担收益低的风险。环境所具有的不确定性要求智能体必须在多次决策之后才能完成正确建模，这为智能体决策带来了更多困难。强化学习中多个算法对利用与探索进行研究，例如 ε-greedy 方法。

ε-greedy 方法的核心是设定 $\varepsilon(\varepsilon \in [0,1])$ 参数用于调节智能体在决策时偏向于利用还是探索。具体而言，智能体在每次决策之前，会随机产生一个小数 n $(n \in [0,1])$。若 $n > \varepsilon$，则在选择动作时遵从贪婪策略，即 $A_t \doteq \arg\max_a Q_t(a)$。反之，则在选择动作时在已知动作中随机选择；通过在不同时期对 ε 值做出调整，帮助智能体在“懵懂无知”时积极对未知世界进行探索，在“经验丰富”时偏好借助已有经验做出判断。

预测与控制。预测与控制困境发生在智能体进行策略优化和策略评估的迭代时期。预测是指对给定的策略进行策略评估，即固定策略参数，并迭代计算当前策略下的所能获得的累积期望回报；控制是指根据价值函数对策略优化，如贪婪更新（将策略调整为选择所有动作中当前动作价值函数值最大者），即根据值函数改变策略参数，多次反复找到最优策略。智能体在训练时，始终固定策略参数和频繁随机改变策略参数两者都是不可取的训练技巧，如何在预测与控制之间找到平衡点成为强化学习中的重要问题。

2.1.3 同策略与异策略算法

强化学习得以依靠数据驱动获得优质策略的原因之一是，其对收集数据集所用策略和部署智能体所用策略之间并无强制等价关系约束。在强化学习中，用于与环境交互收集信息产生数据的策略被称为行为策略，即训练过程中的策略。而在行为策略产生的数据中不断学习和优化得到的策略被称为目标策略，即学习训练完毕后拿去做行为评估的策略。

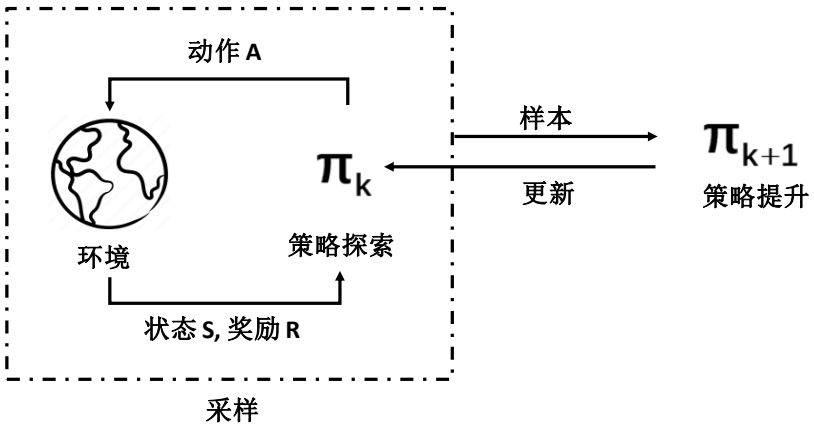

图 2-4 同策略算法

Figure 2-4 On policy algorithms

同策略（On Policy）算法是指行为策略（去产生数据的策略）和目标策略（从数据学习得到的策略）两者相同。同策略是“边学边做”，其核心思想是智能体使用自身策略采样，并根据环境反馈更新价值函数。使用更新后的价值函数进行策略评估，并依照评估结果对当前的策略进行优化。此时，待优化策略等同于当前自身策略。简而言之，智能体的决策阶段和学习阶段相重合，学习者就是决策者。例如，没有任何先验知识的围棋智能体通过对弈学习围棋基本理论。

同策略算法无需借助条件转换等额外操作，所得即所学，帮助智能体快速收敛。但每次更新都需要进行新的采样，且采样时需要平衡探索和利用。倾向于探索，智能体无法获得较高收益，学习效率低下。倾向于利用，智能体会陷入局部最优。

异策略（Off Policy）算法是指，行为策略（去产生数据的策略）和目标策略（从数据学习得到的策略）两者不同，但须满足条件：存在目标策略 $\pi_\alpha(a|s)>0$ 必然有行为策略 $\pi_k(a|s)>0$ 成立，即行为策略的可能出现的动作必须涵盖目标策略的所有可能动作。异策略是“站在巨人的肩膀上”，其基本思想是智能体保留自身策略，采用代替策略进行采样。代替策略范围不受任何限制，如优化成熟的人类策略、已习得的过往策略等。智能体通过代替策略与环境进行交互，获取相应的奖励信号更新自身策略的价值函数。简言之，具体而言，智能体在自身策略所形成的价值函数基础上，通过观察其他策略所引发的行为，

获取有价值的信息。异策略允许智能体从广泛的行为模式中提取经验，用于优化自身算法。此类方法确保了数据全面性，覆盖较大的动作空间，数据来源多样（自行产生或者外来数据均可）[46]。并且迭代过程中允许存在两个策略，将探索和利用分离，支持灵活调整探索概率。但由于目标策略与行为策略并不相同，策略之间的偏差为更新过程引入了高方差。例如通过重要性采样进行调节偏差影响时，对数据点的不同加权导致策略收敛震荡[47]。

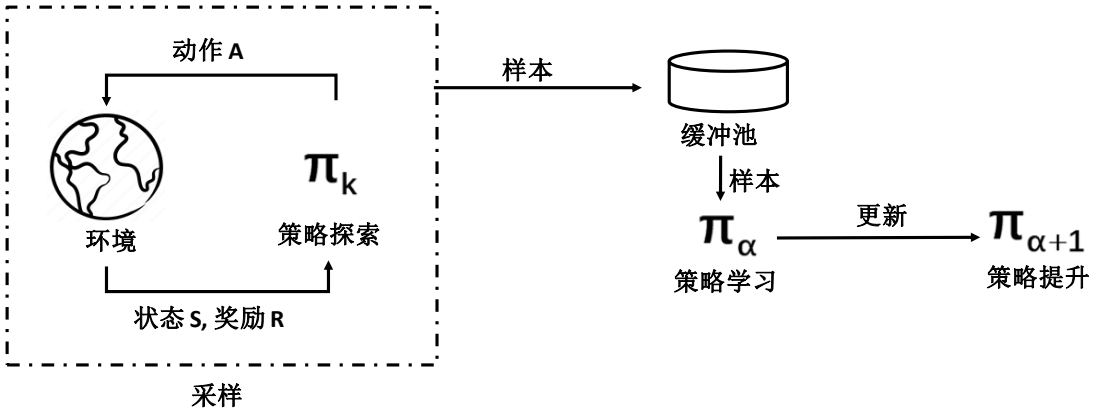

图 2-5 异策略算法

Figure 2-5 Off policy algorithms

2.2 安全强化学习

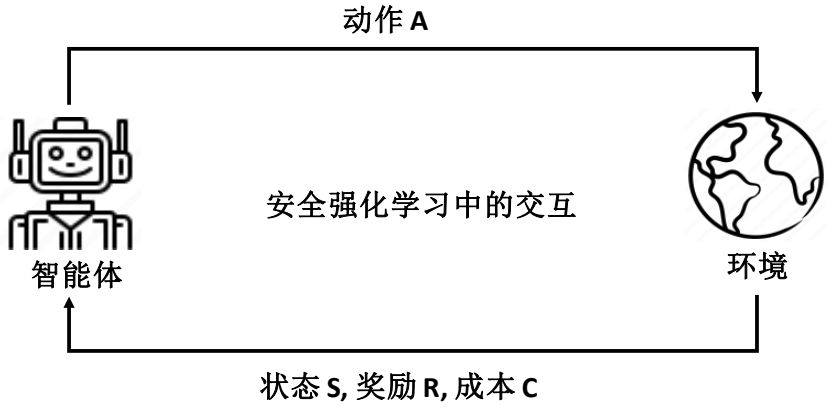

图 2-6 安全强化学习下的智能体与环境交互

Figure 2-6 Agents interact with environment in safe reinforcement learning

安全强化学习中的安全一词有多重定义。García 和 Fernández[48]认为，安全强化学习是考虑安全或风险等概念的强化学习。Hans 等人[49]认为，人类需要

将环境状态标记为“安全”或“不安全”，并且如果“智能体从不踏足不安全的状态”，应该将智能体视为具有“安全性”。总而言之，安全强化学习方法旨在优化成本目标，避免对手攻击，改善不良情况，降低风险以及控制器的安全等[50,51]。

安全强化学习中智能体与环境的交互过程与强化学习相似，但环境对智能体给予的反馈略有不同。如图 2-6 所示。

安全强化学习中的时间步 t、幕、奖励 R、环境、状态 S 等基本构成与强化学习相同，此处不再赘述。

成本 C 即 cost C，是用于衡量当前智能体所选动作安全程度的标量，与奖励相似，可为正值或负值。成本的具体含义在众多研究中存在诸多差异，本章所选用其中被最广泛认可的含义。即成本的数值与智能体当前动作的危险程度呈正相关，智能体当前的动作越危险，环境反馈的成本值越大。反之，环境反馈的成本值越小。对于长序列求解问题，定义回报 G_t^c 用以衡量整条轨迹的累计成本。

$$G_t^c \doteq C_{t+1} + \gamma C_{t+2} + \gamma^2 C_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k C_{t+k+1} \tag{2-12}$$

安全强化学习的优化目标是 $\pi^* = \arg\max \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} G_t^{\mathrm{r}} \mid s_0 = s\right]$，s.t. 约束条件，要求在满足一定约束的条件下，找到获得最大累积回报的策略。根据约束条件不同，所得到的最优策略也各不相同。

2.2.1 安全强化学习的约束

常见的约束条件分为三类，按照约束强度从强到弱分别为，硬约束、概率约束、软约束。在表述约束条件时，常用 b 表示安全阈值。

$$C_i(\pi) = \sum_t C_i(s_t, a_t, s_{t+1}) < b_i \tag{2-13}$$

硬约束是指在整条轨迹中，智能体所发出的每一个动作都能满足安全约束。硬约束要求智能体保证在部署时，必须做出绝对安全的决策。此类约束对安全约束要求严格，智能体习得策略可能会偏向“中庸”。

$$C_i(\pi) = P\left(\sum_t C_i(s_t, a_t, s_{t+1}) \le b_i\right) \ge \eta_i \tag{2-14}$$

概率约束是指在整条轨迹中，智能体所选择的动作在一定概率下违反安全约束。概率约束允许智能体在长序列决策中，有部分动作不满足安全约束。此类约束可以借助概率η_i调整安全约束的严格程度，多用于金融领域。

$$C_i(\pi)=\mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t C_i(s_t,a_t,s_{t+1})\right]\leq b_i \tag{2-15}$$

软约束是指在整条轨迹中，智能体只需要满足成本回报的期望小于某个阈值即可。软约束并未对单个动作进行严格约束，具有一定的灵活性，通常用于多约束的优化问题。

2.2.2 安全强化学习与强化学习的区别

安全强化学习自身特性带来的约束使其求解结果与强化学习可能完全不同。强化学习所使用的马尔可夫决策过程，保证智能体寻找最优策略时无需考虑任何额外条件，理论上一定存在最优策略。而安全强化学习所用受限的马尔可夫决策过程，在马尔可夫决策过程中加入了安全约束，这导致安全强化学习不存在统一的固定最佳策略。

安全约束的多样性阻碍了安全强化学习收敛至统一的固定最佳策略。各异的决策偏好催生了多样的安全约束，例如优先分配、风险把控等，最终影响智能体策略收敛方向。在不同的安全约束条件下，智能体在决策前会依据约束要求对动作进行判别，从而筛选出符合安全要求的动作候选集。即使处于相同的状态，由于安全约束的不同，智能体的决策路径也会出现差异，并最终收敛于不同的方向。

随机策略的必要性阻碍了安全强化学习收敛至统一的固定最佳策略。随机策略支持智能体在决策时，根据概率设定随机挑选动作完成决策。类比于确定性策略，随机策略允许智能体在部署时根据约束要求做出更为灵活的决策。例如，确定性策略为满足约束，会放弃回报优异的动作。随机策略可以通过对概率做出调整，最终产生回报优异且满足期望约束的决策。

成本约束的累积变化阻碍了安全强化学习收敛至统一的固定最佳策略。使用软约束作为安全约束的长序列决策中，成本会随着序列长度不断累积。同一条轨迹不同时间步的相同状态下，智能体为了满足约束，无法始终选择当前回报最高的最优动作。

2.3 在线强化学习

在线强化学习中智能体与环境交互是实时发生的，无需依赖额外数据集。当环境发生变化时，在线强化学习展现出了强大的灵活适应能力。智能体在发出动作后，根据环境的反馈动态调整值函数或策略以适应环境。具体而言，智能体会在接收到新环境的反馈后，通过贝尔曼方程迭代覆盖过往认知。并且对值函数进行微调使其拟合现有经验，借助调整后的值函数对策略进行纠正。在线强化学习适用于资源管理[52]、金融交易[53,54]等领域，便于根据环境的反馈（资源余量、成交数据），动态调整策略偏好。

2.3.1 在线强化学习优势

在线强化学习发展迅速，主要得益于：

在线强化学习具有强大的实时响应能力。在线强化学习支持智能体与环境进行实时交互，并根据反馈实时学习和修正策略。当数据中出现误差或随机噪声时，智能体可以通过与环境的实时交互，对判断失误的策略进行即时修正。在线智能体的实时相应能力有效避免误差沿贝尔曼方程不断积累，从而避免对后续决策的消极影响。

在线强化学习具有处理不确定性的能力。在线强化学习鼓励智能体通过利用与探索，对不确定的环境进行逐步建模，最终找到有效策略。例如在棋牌类游戏中，自我智能体在本幕开始时，仅能对自己的牌面进行建模。当通过不断的利用和探索后，可以通过对方行为计算其牌面，重新对环境进行精确建模。

在线强化学习鼓励智能体进行探索。在线强化学习为智能体提供了在训练过程中持续探索的能力，使其能够突破次优策略的局限，进而寻求并实现最优策略。在数学问题求解领域[55]，人类的思维模式往往受限于过往经验和既定的解题框架。在线智能体允许借助探索对公式进行灵活的嵌套与组合，突破传统思维的束缚，产生创新性的解决方案。

2.3.2 在线强化学习挑战

然而在线强化学习还存在一些挑战：

在线强化学习样本效率低。强化学习常见误区之一是对智能体训练收敛所需样本量和耗时过于乐观。在既定样本数量的需求下，在线智能体只能通过不断与环境进行实时交互以采集样本。在交互过程中，在线智能体还需对利用和

探索做出权衡。在找到最优策略之前，智能体所收集的样本良莠不齐。无效样本引起值函数的震荡，将误差引入策略之中。两者相互影响，共同阻碍智能体的快速收敛。

在线强化学习计算资源需求大。在线的设定要求智能体与环境进行实时交互，这对计算资源提出了极高的需求。复杂多维状态空间下，策略的评估和更新需要通过大量计算进行迭代。例如战胜顶尖人类棋手和其他优秀 AI 的 AlphaGo-Zero[56]，其训练过程对计算资源需求巨大，需要在四个 TPU（Tensor Processing Unit 张量处理单元，谷歌开发专用于进行机器学习和深度学习的芯片）上进行超过 40 天的训练。

在线强化学习需要注意训练安全问题。在线交互的设定虽然支持智能体灵活适应多变的环境，但在训练过程中可能会对环境或自身造成损失。例如在进行实时交互时，由于智能体无法判定未知动作的安全性，可能在探索未知动作时违反安全约束。

2.4 离线强化学习

离线强化学习中智能体不与环境发生交互，而是依赖于已有数据集进行训练。离线强化学习与异策略学习相似，但其本质完全不同。离线强化学习强调智能体仅使用已有数据集进行学习，且所使用的数据集是静态数据集，在训练过程中不会发生任何扩充或缩减操作。而异策略学习强调智能体的目标策略和行为策略不同，目标策略可以使用当前行为策略所产生的实时数据或历史数据，并且行为策略可以随时进行优化和调整增强数据的丰富度。离线强化学习适用于难以实时采样的应用领域，如医疗手术、工业生产等。

2.4.1 离线强化学习优势

离线强化学习具有高效的数据利用率。离线强化学习支持智能体充分利用已有静态数据集，挖掘数据内部联系以提取有效信息。不仅节约了实时采样所需的计算资源和高耗时，还令已有数据集迸发出新的活力。例如，在金融领域，智能体通过对过往交易记录、大盘走向的深度分析，训练得到优质策略。

离线强化学习具有训练安全性。离线强化学习在训练时，无需与环境进行实时交互，规避了智能体陷入利用与探索困境的可能。智能体无需对未知动作进行探索，降低次优策略或不收敛策略带来的安全风险。例如，在自动驾驶中，

使用历史数据训练离线智能体，无需在行驶道路中尝试突然加减速等未知动作以获得反馈。

离线强化学习便于优化和迭代。离线强化学习算法在验证算法性能时，依靠离线数据集和离线集成环境进行实验。算法的优化和迭代仅需消耗计算资源，无需考虑实际环境的限制和影响。多个算法可在同一离线集成环境下进行实时效果对比，有效防范由物理环境的不确定性所产生的误差。

2.4.2 离线强化学习挑战

离线强化学习中的分布偏移问题。分布偏移是离线强化学习中的一个重要问题，是指训练数据和实际数据的分布之间的差异会导致离线训练的智能体在评估未知动作时给出过高的 Q 值。在离线强化学习中，智能体无法通过探索来纠正其错误的判断，令分布偏移问题变得更为致命。例如，离线训练的推荐系统在接收到新兴热点视频时，无法对其进行准确的判别，最终推荐给错误的目标用户。

离线强化学习过分依赖数据集。离线强化学习算法的数据来源有且仅有历史静态数据集，因此数据集的质量会影响到智能体的收敛效果。当所使用数据集有数据缺失、无效数据过多等问题时，智能体无法从数据集中捕捉到足够的有效信息，从而影响智能体对环境的建模和策略的学习。例如，随机策略所产生的数据集中包含过多无效数据，智能体很难从奖励稀疏的长轨迹序列数据中学习到有效的行为。

离线强化学习忽略部署安全。离线强化学习借助数据集训练有效规避在线的实时交互行为，然而，离线的设定也阻碍了智能体从实时交互中获得危险信号或警告。若数据集缺失对安全性约束的表达，离线强化学习在训练中会自然而然忽略策略的安全性要求，最终导致智能体在实际部署时的策略安全缺失问题。例如，未考虑部署安全的离线智能体在进行导航任务时，不会遵守交通法规以求更短耗时。

2.5 扩散模型

扩散模型[57]的灵感源于热力学系统的熵增原理，其本质是通过随机微分方程描述数据分布的渐进式扰动与恢复过程。扩散模型通过构建数据与噪声之间的动态转化系统，开创了生成式人工智能的确定性优化新范式。其核心架构模

拟了物理系统的能量耗散过程：在前向阶段，原始数据历经数百次确定性加噪步骤，逐步退化为各向同性高斯噪声；逆向阶段则通过神经网络学习噪声到数据的梯度场，实现从随机噪声到目标分布的精准重建。这种基于分数匹配的渐进式修正机制，解决了传统生成模型在模式覆盖与采样质量间的固有矛盾。

模型的训练过程基于随机微分方程的正反向时间演化理论。前向扩散通过预设的噪声调度策略，将数据分布线性退化为易处理的简单分布；逆向过程则利用 U-Net 架构的噪声预测网络，逐步剥离数据中的扰动信号。关键创新体现在训练目标的数学构造——通过最小化噪声预测误差而非直接拟合数据分布，系统能够隐式学习数据流形的几何特征，这种去中心化的优化策略显著提升了生成样本的多样性和保真度。

2.4.1 DDPM

前向过程（加噪过程）：DDPM[57]（Denoising Diffusion Probabilistic Models）的前向过程是一个逐步向数据中添加高斯噪声的过程。具体来说，从原始数据点 x_0 开始，经过 T 步的迭代，每一步都根据预设的噪声方差 β_t，将当前数据 x_t 转换为 x_{t+1}，最终得到的 x_T 接近于标准正态分布；反向过程（去噪过程）：反向过程是前向过程的逆操作，目标是从噪声中恢复出原始数据。这个过程同样是一个马尔可夫链，每一步都通过训练一个神经网络来预测并去除当前数据中的噪声，逐步重建原始数据。

损失函数：DDPM 的训练目标是最小化预测噪声与真实噪声之间的均方误差（MSE）。通过优化这个损失函数，模型可以学习到如何准确地去除噪声。

逐点采样：在反向过程中，每一步都基于前一步的结果进行采样，逐步减少噪声，直到得到最终的生成样本。这种方法虽然生成质量较高，但需要较多的计算资源和时间。

2.4.2 EDM

前向过程（加噪过程）：连续噪声调度与参数化。EDM[58]（Elucidating the Design Space of Diffusion-Based Generative Models） 的前向过程摒弃了 DDPM 的离散时间步，转而采用连续时间参数化，实现更精细的噪声控制。噪声强度参数化定义对数信噪比（Log-SNR）为：$\gamma(t) = -\log(\sigma_t^2)$，其中 σ_t 表示时刻 t 的噪

声标准差，通过设计 $\gamma(t)$ 的连续变化轨迹，实现噪声强度的灵活调整。前向加噪的直接扰动公式为：$x_t = x_0 + \sigma_t \cdot \epsilon,\ \epsilon \sim \mathcal{N}(0, \mathbf{I})$，根据预设的调度表动态调整，避免 DDPM 中固定的线性或余弦调度限制；反向过程（去噪过程）：反向过程是概率流 ODE 求解过程。EDM 将逆向过程视为连续时间的常微分方程，采用高阶求解器加速采样 $\frac{dx}{dt} = -\dot{\sigma}(t)\sigma(t)\nabla_x \log p_\theta(x;\sigma(t))$。通过二阶数值积分，将采样步数从 DDPM 的 1000 步压缩至 35 步，同时保持生成质量。

损失函数：基础目标仍为预测噪声的均方误差，但额外引入了权重函数 $\lambda(\sigma)$。EDM 设计为 $\mathbb{E}_{\sigma,x_0,\epsilon}\left[\lambda(\sigma)\| \epsilon - \epsilon_\theta(x_t,\sigma)\|^2\right]$，通过理论分析重新设计损失函数，平衡不同噪声水平的训练权重。其中，$x_t = x_0 + \sigma\epsilon,\ \epsilon \sim \mathcal{N}(0, \mathbf{I})$，且 $\lambda(\sigma) \propto \frac{1}{\sigma^2 + \sigma_{\text{data}}^2}$ 用于避免简单样本（极低或极高噪声）主导训练。

preconditioning 和损失权重：EDM 通过引入 σ 的跳跃连接来对模型进行 preconditioning，并对损失的加权方式和采样 σ 的分布进行了优化。

2.6 本章小结

本章首先介绍了强化学习的基本定义、主要构成、重要问题，然后阐述了安全强化学习的安全约束，并对安全强化学习与强化学习的区别进行解释。此外，分别论述了在线强化学习和离线强化学习的基本概念、优势和挑战。最后，介绍了一些热门生成式模型。

第3章　基于数据增强的双层离线强化学习框架

针对离线强化学习中的分布偏移问题，提出了一种基于数据增强的双层离线强化学习框架 Diffusion-DA。该框架通过提供低熵数据收敛层与高熵扩散探索层两个训练层，支持智能体通过感知不同构成的数据集对状态、动作空间进行复杂建模，从而缓解分布偏移问题。其中，低熵数据收敛层使用原始数据，帮助智能体在决策时贴合原始数据分布，以避免分布外偏移问题。高熵扩散探索层使用经过扩散生成的增强数据集，借助扩散模型对数据进行建模再表达。此外，D4RL 数据集中的大量实验证明，使用 Diffusion-DA 算法所得数据集训练的智能体在多个复杂离线任务中的优良表现。

3.1 引言

随着数字化转型和智能化发展，历史数据呈现出指数级的增长趋势。如何从大量历史数据中成功习得可用经验，成为人工智能热门课题之一。离线强化学习作为数据驱动型的人工智能领域之一，支持智能体从历史数据中找寻潜在的结构与关系[15]。具体而言，离线强化学习将智能体的训练数据来源限制在静态的历史数据集范围之内，但对历史数据集的构成并无明确约束。此性质不仅为离线强化学习提供了大量可用数据，并支持离线强化学习更易于完成从模拟环境到落地应用的转变，两者相辅相成，助力离线强化学习学习在基于历史数据进行决策的领域崭露头角。例如，推荐领域[10,11]通过用户过往交互数据学习个人偏好，进行个性化推荐；自然语言处理领域[59]通过语义学习和令牌分解，从大量历史数据中提取经验，生成准确且符合人类逻辑的答案。

然而，研究发现[14]，智能体仅依赖静态历史数据完成策略的学习和提取，在复杂决策问题中无法做出良好的决策。具体表现在智能体中的价值函数会对未知的状态动作对给出不符合实际的过高评价，从而影响决策效果。这是由于离线强化学习算法的价值函数仅对静态数据集中的数据进行正确的拟合，当数据集中的数据与实际部署数据不同时，智能体面对未知数据价值函数会做出错误评价。当智能体被允许进行在线交互时，可以通过探索与利用不断试错，对价值函数中的错误拟合根据环境反馈进行修正。然而，离线强化学习要求智能

体离线不交互，仅使用静态数据集进行训练，剥夺了智能体通过交互消除价值函数过估误差的可能。

已有研究从多角度提出解决方案取得显著成果，但分布偏移问题仍然阻碍离线强化学习的进一步发展。（1）对智能体的动作空间进行约束[14,21]。该类算法通过修改价值更新方程或策略提取方程对动作施加一定约束，如动作相似度约束或惩罚不相似动作，鼓励智能体在决策时偏好静态数据集中已有动作。通过将智能体的决策结果限制在已知动作空间下，避免评估未知动作带来的价值函数高估问题，但也错过未知动作中的优质动作。（2）通过内置环境模型对智能体的决策结果进行修正[32,33,36]。该类算法考虑额外设置模型用于拟合环境对智能体动作的反馈，智能体通过反馈对策略进行更新。当环境模型能够完全拟合真实环境时，即可为离线智能体修正价值函数的误差。但考虑到随机事件、环境复杂、数据维度高等因素影响，环境模型往往建模难度大，难以对真实环境完全拟合。（3）智能体采用生成式决策[24,27,38,60]。随着生成式大模型的发展，离线强化学习领域考虑使用生成式模型为智能体骨架进行决策。智能体依靠生成式模型，对历史数据序列进行建模，生成与原有数据相似但更具多样性的优质数据。进行长序列预测时，生成式决策在连续多步决策中可能会产生误差的累积，从而影响最终生成的轨迹质量。

基于已有研究，考虑设计一个通用离线强化学习框架，旨在为各类算法提供避免分布偏移的有效方案。借助扩散模型的数据生成能力，提出了基于数据增强的双层离线强化学习框架 Diffusion-DA，如图 3-1 所示。为了确保智能体在训练过程中稳定收敛的同时避免分布偏移问题，框架根据训练所用数据的信息熵高低，提供低熵数据收敛层、高熵扩散探索层两个训练层。低熵数据收敛层采用原始数据集对算法进行训练，鼓励算法在稳定的原始数据集中捕捉原始数据分布，快速收敛。随后算法会经过高熵扩散探索层，根据扩散模型生成的扩散数据集进行伪探索。扩散数据集以原始数据集为摹本，对数据片段进行分别增强，具有更高的不确定性。这项工作的主要贡献如下：

（1）Diffusion-DA 提供了一个通用的离线强化学习框架，适配各类离线强化学习算法，帮助算法在原有基础上避免外推误差的影响。

（2）使用扩散模型支持离线算法进行伪探索。离线算法固有离线不交互特

性使算法跳出探索与利用困境，但是也限制了离线算法的动作空间。借助扩散模型的强大表达能力，对数据进行扩充。

（3）在对比实验中，Diffusion-DA 性能优良，多个实验结果体现了 Diffusion-DA 设计的合理性和有效性。

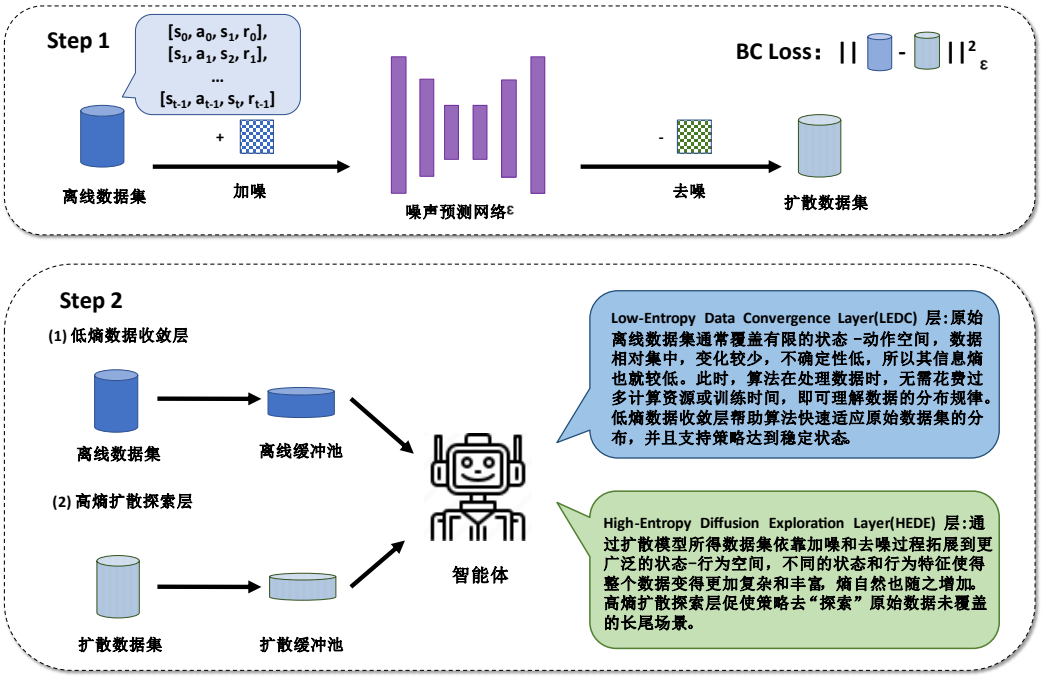

图 3-1 Diffusion-DA 框架图

Figure 3-1 The framework of Diffusion-DA

3.2 方案设计

离线强化学习通常借助马尔可夫决策过程进行建模，其中智能体在每个时间步 t 根据当前状态 s_t 进行决策，选择要执行的动作 a_t，最终会收到环境的反馈：下一个状态 s_{t+1} 和奖励 r_t。轨迹 $\tau=\{s_0,a_0,s_1,r_0,,...,s_t,a_t,s_{t+1},r_t,,...\}$ 记录一幕，从状态 s_0 开始，到最终此幕结束。智能体的优化目标是在一幕之中，获得更多的累积奖励 $G^r=\sum_{t=0}^{\infty}\gamma^t r_t$。其中 γ 是折扣率，用于表示奖励 r_t 对当前状态 s_t 或当前状态动作对 (s_t,a_t) 的影响随时间 t 的变化。在离线强化学习算法中，智能体从静态数据集 $\mathcal{B}(s_t,a_t,s_{t+1},r_t)$ 中采样对策略 $\pi(a|s)$ 进行学习和更新。离线强化学习算法使用动作价值函数 $Q^r(s_t,\mathbf{a}_t)=\mathbb{E}_{\tau\sim\pi}[G_r|s_t,a_t]$ 评价当前动作优劣，用状态价值函数 $V^r(s_t)=\mathbb{E}_{\tau\sim\pi}[G_r|s_t]$ 评价当前状态优劣。基于数据增强的双层离线强化学习

框架的目标是，帮助只能找到最优策略 π^*。

$$\pi^* = \arg\max_{\pi} G^r \tag{3-1}$$

为方便叙述，文中以基于 TD3 算法[61]改进的典型离线强化学习算法 TD3-BC[62]为例，TD3 算法在进行决策时，贪婪的选择当前价值函数最高的动作，如（3-2）所示。

$$\pi = \arg\max \mathbb{E}_{(s,a)\sim D}\left[Q(s,\pi(s))\right] \tag{3-2}$$

TD3-BC 借助演员-评论家框架，使用两个演员用于对当前动作和下一个动作分别做出决策，并且每个演员对应两个评论家以解决 Q 值高估问题。此外，为避免离线强化学习中的分布偏移问题，TD3-BC 在原有策略提取方程中添加了行为克隆的正则项。TD3-BC 算法框架如图 3-2 所示。

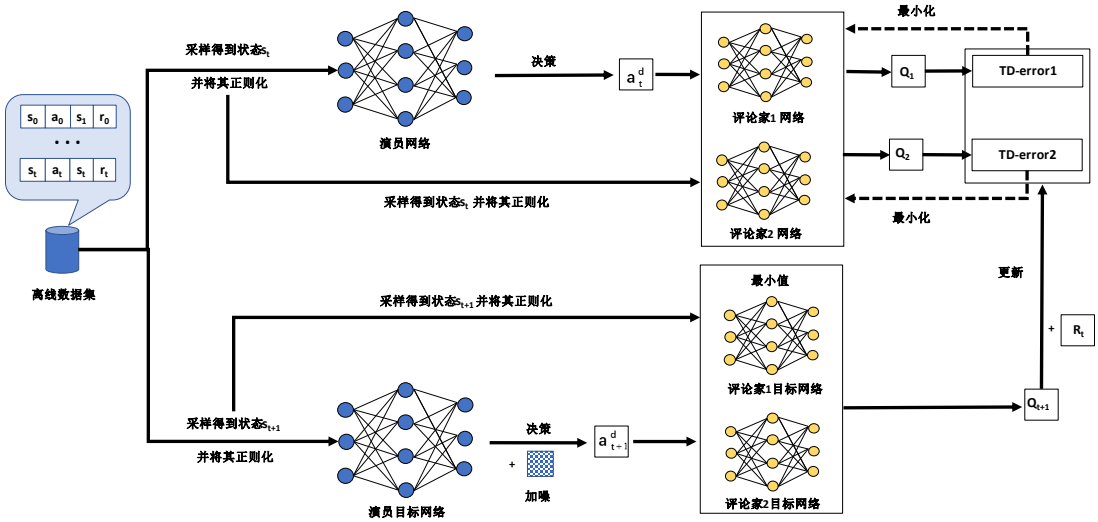

图 3-2 TD3-BC 算法框架图

Figure 3-2 Framework diagram of the TD3-BC algorithm

TD3-BC 算法作为离线强化学习算法，在 TD3 算法的策略提取方程的基础上，对策略施加动作相似度的正则项。其策略提取公式，如（3-3）所示。

$$\pi = \arg\max \mathbb{E}\left[\lambda Q(s,\pi(s)) - (\pi(s)-a)^2\right],$$
$$\lambda = \frac{\alpha}{\frac{1}{N}\sum_{s_t,a_t}|Q(s_t,a_t)|} \tag{3-3}$$

其中 λ 为 $Q(s_t,a_t)$ 值函数的权重，用以平衡 $Q(s_t,a_t)$ 值与正则项之间的关系。

3.2.1 低熵数据收敛层

原始静态数据中数据分布相对明确和集中，具有信息熵较低的特点，故第一层训练层为低熵数据收敛层。低熵数据收敛层支持 TD3-BC 算法根据信息熵较低的原始静态数据集 $B(s_t,a_t,s_{t+1},r_t)$ 进行学习，达到稳定状态。在此训练层中，由于 TD3-BC 算法处于离线的训练状态下，智能体可感知的数据被限制在数据集覆盖范围内，数据分布相对明确和集中。动作价值函数 $Q^{\pi}(s_t,a_t)$ 通过对数据集中的真实片段进行采样，拟合数据集内数据的真实 Q 值。

$$\begin{aligned} Q^{\pi}(s_t,a_t)=\mathbb{E}_{\pi}[r_t+\gamma Q^{\pi}(s_{t+1},a_{t+1}) \mid S=s_t,A=a_t], \\ \text{s.t. } (s_t,a_t,r_t,s_{t+1},a_{t+1})\sim\mathcal{B} \end{aligned} \tag{3-4}$$

然后 TD3-BC 算法根据 $Q^{\pi}(s_t,a_t)$ 对动作状态对的判断，形成初步策略。

$$\begin{aligned} \pi=\arg\max \mathbb{E}_{(s_t,a_t)\sim\mathcal{B}}\left[\lambda\, Q^{\pi}(s,\pi(s))-(\pi(s)-a)^2\right], \\ \lambda=\frac{\alpha}{\frac{1}{N}\sum_{s_t,a_t}\mid Q^{\pi}(s_t,a_t)\mid} \end{aligned} \tag{3-5}$$

3.2.2 高熵扩散探索层

低熵数据收敛层帮助 TD3-BC 算法迅速收敛至稳定，但由于仅能对数据集分布内动作做出正确判断，所以经过此训练层训练的 TD3-BC 算法无法正确评估分布外动作，使策略趋向于次优。故考虑添加第二个训练层，即高熵扩散探索层。

（1）数据增强器

高熵扩散探索层使用扩散数据集帮助 TD3-BC 算法进行伪探索，通过提供额外数据支持算法在离线训练中接触未知动作，其网络架构如图 3-3 所示。为了避免扩散数据集为算法的训练引入更多误差，Diffusion-DA 要求扩散数据集与原始数据集分布相似并且具备一定多样性。基于此，Diffusion-DA 为高熵扩散探索层设计了作用于数据片段的数据增强器 D_{ω}。具体而言，Diffusion-DA 从原始数据集 $\mathcal{B}(...,s_t,a_t,s_{t+1},r_t,...)$ 中采样得到数据片段 $F_t=\{s_t,a_t,s_{t+1},r_t\}$ 作为训练样本，向训练样本 F_t 中添加噪声直至其变为纯噪声，并使用数据增强器 D_{ω} 预测所添加的噪声。

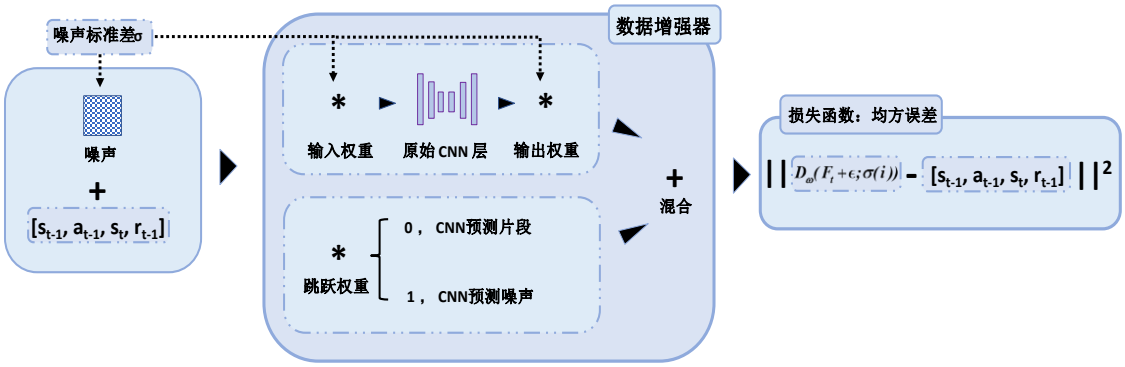

图 3-3 数据增强器网络架构

Figure 3-3 Data Enhancer Network Architecture

由于数据片段 $F_t=\{s_t,a_t,s_{t+1},r_t\}$ 中隐含复杂的状态转移矩阵和奖励函数，为了避免随机性带来的误差，考虑使用基于概率流常微分方程的方式而不是传统的随机微分方程的离散化形式对数据片段 F_t 随扩散步 i 的变化进行建模。

$$dF_t=\left[\frac{\dot{s}(i)}{s(i)}F_t-s(i)^2\dot{\sigma}(i)\sigma(i)\nabla_{F_t}\log p\left(\frac{F_t}{s(i)};\sigma(i)\right)\right]di \tag{3-6}$$

其中 $s(i)$ 是用于调整数据片段尺度的缩放因子，$\dot{s}(i)$ 是 $s(i)$ 关于扩散步 i 的导数，$\sigma(i)$ 是用于表示噪声强度的噪声标准差，$\dot{\sigma}(i)$ 是 $\sigma(i)$ 关于扩散步 i 的导数。$p\left(\frac{F_t}{s(i)};\sigma(i)\right)$ 是扩散步 i 时的数据片段分布，$\nabla_{F_t}\log p(.)$ 为分数方程，用于表示如何移动数据片段使其出现的概率增加。$\frac{\dot{s}(i)}{s(i)}F_t$ 用于控制数据片段的变化趋势，而 $s(i)^2\dot{\sigma}(i)\sigma(i)\nabla_{F_t}\log p\left(\frac{F_t}{s(i)};\sigma(i)\right)$ 引导生成的数据样本逐渐逼近真实的数据样本分布。

受到 EDM[58]和 SYNTHER[63]中启发，考虑从高噪声的初始状态，通过减小噪声标准差 $\sigma(i)$ 和沿着概率密度增加最快的方向（即对数梯度方向）移动，生成低噪声的清晰数据片段。

$$dF_t=\left[-\dot{\sigma}(i)\sigma(i)\nabla_{F_t}\log p\left(F_t;\sigma(i)\right)\right]di \tag{3-7}$$

其中 $\nabla_{F_t}\log p\left(F_t;\sigma(i)\right)=\frac{(D_\omega(F_t;\sigma(i))-F_t)}{\sigma(i)^2}$，最终数据增强器 D_ω 通过下式进行训练，且 ϵ 为人为添加的已知噪声。

$$\mathbb{E}_{F_t\sim p_{data}}\mathbb{E}_{\epsilon\sim\mathcal{N}(\mathbf{0},\sigma^2\mathbf{1})}\left\|D_\omega(F_t+\epsilon;\sigma(i))-F_t\right\|_2^2 \tag{3-8}$$

（2）扩散数据集的获取与 TD3-BC 算法的训练

数据增强器经由（3.8）式训练后，即可为高熵扩散探索层生成所需扩散数据集 $\hat{\mathcal{B}}$。扩散数据集通过数据增强器生成。凭借数据增强器的拟合与表征能力，该数据集逐渐呈现出有别于原始数据的多样性。同时，数据增强器以数据片段的相似性为约束条件，将扩散数据集与原始数据集的相似度限定在可接受范围内。

$$\hat{\mathcal{B}}=(...,s_t^g,a_t^g,s_{t+1}^g,r_t^g,...)\quad \text{s.t. }\{(s_t^g,a_t^g,s_{t+1}^g,r_t^g)\sim D_\omega(\epsilon;\sigma(i))\}_{t=1}^n \tag{3-9}$$

TD3-BC 算法依靠扩散数据集 $\hat{\mathcal{B}}$ 对训练过的状态价值函数 Q^π 进行微调。扩散数据集支持 TD3-BC 算法通过学习原始数据集外的数据片段（$s_t^g,a_t^g,s_{t+1}^g,r_t^g$），对曾经的未知状态动作对 (s_t^g,a_t^g) 做出正确估值。

$$\begin{gathered}Q^\pi(s_t^g,a_t^g)=\mathbb{E}_\pi[r_t^g+\gamma Q^\pi(s_{t+1}^g,a_{t+1}^g)\mid S=s_t^g,A=a_t^g],\\ \text{s.t. }(s_t^g,a_t^g)\sim\hat{\mathcal{B}}\end{gathered} \tag{3-10}$$

最终，TD3-BC 算法根据状态价值函数 Q^π 对动作状态对的判断，形成策略。

$$\begin{gathered}\pi=\arg\max\mathbb{E}_{(s_t,a_t)\sim B\ or\ \hat{\mathcal{B}}}\left[\lambda\,Q^\pi(s,\pi(s))-(\pi(s)-a)^2\right],\\ \lambda=\frac{\alpha}{\frac{1}{N}\sum_{s_t,a_t}|\,Q^\pi(s_t,a_t)\,|}\end{gathered} \tag{3-11}$$

TD3-BC 算法在基于数据增强的双层离线强化学习框架中的训练过程如框架 1 所示。

表 3-1 基于数据增强的双层离线强化学习框架

Table 3-1 A two-layer offline reinforcement learning framework based on data augmentation

框架 1 基于数据增强的双层离线强化学习框架
需要:离线数据集 $\mathcal{B}$;
初始化: 数据增强器 D_ω,动作价值网络 Q^π;
1： 初始化 TD3-BC 算法
2： # 低熵数据收敛层训练
3： **for** 每次迭代 **do**:
4： 随机采样 $(s_t, a_t, s_{t+1}, r_t) \sim \mathcal{B}$
5： 根据（3-4）式和（3-5）式使用离线数据集 $\mathcal{B}$ 对动作价值网络 Q^π 进行更新
6： **end for**
7： # 高熵数据集探索层训练
8： 随机采样得到数据片段 $F_t = \{s_t, a_t, s_{t+1}, r_t\}$ s.t. $(s_t, a_t, s_{t+1}, r_t) \sim \mathcal{B}$
9： 根据 （3-8）式训练数据增强器 D_ω
10：根据（3-9）式获得扩散数据集 $\hat{\mathcal{B}}$
11：**for** 每次迭代 **do**:
12： 随机采样 $(s_t^g, a_t^g, s_{t+1}^g, r_t^g) \sim \hat{\mathcal{B}}$
13： 依照（3-10）式和（3-11）式使用离线数据集 $\mathcal{B}$ 对动作价值网络 Q^π 进行更新
14: **end for**

3.3 实验评估

3.3.1 数据集及实验环境

实验所用数据集为 Datasets for Deep Data-Driven Reinforcement Learning（D4RL）[64]数据集，由 DeepMind 于 2020 年创建。为了拟合现实环境中的复杂数据，该数据集涉猎多个领域和任务，可为算法提供不同构成的非独立同分布的复杂数据子集。如，medium 数据集由对未训练结束时的算法进行随机采样所得到的一百万条数据样本组成；medium-replay 数据集则是先将 Soft Actor-Critic 算法智能体通过在线训练使其达到中等性能，然后将其在训练期间缓冲区中所有数据进行随机采样所得数据集。

实验所用 PC 机配置为 NVIDIA Corporation、64G 内存、13 代英特尔 i5-13600KF 处理器、64 位 Ubuntu20.04.6 LTS 操作系统作为实验平台。

实验使用多个实验环境对算法性能进行测试，如 MUJOCO[65]、Gymnasium[66]等。其中，MUJOCO 旨在促进机器人控制、生物力学等需模拟多关节铰接结构与环境作用过程的领域的开发。此外，MUJOCO 提供 GUI 交互式可视化界面用于分析记录控制数据。如图 3-4 所示。

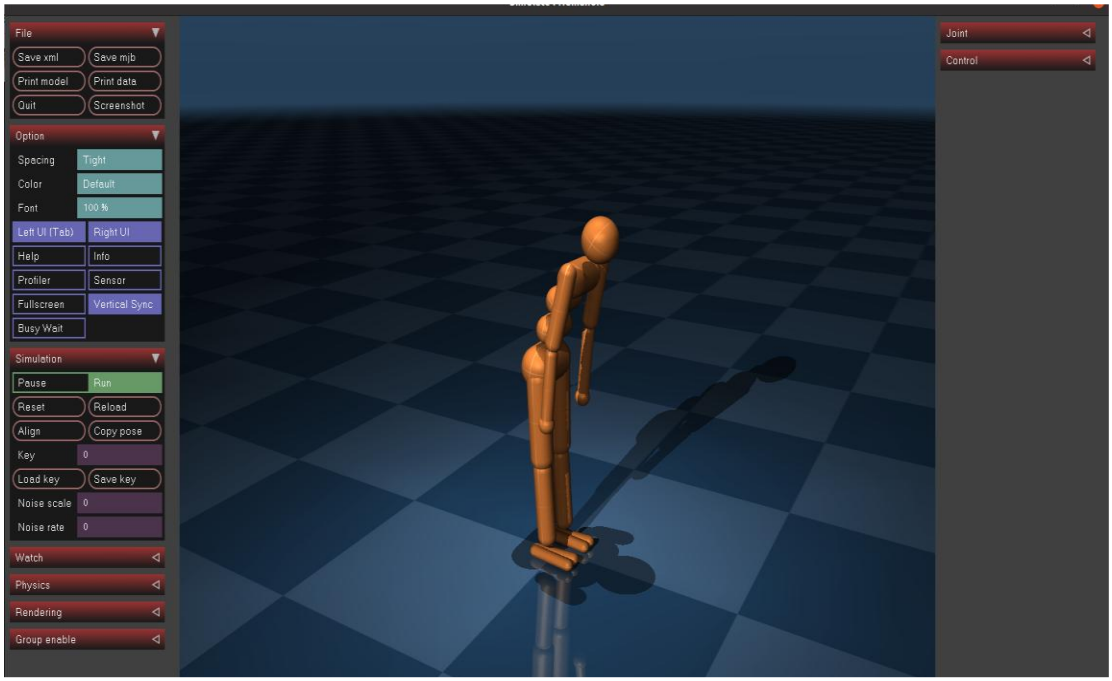

图 3-4 MUJOCO 可视化界面

Figure 3-4 MUJOCO visual interface

3.3.2 实验评估指标

离线强化学习的优化目标是令智能体找到累积奖励最大的策略或动作，故采用奖励作为算法的评估指标。此外，为捕捉算法的性能变化，对算法共进行 200 次评估，每次评估使用 10 幕平均奖励作为评估数据。

实验性能对比中使用了多个基线算法进行比较。IQL 算法[28]使用单步规划的思想使用期望回归的方式拟合 Q 值函数以避免引入分布外动作，并借助优势加权行为克隆完成策略的提取；TD3-BC 算法[62]以 TD3 算法为基础，额外设置了行为克隆项对策略更新进行约束，并对数据中的状态进行了归一化处理；CQL 算法[14]通过对数据集分布外的动作施加惩罚，从而引导策略在学习时偏向学习数据集内的数据。AWAC 算法[67]根据动作状态对的优势值函数对策略进行加权，令策略关注当前状态下，高于平均动作价值函数的优质动作。

3.3.3 基准性能对比

对比试验中，将各离线强化学习算法使用原始离线数据集训练后所得结果，作为其基线性能结果。此外，将 IQL 算法与 TD3-BC 算法使用 Diffusion-DA 框架训练（依次使用低熵数据收敛层、高熵扩散探索层对算法进行训练）所得结果，作为 Diffusion-DA +IQL 和 Diffusion-DA +TD3-BC 的性能结果。

表 3-2 Diffusion-DA 及基线性能对比结果

Table 3-2 Diffusion-DA and baseline performance comparison results

Task	Diffusion-DA +IQL	Diffusion-DA +TD3-BC	BC
halfcheetah-medium-v2	**49.86**	49.79	43.64
halfcheetah-medium-replay-v2	**46.66**	46.37	40.52
walker2d-medium-v2	**88.04**	87.31	80.64
walker2d-medium-replay-v2	**95.18**	94.58	54.41
hopper-medium-v2	76.82	75.36	69.04
hopper-medium-replay-v2	**102.48**	100.58	77.52
Average	76.51	75.67	60.96

表 3-3 Diffusion-DA 及基线性能对比结果（续表）

Table 3-3 Diffusion-DA and baseline performance comparison results(Continued)

IQL	TD3-BC	CQL	AWAC
48.55	48.23	47.31	49.53
44.83	45.43	46.49	46.43
87.68	86.57	84.87	84.82
89.02	91.36	89.85	87.51
77.37	71.44	71.87	**98.11**
101.45	98.52	101.11	100.95
74.82	73.59	73.58	**77.89**

如表 3-2 和表 3-3 所示，通过对隐式策略优化的 IQL 和施加行为克隆约束的 TD3-BC 两类典型算法进行训练，所得 Diffusion-DA +IQL 和 Diffusion-DA +TD3-BC 在众多基线算法中表现优异。虽然在 hopper-medium-v2 和 Average 的性能比较中，AWAC 略优于 Diffusion-DA +IQL。但在其他数据集下，AWAC 性能均劣于 Diffusion-DA +IQL，即 Diffusion-DA +IQL 具有更强的泛化性。

Diffusion-DA 通过提供不同构成的数据集对算法给予支持，数据增强后的样本支持算法在原始数据分布边缘进行伪探索以拓展可感知数据空间，而原始数据帮助算法在随机起始点找到策略优化方向稳定收敛。两者相辅相成，帮助算法在决策时展现更优性能。BC 算法在训练时，仅考虑与数据集中数据分布的相关性，并未对分布偏移采取任何措施。故当 BC 算法采用不同质量的数据集进行训练时，表现出的性能差异巨大；TD3-BC 算法虽然在 BC 算法中提出改进，但仍然依赖于原始数据集质量；IQL 算法和 AWAC 算法通过隐式约束或重要性采样对策略进行优化，但两者的网络架构导致其无法满足拟合高维复杂数据的策略表达需求。CQL 算法面对分布偏移采用保守的思想对值函数进行更新，然

而惩罚项导致值函数错失存在潜在奖励的优质未知动作。

3.3.4 消融性能对比

为证明 Diffusion-DA 算法有效性，对 Diffusion-DA 模块进行拆解并重组。其中 LEDC 是低熵数据收敛层，HEDE 是高熵数据扩散层。Mixed 为设置 random 函数，随机选择当前一幕使用低熵数据收敛层或高熵数据扩散层对算法进行训练。消融实验中以 TD3-BC 算法为例，使用不同框架与其进行组合以求解以下问题。（1）Diffusion-DA 中是否存在冗余组件（2）Diffusion-DA 框架所提供训练顺序是否合理

表 3-4 Diffusion-DA 消融实验结果

Table 3-4 Ablation experiment results of Diffusion-DA

Task	Diffusion-DA	Mixed
halfcheetah-medium-v2	**49.79**	48.86
halfcheetah-medium-replay-v2	46.37	**46.51**
walker2d-medium-v2	**87.31**	87.15
walker2d-medium-replay-v2	**94.58**	94.07
hopper-medium-v2	**75.36**	72.48
hopper-medium-replay-v2	**100.58**	97.33
Average	**75.67**	74.4

表 3-5 Diffusion-DA 消融实验结果（续表）

Table 3-5 Ablation experiment results of Diffusion-DA(Continued)

Diffusion-DA except HEDE	Diffusion-DA except LEDC	LEDC + HEDE
48.23	49.43	49.3
45.43	46.22	46.48
86.57	87.28	86.22
91.36	94.25	89.88
71.44	74.36	75.16
98.52	98.33	100.27
73.59	74.98	74.55

消融实验结果如表 3 4 和表 3 5 所示，现对上述问题进行讨论。

（1）Diffusion-DA、Diffusion-DA except HEDE 和 Diffusion-DA except LEDC 性能表现证明，低熵数据收敛层和高熵数据扩散层均为 Diffusion-DA 提供了一定助力；Diffusion-DA except HEDE 通过帮助算法拟合原始数据分布支持，减少震荡稳定收敛。但仅使用原始数据集进行训练缺乏数据多样性，算法无法摆脱分布偏移的影响；Diffusion-DA except LEDC 通过对数据进行分段增强再表

达，支持算法完成离线下伪探索。但完全使用扩散数据集，在追求样本多样性的同时，需牺牲分布相似性。算法受到噪声或离群点影响，无法学习到精确的数据特征。此外，分析发现，三者性能与数据集质量具有一定相关性。

（2）通过对低熵数据收敛层和高熵数据扩散层进行拆解和交换，对 Diffusion-DA 的训练步骤进行分析。LEDC + HEDE 将两个训练层顺序交换，性能略低于 Diffusion-DA。LEDC 所用扩散数据集中存在一定噪声和离群点，导致算法混淆真实数据和噪声。在后续 HEDE 的训练中，优化器中的历史梯度已经受到扩散数据集影响，难以收敛至更优；Mixed 中将两个训练层所用数据集进行混合，随机选择当前训练所用训练层。在训练时，算法长期处于复杂混合数据特征中，导致模型出现可塑性损失[68]。Q 值函数对新数据的拟合能力受到影响，最终无法泛化到真实数据上。

3.3.5 数据质量分析

为验证 Diffusion-DA 的数据增强效果，将低熵数据收敛层和高熵数据扩散层中所用数据进行可视化分析。以 halfcheetah-medium-v2 环境为例，进行分析。其中蓝色为低熵数据收敛层所用数据，粉色为高熵数据扩散层，紫色为两种数据重叠部分。

halfcheetah-medium-v2 环境的状态空间为 17 维，动作空间为 6 维，奖励为 1 维。状态空间是[-inf, inf]的 float64 连续空间，动作空间是[-1, 1]的 float32 连续空间。奖励表示为前向奖励与控制成本之差，前向奖励定义为 $forward\ reward = w_{\text{forward}} \times \frac{dx}{dt}$，即 halfcheetah 沿 x 轴正方向（向右）移动时获得的激励，其中 w_{forward} 前向奖励权重默认为 1，dx 是动作前后尖端位移差，dt 为动作时间间隔默认 0.05。控制成本定义为 $ctrl\ cost = w_{\text{control}} \times \| \text{action} \|_2^2$，即对动作幅度的惩罚项（负奖励），其中 w_{control} 控制成本权重默认为 1。

此外，实验参数设置为采样批次 256、缓冲池大小 10000000、折扣率 0.99、评估次数 5000、原始算法 TD3-BC、最大时间步 1000000、评估频率 10、随机种子选用 0。

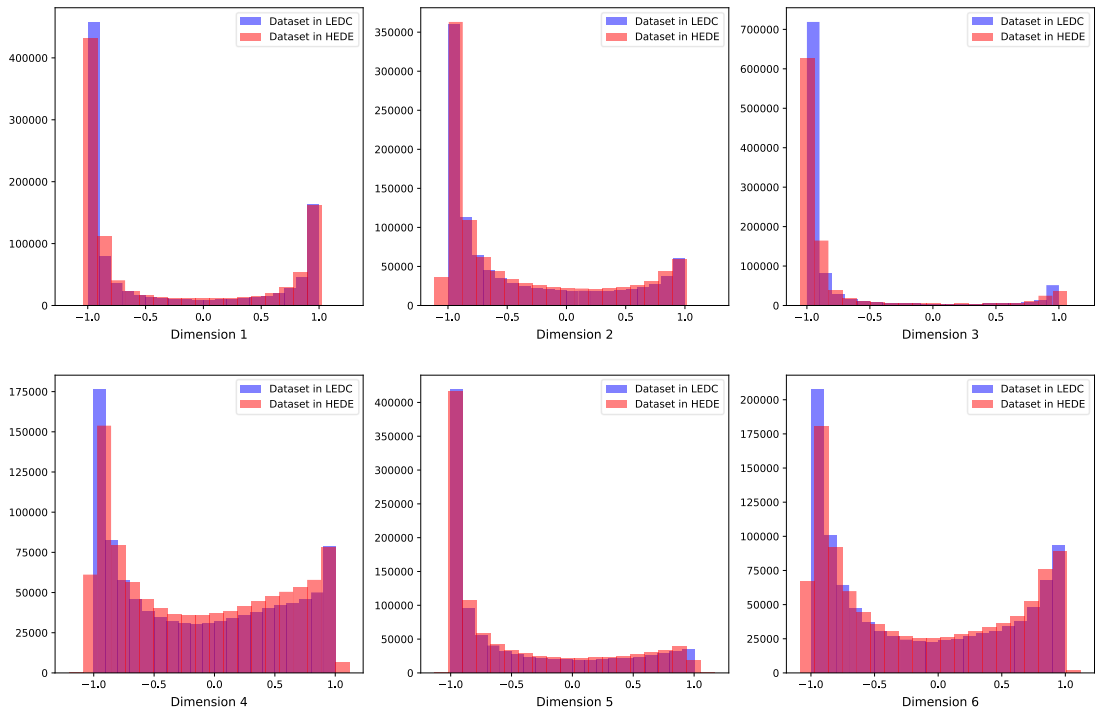

图 3-5 不同训练层所用数据的动作空间对比

Figure. 3-5 Comparison of action space of data used in different training layers

如图 3-5 所示，Diffusion-DA 对数据的动作空间不同维度均能在拟合大致分布的同时，对数据进行伪探索。动作空间不同维度中数据分布多为双峰分布，在 6 个维度上，两个训练层所用数据集的分布都显示出明显的偏斜，大多数数据集中在接近 0 的区域，随着数据值远离 0，频数迅速下降。两个数据集在所有维度上的分布都显示出高度的相似性，这表明在这些维度中两个数据集具有相似的特征或模式。尽管总体相似，但在某些维度上（如维度 3 和维度 6），HEDE 数据集的分布稍微分散，频数在两端略高，所以在这些维度上 HEDE 具有更广泛的特征或更大的变异性。

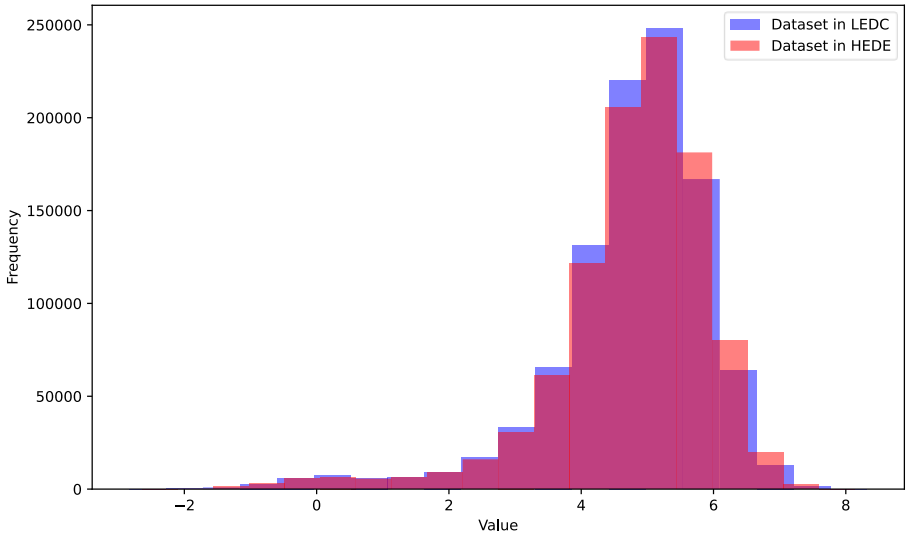

图 3-6 不同训练层所用数据的奖励空间对比

Figure. 3-6 Comparison of reward space of data used in different training layers

而具有单峰分布的奖励数据分布，Diffusion-DA 也能做到完美拟合，如图 3-6 所示。两个数据集的分布都集中在正值区域，且呈现出右偏（正偏）的特性，即数据值的尾部向右延伸，存在一些较大的数据值。

图 3-7 中为状态空间中不同维度间数据对比。状态空间比动作空间具有更多变丰富的数据特征，因此具有更加复杂的数据分布。大多数维度上的数据分布集中在 0 附近，且分布多为对称分布，但部分维度存在偏态。数据分布范围较广，数据点的离散程度较大。

综上所述，Diffusion-DA 对数据进行分段增强，通过随机噪声控制不同维度间具有一定差异。两个训练层数据分布的直方图重叠部分保证了数据与数据之间的相似性，约束数据中潜在状态转移矩阵符合真实规律。非重叠部分为支持算法进行离线下伪探索提供了更丰富的数据特征。具体而言，扩散数据集在原始数据分布边缘进行拓展，支持算法捕捉基础特征后，具有更高的泛化能力。

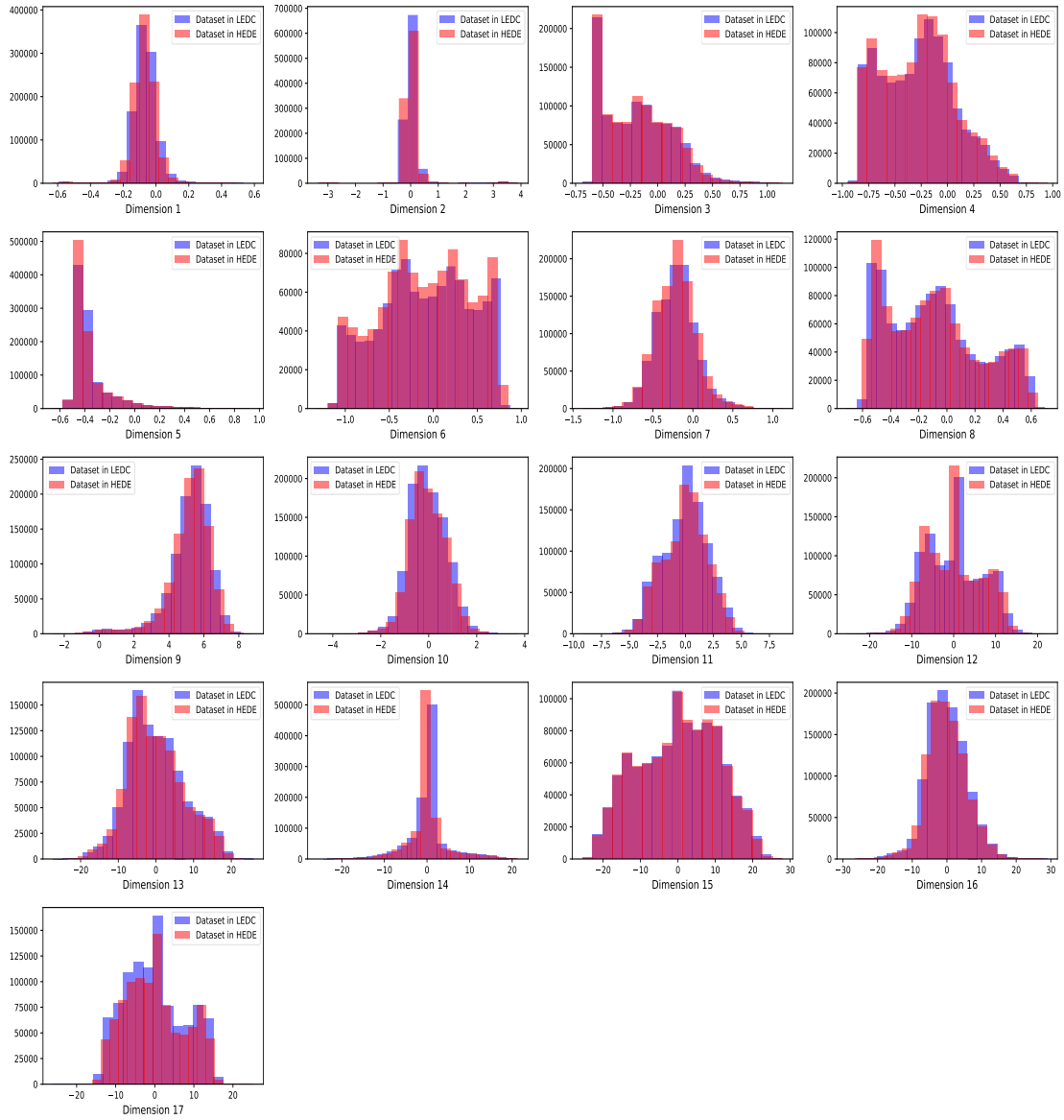

图 3-7 不同训练层所用数据的状态空间对比

Figure. 3-7 Comparison of observation space of data used in different training layers

3.4 本章小结

本章主要对基于数据增强的双层离线强化学习框架进行介绍。首先，介绍了离线强化学习的背景和意义，解释基于数据增强的双层离线强化学习框架的研究问题。然后对研究问题进行建模，并设计具体框架。该框架通过两个训练层来确保框架的有效性，低熵数据收敛层和高熵扩散探索层。Diffusion-DA 利用原始数据集帮助算法稳定收敛，然后借助扩散数据集协助算法进行伪探索以消弱分布偏移的影响。D4RL 中多个基线任务中，Diffusion-DA 均帮助算法表现出优异性能。

第4章　基于双 Q 引导的离线强化学习安全策略增强算法

针对离线数据集良莠不齐、忽略环境不确定性所带来的安全问题，提出一种基于双 Q 引导的离线强化学习安全策略增强算法——SPI（Safe Policy Improvement）。SPI 的核心思想是通过扩散模型引导生成满足安全约束的高奖励动作，在相同数据分布下采样得到安全的数据。

具体而言，SPI 算法首先对离线数据集进行重新映射，以增强其成本属性。随后，将扩散模型、动作值函数和安全值函数耦合至同一框架中，该框架能够促使生成的动作收敛到符合安全约束的分布。通过在 DSRL 基准下的广泛实验验证，SPI 在大多数任务中的表现均优于相关基线方法，证明 SPI 在提升策略安全性方面的有效性。

4.1 引言

离线强化学习是一种不与环境交互，利用离线数据集来学习策略的范式，其优化目标是获得最大化期望累积奖励的策略[69,70]。得益于从实时交互式在线采样到多批次数据集离线采样的转变，训练智能体学习策略变得更加高效和便利。这种转变推动了离线强化学习在某些采样成本高且耗时领域的广泛研究，例如：自动驾驶[71,72]、推荐系统[73,74]、竞技游戏[75,76]和机器人控制[77-79]。然而，在使用离线数据集进行策略学习时，通过最大化期望累积奖励训练的离线强化学习智能体在决策中不会考虑决策结果是否满足安全约束。例如，在导航任务中，机器狗为了获得与持续时间负相关的高奖励，不会对行进路径上的障碍物采取任何规避动作，而是尽快直奔前方。训练所得机器狗在实际部署时，会因为忽视安全性而造成危险事件的发生。

安全离线强化学习是一个很好的解决方案，它将安全约束纳入到离线强化学习的优化目标中。通过采用约束马尔可夫决策过程，安全离线强化学习可以确保在评估策略时同时考虑奖励和安全属性。然而，训练所使用的数据集可能来自多个具有不同目标任务的行为策略[80]，其复杂的构成来源很难确保所收集的高奖励动作能够满足安全约束。现有工作主要集中在如何在满足安全约束的情况下找到最优策略。然而，现有工作还具有一定局限性。（1）一种方法是直

接对策略本身施加约束[81-83]。例如，在现有的离线强化学习算法中引入拉格朗日乘数，对违反安全约束的行为施加惩罚。这将原来的约束最小化问题转化为无约束优化问题，同时对不满足约束的状态-动作对进行惩罚。但在处理高维复杂数据时，拉格朗日乘数法容易收敛到局部最优。（2）另一种方法是将正则化应用于价值函数，使其安全性评估更加保守[84,85]。通过修改贝尔曼更新方程，建立悲观近似的下限，这有助于减少高估的影响。然而，算法需要额外权衡策略的激进或保守的程度。过度保守或过度激进的策略会错过低成本且高回报的优质状态动作对。

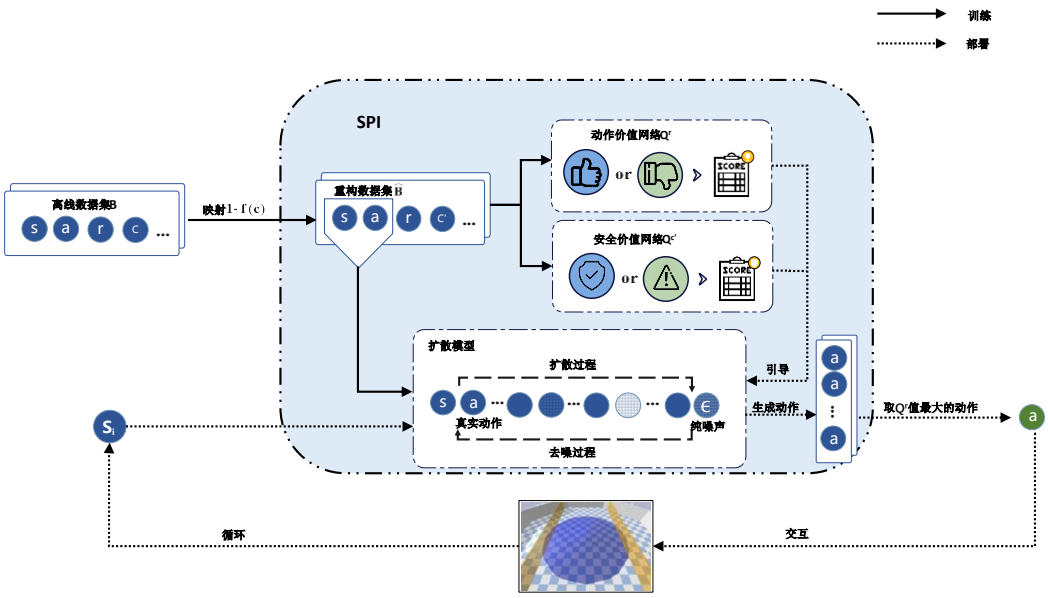

图 4-1 SPI 框架

Figure 4-1 The framework of SPI

针对上述问题，提出了通过扩散模型进行离线强化学习安全策略改进的算法 SPI，如图 4-1 所示。受 Diffuison-ql 算法[86,87]的启发，提出使用扩散概念扩展原始离线数据集，确保 SPI 通过三个过程收敛到安全、高回报的策略。

具体来说，首先将离线数据集的成本特征重新映射，以避免数据稀疏相关问题的发生。其次，应用扩散模型对处理后的数据分布进行去噪和拟合，并学习动作值函数和安全值函数。此外，将两个值函数的值纳入噪声扩散过程的优化目标中，确保生成的动作在满足安全约束的同时与离线数据集中的策略分布紧密一致。这项工作的主要贡献如下：

（1）SPI 通过离线数据集内的数据映射和处理促进安全策略的评估。考虑到安全值函数的时间累积效应，将数据集中的成本属性重新映射，以避免成本效益消失等事件。

（2）耦合扩散模型与动作值函数和安全值函数，使 SPI 具备平衡拟合分布和生成安全动作的能力。耦合操作可以支持 SPI 能够学习复杂的高斯分布并开发新的、多样化的安全动作。

（3）在众多与安全相关的离线强化学习任务中，SPI 在高奖励和安全性方面都表现出比基线算法更好的整体性能。

4.2 方案设计

安全离线强化学习与离线强化学习不同，使用受限的马尔可夫决策过程 $(\mathcal{S},\mathcal{A},\mathcal{P},r,c,\gamma)$ 进行建模。其中 $\mathcal{S}$ 表示状态的集合，$\mathcal{A}$ 是动作的集合，$\mathcal{P}(s_{t+1}\mid s_t,a_t)\to[0,1]$ 表示环境中的状态转移矩阵。此外，$r_t(s_t,a_t)\to\mathbb{R}$ 与 $c_t(s_t,a_t)\to\mathbb{R}$ 是两个函数，前者表示当前状态动作的奖励，后者表示当前状态动作对的成本。γ 是一个用于计算时间相关性的折扣率，其取值范围是[0，1]。在受约束的马尔可夫决策过程中，轨迹表示为 $\tau=\{s_0,a_0,r_0,c_0,...,s_t,a_t,r_t,c_t,...\}$。累积奖励 $G^r=\sum_{t=0}^{\infty}\gamma^t r_t$ 和累积成本 $G^c=\sum_{t=0}^{\infty}\gamma^t c_t$ 是用于评估策略表现的重要指标。安全离线强化学习的目标是利用离线数据集 $B(s_t,a_t,r_t,c_t,s_{t+1})$ 训练策略 $\pi(a\mid s)$，在满足安全约束的前提下，实现累积奖励的最大化。安全约束的具体设置通常与应用领域密切相关。在实际应用中，常见的安全约束形式是令期望累积成本小于预设的成本阈值 b。Q 值函数又名动作价值函数 $Q^r(s_t,\mathbf{a}_t)=\mathbb{E}_{\tau\sim\pi}[G_r\mid s_t,a_t]$ 用于评估策略 π 在当前状态 s 下选择的动作 a 的优劣。最优策略定义为 π^*。

$$\pi^*=\arg\max_{\pi} G^r \quad \text{s.t.} G^c\le b. \tag{4-1}$$

SPI 通过三个步骤来实现安全策略提升：增强安全特征，安全分布建模和提取安全策略。

4.21 增强安全特征

安全离线强化学习依靠受限的马尔可夫决策过程对轨迹进行建模，借助成本特征将潜在的成本因素量化为确定性值。为了提高数据质量，首先考虑将数据集进行重构。这种重构旨在提高成本特征并鼓励重构后的数据满足算法收敛

的需求。定义c'来衡量策略在当前状态下所采取行动的安全性。

$$c' = 1 - f(c) = \begin{cases} 0, & \text{if unsafe} \\ else, & \text{if safe} \end{cases} \tag{4-2}$$

其中 $f(c) \to [0,1]$。通过对成本的映射，从原始数据集 $\mathcal{B}(s_t, a_t, r_t, c_t, s_{t+1})$ 中获得了特征增强的重构数据集 $\hat{\mathcal{B}}(s_t, a_t, r_t, c'_t, s_{t+1})$。类比动作价值函数 $Q^r(s_t, a_t)$，重新定义了一个安全价值函数以评估策略在当前状态下采取的行动的安全性。

$$Q^{c'}(s_t, a_t) = \mathbb{E}_{\tau \sim \pi}[\sum_{t=0}^{\infty} \gamma^t c'_t] \tag{4-3}$$

正如上式所示。在计算动作状态对的安全价值函数时，为确保策略在长轨迹序列任务中的出色安全性，需要将折扣率 γ 纳入考虑范围之中。当折扣率 γ 较小时，该策略会专注于即时成本的获取。而当折扣率 γ 趋近于 1 时，策略会更“远视”，考虑未来可获得的成本。但当成本数据稀疏时，不合理的折扣率 γ 可能会触发恶性事件（例如成本效益消失）。

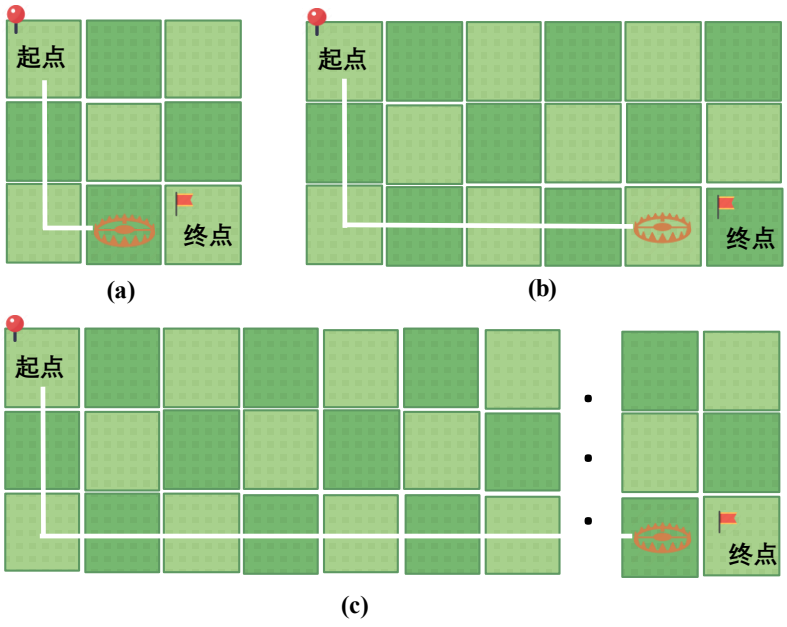

图 4-2 像素世界

Figure 4-2. Pixel world.

如图 4-2 所示，智能体始终从初始位置（0,0）出发，试图安全到达重点。如果智能体踏入陷阱，会将获得 1 的成本。如果达到终点，它将获得 1 的奖励。其余网格代表草地，踏足之上，智能体不会获得任何奖励或成本。

当使用 C 约束时，除了成本为 1 的陷阱位置外，所有其他正方形的成本为 0。该策略的优化目标是最小化累积期望成本值。当考虑使用白色轨迹来评估初始位置的状态动作对的安全价值函数时，在图 4-2（a）3*3 像素世界中估值为 $1*\gamma^3$，在图 4-2（b）3*6 像素世界中估值为 $1*\gamma^6$。如果考虑图 4-2（c）3*n $(n\to\infty)$ 此类的较大像素世界，则可能产生因折扣率不合理导致多次幂操作使得成本效益消失。

考虑引入 c' 通过映射解决成本消失的问题。经过映射之后，除陷阱 c' 为 0 以外，其余所有正方形的 c' 为 1。该策略的优化目标是最大化累积期望 c'。同样，使用白色轨迹来测量初始位置的状态动作对的安全价值函数，在 3*n 像素世界中会获得 $1*\gamma^5+1*\gamma^4+1*\gamma^3+1*\gamma^2+1*\gamma^1+1*\gamma^0$。此时，对于较大的像素世界，可以获得更大的价值。

4.2.2 安全分布建模

扩散模型表现出强大的数据建模能力和策略表达能力。然而，天真地将其直接应用于重构数据集 $\hat{\mathcal{B}}$，只能产生与原始数据集相似的策略，其本质类似于模仿学习。为了使模型能够捕获安全数据的分布，将扩散模型与两个值函数集成在一起。这使得扩散模型在训练过程中受到两种价值函数的指导，支持智能体产生丰富的策略。该耦合结构被定义为安全分布模型 D_ω，它使用从重构的离线数据集 $\hat{\mathcal{B}}$ 中采样所得数据 $(s_t,a_t,r_t,c_t',s_{t+1})$ 进行训练。

$$\begin{gathered}\omega\leftarrow arg\min_\omega[\eta KL_\omega((s_t,a_t)\,|\,(s_t,a_t^g))-\mu Q_\theta^r(s_t,a_t^g)-\phi Q_\lambda^{c'}(s_t,a_t^g)]\\ \text{s.t. } (s_t,a_t)\sim\hat{\mathcal{B}},(s_t,a_t^g)\sim\mathcal{D}_\omega\end{gathered}\tag{4-4}$$

安全分布模型 D_ω 由 KL 散度项和值函数项耦合在一起。其中，KL 散度项用于确保生成与原始数据集分布相似的策略，以避免外推错误。值函数项鼓励安全分布模型选择令 Q^r 与 $Q^{c'}$ 均给出高评价的动作。为了鼓励智能体在决策时考虑多样化策略，将 KL 散度项与两个值函数之间的参数关系定义为 $\eta+\mu+\phi=1$，其中 $\eta\in[0,1],\mu\in[0,1],\phi\in[0,1]$。

（1）KL 散度项

为了避免外推误差，使用 KL 散度来度量所捕获数据的分布与离线数据集

中数据分布之间的差异。

在正向过程或扩散过程中，SPI 对状态动作对进行采样，并不断加入高斯噪声，直到状态动作对变成纯噪声。根据马尔可夫性质，可以推断：

$$\begin{aligned}(\mathbf{s_0},\mathbf{a}_i) &= \sqrt{\alpha_i}\,(\mathbf{s_0},\mathbf{a}_{i-1}) + \sqrt{1-\alpha_i}\,\mathbf{y}_{i-1} \\ &= \sqrt{\alpha_i\alpha_{i-1}}\,(\mathbf{s_0},\mathbf{a}_{i-2}) + \sqrt{1-\alpha_i\alpha_{i-1}}\,\bar{\mathbf{y}}_{i-2} \\ &= \dots \\ &= \sqrt{\bar{\alpha}_i}\,(\mathbf{s_0},\mathbf{a}_0) + \sqrt{1-\bar{\alpha}_i}\,\mathbf{y}\end{aligned} \tag{4-5}$$

其中，$\alpha_i = 1-\beta_i$，$\bar{\alpha}_i = \prod_{j=1}^{t}\alpha_j$，并且 $\mathbf{y}_i$ 是一个标准正态分布，是两个高斯分布组合后的结果。此外，使用 $\{1,2,3,\dots,i\}$ 来表示扩散步骤。加噪过程是通过从原始状态动作对 $(\mathbf{s_0},\mathbf{a}_0)$ 和从标准正态分布中抽取的噪声之间进行插值操作完成。

对于反向过程或去噪过程，使用参数化神经网络 p_ω 进行拟合。

$$p_\omega((\mathbf{s_0},\mathbf{a}_{i-1})\,|\,(\mathbf{s_0},\mathbf{a}_i)) = \mathcal{N}(\boldsymbol{\mu}_\omega((\mathbf{s_0},\mathbf{a}_i),i),\boldsymbol{\Sigma}_\omega((\mathbf{s_0},\mathbf{a}_i),i)) \tag{4-6}$$

具体而言，SPI 使用噪声模型 $\boldsymbol{\epsilon}_\omega$ 来预测附加噪声的分布，然后从纯高斯噪声中逐步去噪，最终拟合回状态动作对。在变分推理中使用梯度下降法对证据下界（ELBO）进行优化，其中添加的噪声标签是由神经网络产生的噪声。两者之间的拟合确保了生成的状态动作对 (s_t,a_t^g) 与数据集中的 (s_t,a_t) 一致性。

$$KL_\omega((s_t,a_t)\,|\,(s_t,a_t^g)) = \|\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_\omega(\sqrt{\bar{\alpha}_i}\,(\mathbf{s}_t,\mathbf{a}_t) + \sqrt{1-\bar{\alpha}_i}\,\boldsymbol{\epsilon},i)\|^2 \tag{4-7}$$

（2）值函数项

SPI 利用两个价值函数来指导扩散模型获取安全且高回报的分布。Q_θ^r 鼓励在高奖励的动作空间抽样，同时 $Q_\lambda^{c'}$ 鼓励在安全性高的动作空间抽样。Q_θ^r 与 $Q_\lambda^{c'}$ 共同指导扩散模型。

$$\theta \leftarrow arg\min_\theta \| r_t + \gamma Q_\theta^r(s_{t+1},a_{t+1}) - Q_\theta^r(s_t,a_t)\|^2 \tag{4-8}$$

$$\lambda \leftarrow arg\min_\lambda \| c_t' + \gamma Q_\lambda^{c'}(s_{t+1},a_{t+1}) - Q_\lambda^{c'}(s_t,a_t)\|^2 \tag{4-9}$$

值函数项与 KL 散度项共同组成安全分布模型 D_ω，在扩散过程中对数据进行调整。扩散引导过程如图 4-3 所示。

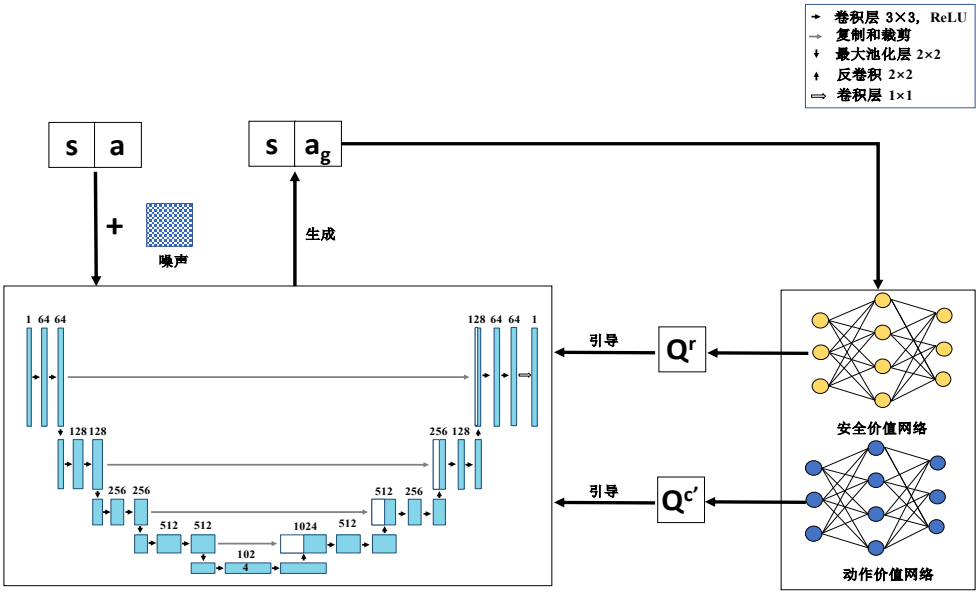

图 4-3 扩散引导过程

Figure 4-3 Network architecture used for diffusion

4.2.3 提取安全策略

表 4-1 基于双 Q 引导的离线强化学习安全策略增强算法

Table 4-1 Offline reinforcement learning safe policy improvement algorithm based on double-Q guidance

算法 1 基于双 Q 引导的离线强化学习安全策略增强算法

需要: 安全分布模型 D_ω,离线数据集 $\mathcal{B}$, 参数 η, μ, ϕ;

初始化: 噪声模型 ϵ_ω, 动作价值网络 Q_θ^r， 安全价值网络 $Q_\lambda^{c'}$

for 每次迭代 **do**:
  # 安全特征增强
  根据（4-2）式对数据集 B 进行重构获得扩散数据集 $\hat{\mathcal{B}}$
  # 安全分布建模
  随机采样 $(s_t, a_t, s_{t+1}, r_t, c_t') \sim \hat{\mathcal{B}}$
  根据（4-8）式使用数据片段 (s_t, a_t, s_{t+1}, r_t) 训练 Q_θ^r
  根据（4-3）式和（4-9）式使用数据片段 $(s_t, a_t, s_{t+1}, c_t')$ 训练 $Q_\lambda^{c'}$
  根据（4-7）式使用 (s_t, a_t) 训练 ϵ_ω
  更新噪声模型 ϵ_ω, 动作价值网络 Q_θ^r, 安全价值网络 $Q_\lambda^{c'}$
    根据（4-4）式训练 D_ω
    更新安全分布模型 D_ω.
    # 安全策略提取
    根据（4-10）式提取安全策略
end for

经过安全特征提升和安全分布捕获，SPI 具有对安全数据建模的能力。最

后，使用动作价值函数对生成的动作进行筛选，提取安全策略。

$$\pi(s) = arg\max_{a_t^g} Q_\theta^r(s, a_t^g) \quad \text{s.t. } \{a_t^g \sim D_\omega(s)\}_{t=1}^n \tag{4-10}$$

在实验中，为了减轻随机噪声和离群点的影响，将状态复制 50 次并输入到 SPI 中，在生成的 500 个动作值中进行筛选，最终使用贪婪算法选择动作价值函数值最大者。整个过程在算法 1 中说明。

4.3 实验评估

4.3.1 数据集及实验环境

实验所用数据集为 Datasets for Safe Reinforcement Learning（DSRL）[88]数据集，这是专为离线安全强化学习设计的数据集，通过评估不同难度级别的基线任务来评估所测试算法的性能。实际上，DSRL 不仅包括离线安全强化学习专用数据集，还囊括一些经典的安全离线强化学习基线算法和安全强化学习算法用以在在 38 个流行的安全任务中生成各种数据集。

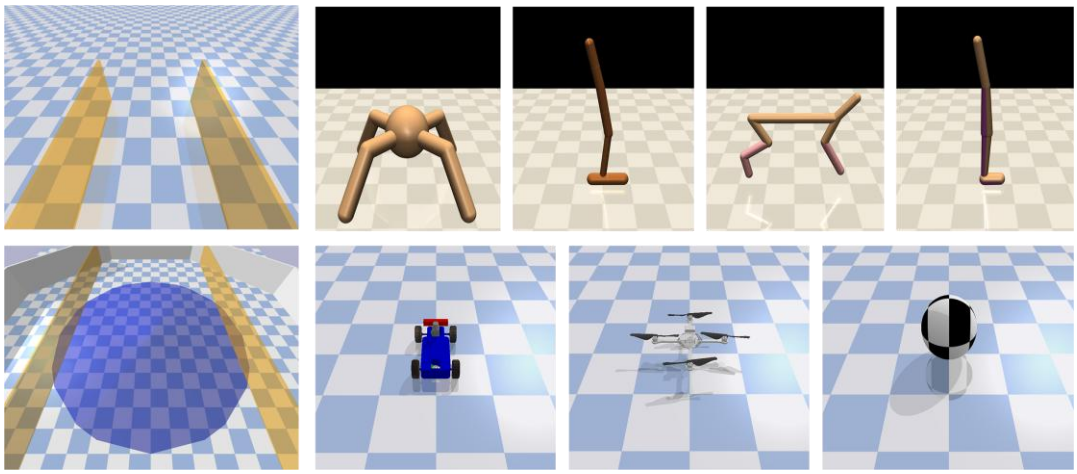

图 4-4 实验环境及智能体

Figure 4-4. Experimental environments and agents

在离线强化学习算法的安全应用背景下，实验所选用环境为经典安全环境，如 Bullet Safety Gym[89]和 Safety Gym[90]，其中包含了各种应用智能体，包括无人机、蚂蚁和跳跃者。智能体要执行的任务可以分为三类。所使用的实验环境和智能体类型如图 4-4 所示。

Circle 任务要求智能体在定义的圆圈内顺时针方向移动，如果偏离指定的安全边界，智能体将获得惩罚；在 Run 任务场景中，训练智能体在安全边界之间移动，智能体会因为超出安全界限会移动速度过快获得惩罚。Velocity 任务要

求智能体在满足速度约束条件下平稳行走。

实验所用 PC 机配置为 NVIDIA Corporation、64G 内存、13 代英特尔 i5-13600KF 处理器、64 位 Ubuntu20.04.6 LTS 操作系统作为实验平台。

4.3.2 实验评估指标

安全离线强化学习领域的主要目标是识别满足约束条件的最优策略。用于评估的关键指标是奖励最大化和成本最小化。为了减轻随机误差的影响，收集的数据经过标准化处理，并使用 20 幕的指标均值作为评估结果。此外，用粗体表示在同一任务下安全性最高者和奖励最高者。

此外，使用五个不同的基线进行比较：

CPQ 算法[85]涉及对值函数的约束，通过对分布外动作的动作值函数进行惩罚来满足安全约束；COptiDICE[82]算法通过利用最优策略的稳态分布与离线数据集分布之间的差异来对问题进行优化，从而对策略进行正则化。BEARL 和 BCQL 算法与 COptiDICE 属于同一类别，它们在离线强化学习算法中加入了拉格朗日约束[91]来对策略进行正则化。BC[92]表示行为克隆，其性能可以看作是采集离线数据集的所用行为策略的相似水平。

4.3.3 基准性能对比

表 4-2 SPI 及基线性能对比结果

Table 4-2 SPI and baseline performance comparison results

Task	Ours		CPQ	
	reward⬆	cost⬇	reward⬆	cost⬇
OfflineBallCircle-v0	**0.95**	5.65	0.66	**0.00**
OfflineCarCircle-v0	**0.71**	0.19	0.70	**0.00**
OfflineCarRun-v0	**0.97**	**0.00**	0.93	**0.00**
OfflineDroneCircle-v0	**0.64**	6.16	-0.26	**0.00**
OfflineAntVelocityGymnasium-v1	0.98	0.57	-1.01	0.08
OfflineHalfCheetahVelocityGymnasium-v1	**0.97**	**0.00**	-0.22	**0.00**
OfflineHopperVelocityGymnasium-v1	**0.89**	4.86	0.03	**0.00**
OfflineWalker2dVelocityGymnasium-v1	0.79	0.01	0.01	**0.00**
Average	**0.86**	2.18	0.11	**0.01**

表 4-3 SPI 及基线性能对比结果（续表）

Table 4-3 SPI and baseline performance comparison results(Continued)

COptiDICE		BEARL		BCQL		BC	
reward⬆	cost⬇	reward⬆	cost⬇	reward⬆	cost⬇	reward⬆	cost⬇
0.69	3.87	0.64	2.64	0.72	2.46	0.63	2.57
0.42	2.37	0.66	0.34	0.56	1.22	0.40	3.94
0.95	**0.00**	0.42	**0.00**	0.96	**0.00**	0.96	**0.00**
0.27	1.74	-0.26	**0.00**	0.48	0.35	0.60	4.90
0.97	2.73	-0.51	**0.00**	**0.99**	3.85	**0.99**	6.30
0.69	**0.00**	-0.27	0.12	0.96	9.89	0.93	1.93
0.40	1.22	0.17	9.44	0.83	9.37	0.70	3.37
0.13	2.43	0.79	**0.00**	**0.80**	**0.00**	0.77	0.02
0.57	1.80	0.21	1.57	0.79	3.39	0.75	2.88

表 4-2，表 4-3 显示了 8 种不同实验任务中 SPI 和其他基线之间的性能比较。由于其强大的数据拟合能力和权衡能力，SPI 算法在多个任务中表现出令人满意的性能。通过对比平均指标，SPI 方法在奖励和成本方面显示出良好的平衡能力,如图 4-5 和图 4-6 所示。

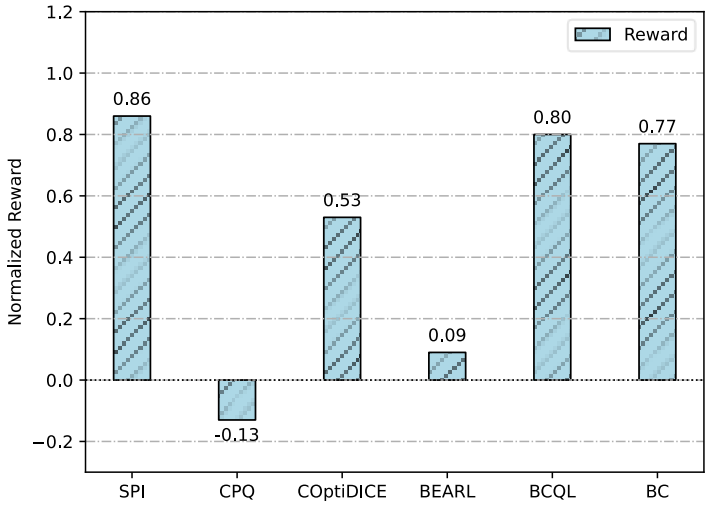

图 4-5 对比实验中各算法的平均奖励

Figure 4-5. The average reward of each algorithm in the compare experiment

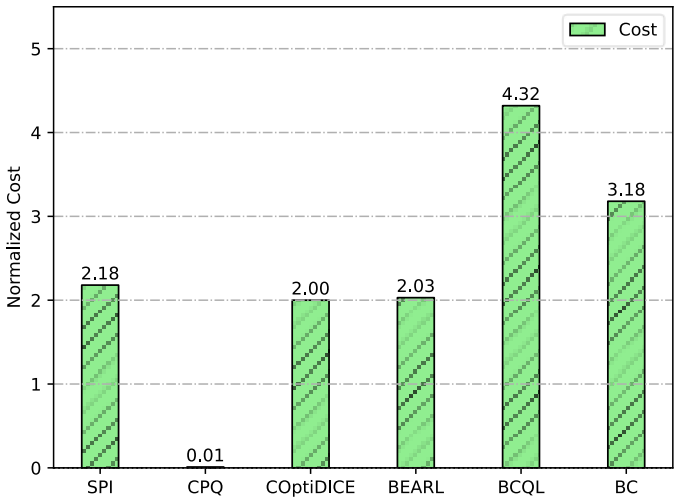

图 4-6 对比实验中各算法的平均成本

Figure 4-6. The average cost of each algorithm in the compare experiment

此外，在一些 SPI 没有获得最高回报的任务中，SPI 的性能仍然可以与最优算法相媲美。具体来说，在 OfflineAntVelocityGymnasium-v1 任务中，对比各算法的奖励指标，SPI 比 BCQL 和 BC 算法低大约 1 个百分点，但 BCQL 和 BC 的成本指标分别比 SPI 高 328 和 573 个百分点；对于 OfflineWalker2dVelocity Gymnasium-v1 任务，CPQ 算法虽然具有与 SPI 相似的成本值，但在奖励指标上表现显著差异，比 SPI 低 78 个百分点。此外，BEAR 和 BCQL 虽然在此特定任务上与 SPI 性能相当，但在其他任务上的性能并不优于 SPI。例如，在某些任务（如 OfflineHopperVelocityGymnasium-v1）中性能过低，这表明算法的鲁棒性存在潜在缺陷。

相比之下，基线算法很难在奖励与成本之间取得平衡。具体来说，一些基线算法（CPQ、COptiDICE 和 BEAR）过于保守，过分关注安全性，不愿执行具有更高奖励的策略。相反，BCQL 算法过于激进，在试图追求高奖励策略的同时违反了成本约束。此外，BC 算法的性能在很大程度上取决于所使用数据集的质量；因此，其鲁棒性相对较低。SPI 在所有数据集中都表现出卓越的性能，因为它能够捕获安全特征并有效地阐明策略。

4.3.4 消融性能对比

为了评估 SPI 算法的有效性，将 SPI 算法分解为几个功能组件，并通过选择性地删除或替换某些模块对模型做出修改。完整的 SPI 算法由 KL 散度项和

值函数项组成。KL 散度项与函数的结合侧重于产生高回报的策略，而 KL 散度项与函数的结合则确保产生满足安全约束的策略。

消融研究的目标是解决以下研究问题：

（1）SPI 算法中的各个模块是否按预期运行并对实验结果做出相应贡献？

（2）在学习过程中只关注安全约束而忽略原始数据分布此类使用单一约束解决混合问题的方法是否具有可行性？

表 4-4 SPI 消融实验结果

Table 4-4 Ablation experiment results of SPI

Task	SPI		SPI except $Q^{c'}$		SPI except Q^{r}	
	reward ↑	cost ↓	reward ↑	cost ↓	reward ↑	cost ↓
OfflineBallCircle-v0	**0.95**	5.65	0.94	5.88	0.04	**0.00**
OfflineCarCircle-v0	0.71	0.19	**0.93**	9.08	-0.01	**0.00**
OfflineCarRun-v0	**0.97**	**0.00**	0.96	0.18	0.95	**0.00**
OfflineDroneCircle-v0	**0.64**	6.16	0.56	8.22	-0.26	**2.89**
OfflineAntVelocityGymnasium-v1	0.98	0.57	**1.04**	19.70	-1.01	**0.00**
OfflineHalfCheetahVelocityGymnasium-v1	0.97	**0.00**	**1.01**	2.99	-0.11	**0.00**
OfflineHopperVelocityGymnasium-v1	**0.89**	4.86	0.17	8.26	0.06	**0.00**
OfflineWalker2dVelocityGymnasium-v1	**0.79**	0.01	0.17	6.10	-0.01	**0.00**
Average	**0.86**	2.18	0.72	7.55	-0.04	**0.36**

根据消融实验结果，如表 4-4 所示，分析结果如下：

（1）结果清楚地表明，完整的 SPI 算法优于删除或替换某些模块的版本。Q^{r} 函数和 $Q^{c'}$ 函数的表现分别证实了它们在探索和满足安全约束方面的有效性。在不同实验中所得指标之间的差异表明，在大多数环境中，高奖励数据分布、低成本数据分布与原始数据分布的区域并不完全重叠。

（2）KL 散度项+$Q^{c'}$ 组合与 SPI 之间奖励指标具有明显差异，这表明了仅依赖单一约束不能达到与多约束方法相同的性能水平。在训练过程中，单纯地使用安全约束引导算法采样，会导致 KL 散度项+$Q^{c'}$ 组合在评估过程中难以拟合分布外数据，导致算法收敛到局部最优。

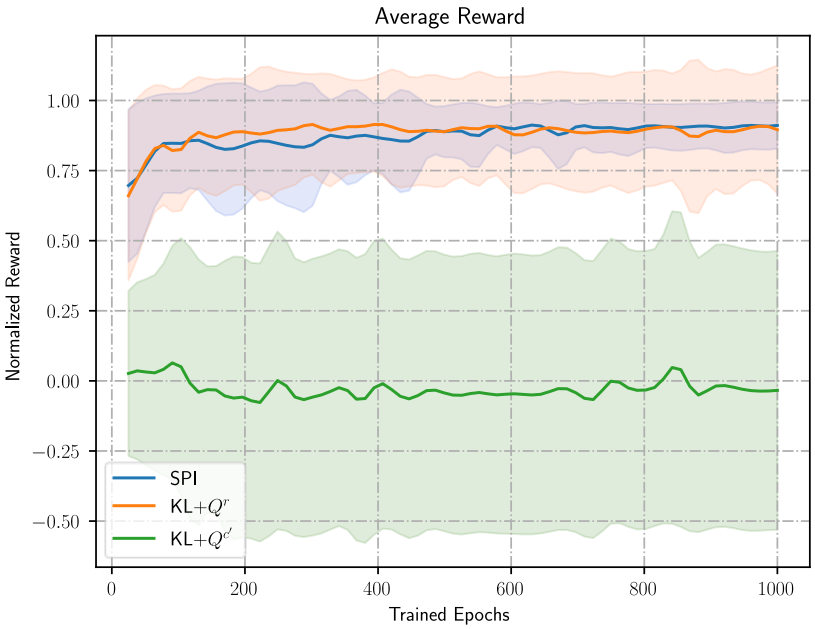

图 4-7 消融实验中各组合的平均奖励

Figure 4-7 The average reward of each combination in the ablation experiment

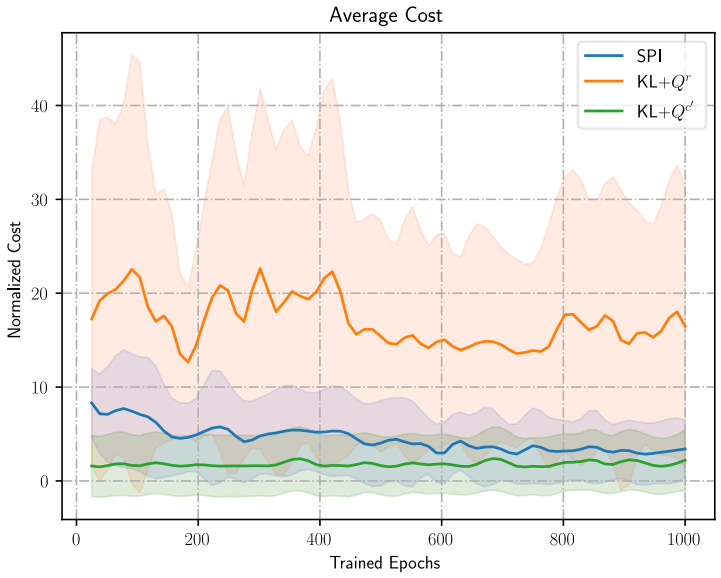

图 4-8 消融实验中各组合的平均成本

Figure 4-8 The average cost of each combination in the ablation experiment

此外，通过对 8 个任务中各种消融模块的性能进行综合分析发现，SPI 展示出了稳健的性能，在追求优异高回报的同时也能考虑动作的成本约束。如图 4-7 和图 4-8 所示，SPI 算法的整体性能优于其简化版本，这表明当 SPI 的各个组件

一起工作时，它们能够比简化版本更有效地权衡多个目标。

4.3.5 参数敏感性分析

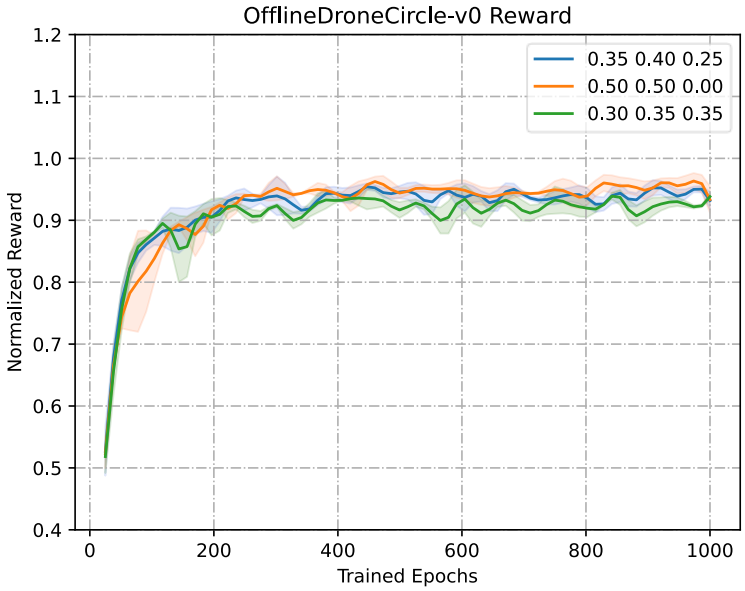

图 4-9 SPI 在随机参数下的平均奖励

Figure 4-9. Average reward of SPI under random parameters

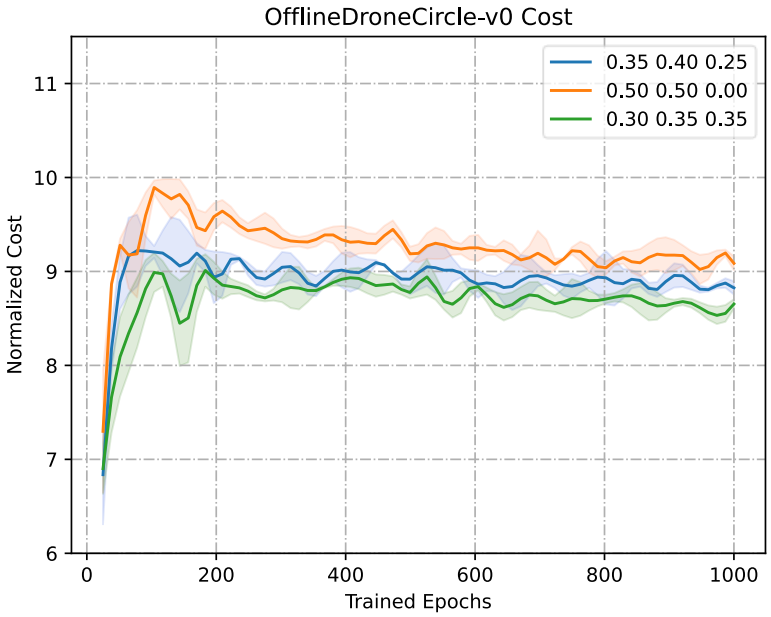

图 4-10 SPI 在随机参数下的平均成本

Figure 4-10. Average cost of SPI under random parameters

η, μ 和 ϕ 用于调节 KL 散度项与两个值函数之间的关系。为了满足不同的策

略需求，使用随机函数确定超参数数值。在三个超参数之和必须为 1 的约束下，进行了多组随机抽样实验。此外，实验参数设置为学习率 3e-4、评估次数 1000，评估频率 25、每次选择 top 1 为决策结果。

SPI 的性能如图 4-9 和图 4-10 所示。无论策略是倾向于更保守还是更激进的需求， SPI 算法都能够快速收敛。这种适应不同参数设置的能力证明了 SPI 方法的鲁棒性和灵活性。通过调整组成结构之间的平衡，该算法可以满足从高成本到低成本场景的广泛策略需求。

4.4 本章小结

本章主要对基于双 Q 引导的离线强化学习安全策略增强算法进行介绍。首先，对安全离线强化学习的背景和意义进行阐述，解释基于双 Q 引导的离线强化学习安全策略增强算法的研究问题。然后利用数学符号将研究问题进行建模，针对研究问题构建了算法框架。该算法通过三个步骤来确保算法的有效性，安全特征增强、安全分布建模、安全策略提取。SPI 通过在决策过程中，将成本和奖励纳入去噪过程，鼓励智能体决策时考虑奖励和成本因素。通过在 DSRL 中与多个安全离线强化学习基线表现进行比较，SPI 展示出了优异的综合能力。实验结果证明，SPI 在满足成本约束下，依然有良好的决策性能。

第5章　总结与展望

5.1 工作总结

离线环境下的强化学习方法在提高数据利用率，规避训练风险，降低成本花销等多个方面具有重要现实意义。但其发展仍存在一定阻碍，针对不同问题本文的主要工作总结如下：

（1）针对分布偏移所带来的未知动作 Q 值过估等问题，提出了名为 Diffusion-DA 的基于数据增强的双层离线强化学习框架。该框架通过使用原始数据集和扩散数据集辅助算法稳定收敛和离线环境下的伪探索，其中扩散数据集由扩散模型通过数据增强所得。Diffusion-DA 对算法可感知空间进行有效扩充，从而缓解离线环境下的分布偏移所带来的未知动作 Q 值过估问题。D4RL 的基线实验证实了 Diffusion-DA 的有效性。

（2）针对离线数据集良莠不齐、忽略环境不确定性所带来的安全问题，提出了一种基于双 Q 引导的离线强化学习安全策略增强算法，命名为 SPI。为了捕获离线数据集中的数据分布并生成安全策略，将扩散模型集成到离线强化学习框架中。通过用两个值函数指导扩散过程，使 SPI 产生的策略逐步逼近安全策略。此外，在数据集中重新映射成本，以避免成本效益消失，鼓励策略满足安全约束。SPI 在各个指标中几乎达到了最佳性能，展示出远超基线算法的优异性能。

5.2 工作展望

本文针对离线强化学习中的不同问题从两个角度提出相应的解决方案，所提方案具有一定可行性的同时，仍存在不足之处。

（1）Diffusion-DA 框架通过不同训练层，辅助算法收敛和实现离线环境下的伪探索。但面对特定偏好或需求，无法提供可引导式的性能增强服务。在未来的工作中将考虑如何在满足状态转移矩阵的前提下，对数据集进行偏好增强以符合任务需求。如，使用随机数据集为算法提供高奖励高覆盖面积的策略协助。

（2）SPI 专注于解决离线强化学习中的部署安全问题，但整体框架设计以

常量约束为优化目标，无法实时响应训练过程中调整约束的需求。在未来的工作将聚焦于如何将 SPI 算法扩展到实时预算约束领域，在满足安全需求的同时，能够根据部署过程中的反馈动态调整约束参数。

参考文献

[1] Liu C, Liu D. Deep reinforcement learning algorithm based on multi-agent parallelism and its application in game environment[J]. Entertainment Computing, 2024, 50: 100670.
[2] Silver D, Huang A, Maddison C J, et al. Mastering the game of Go with deep neural networks and tree search[J]. Nature, 2016, 529(7587): 484-489.
[3] Silver D, Hubert T, Schrittwieser J, et al. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play[J], 2018, 362(6419): 1140-1144.
[4] 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.
[5] Degrave J, Felici F, Buchli J, et al. Magnetic control of tokamak plasmas through deep reinforcement learning[J]. Nature, 2022, 602(7897): 414-419.
[6] Schiewer R, Subramoney A, Wiskott L. Exploring the limits of hierarchical world models in reinforcement learning[J]. Scientific Reports, 2024, 14(1): 26856.
[7] Nievas N, Espinosa-Leal L, Pagès-Bernaus A, et al. Offline Reinforcement Learning for Adaptive Control in Manufacturing Processes: A Press Hardening Case Study[J]. Journal of Computing and Information Science in Engineering, 2024, 25(1).
[8] Wang H, Liu Z, Hu G, et al. Offline Meta-Reinforcement Learning for Active Pantograph Control in High-Speed Railways[J]. IEEE Transactions on Industrial Informatics, 2024, 20(8): 10669-10679.
[9] 冯涣婷, 程玉虎, 王雪松. 基于不确定性估计的离线确定型 Actor-Critic[J]. 计算机学报, 2024, 47(04): 717-732.
[10] Iftikhar A, Ghazanfar M A, Ayub M, et al. A reinforcement learning recommender system using bi-clustering and Markov Decision Process[J]. Expert Systems with Applications, 2024, 237: 121541.
[11] Wang S, Chen X, Jannach D, et al. Causal Decision Transformer for Recommender Systems via Offline Reinforcement Learning[C]. Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2023: 1599–1608.
[12] Wang G, Liu X, Ying Z, et al. Optimized glycemic control of type 2 diabetes with reinforcement learning: a proof-of-concept trial[J]. Nature Medicine, 2023, 29(10): 2633-2642.
[13] An B, Sun S, Wang R. Deep Reinforcement Learning for Quantitative Trading: Challenges and Opportunities[J]. IEEE Intelligent Systems, 2022, 37(2): 23-26.
[14] Kumar A, Zhou A, Tucker G, et al. Conservative Q-Learning for Offline Reinforcement Learning[C]. Neural Information Processing Systems, 2020: 1179-1191.
[15] 王雪松, 王荣荣, 程玉虎. 基于表征学习的离线强化学习方法研究综述[J]. 自动化学报, 2024, 50(06): 1104-1128.

[16] 王硕汝, 牛温佳, 童恩栋等. 强化学习离线策略评估研究综述 [J]. 计算机学报, 2022, Volume(Issue): 1926-1945.
[17] 乌兰, 刘全, 黄志刚等. 离线强化学习研究综述 [J]. 计算机学报, 2025, Volume(Issue): 156-187.
[18] Fujimoto S, Meger D, Precup D. Off-Policy Deep Reinforcement Learning without Exploration[C]. Proceedings of the 36th International Conference on Machine Learning, 2019: 2052--2062.
[19] Wang Z, Novikov A, Żołna K, et al. Critic regularized regression[C]. Proceedings of the 34th International Conference on Neural Information Processing Systems, 2020: Article 651.
[20] Nadjahi K, Laroche R, Tachet Des Combes R. Safe Policy Improvement with Soft Baseline Bootstrapping[C]. Machine Learning and Knowledge Discovery in Databases, 2020: 53-68.
[21] Lyu J, Ma X, Li X, et al. Mildly conservative q-learning for offline reinforcement learning[J]. Advances in Neural Information Processing Systems, 2022, 35: 1711-1724.
[22] Wang Z, Hunt J J, Zhou M. Diffusion Policies as an Expressive Policy Class for Offline Reinforcement Learning[C]. The Eleventh International Conference on Learning Representations, 2023.
[23] Janner M, Du Y, Tenenbaum J, et al. Planning with Diffusion for Flexible Behavior Synthesis[C]. Proceedings of the 39th International Conference on Machine Learning, 2022: 9902--9915.
[24] Ajay A, Du Y, Gupta A, et al. Is Conditional Generative Modeling all you need for Decision Making?[C]. The Eleventh International Conference on Learning Representations, 2023.
[25] Lin Q, Tang B, Wu Z, et al. Safe offline reinforcement learning with real-time budget constraints[C]. Proceedings of the 40th International Conference on Machine Learning, 2023: Article 871.
[26] Jin Y, Yang Z, Wang Z. Is Pessimism Provably Efficient for Offline RL?[C]. Proceedings of the 38th International Conference on Machine Learning, 2021: 5084--5096.
[27] Zhu Z, Liu M, Mao L, et al. MADiff: Offline Multi-agent Learning with Diffusion Models[M]. 2024.
[28] Kostrikov I, Nair A, Levine S. Offline Reinforcement Learning with Implicit Q-Learning[C]. International Conference on Learning Representations, 2022.
[29] 王天久 乌 刘. 基于不确定性权重的保守 Q 学习离线强化学习算法[J]. 计算机科学, 2024, 51(9): e265.
[30] Huang L, Dong B, Zhang W. Efficient Offline Reinforcement Learning With Relaxed Conservatism[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(8): 5260-5272.
[31] Kidambi R, Rajeswaran A, Netrapalli P, et al. MOReL: Model-Based Offline Reinforcement Learning[C]. Neural Information Processing Systems, 2020: 21810-21823.
[32] Yu T, Thomas G, Yu L, et al. MOPO: Model-based Offline Policy

Optimization[C]. Neural Information Processing Systems, 2020: 14129-14142.
[33] Yu T, Kumar A, Rafailov R, et al. COMBO: Conservative Offline Model-Based Policy Optimization[C]. Neural Information Processing Systems, 2021: 28954-28967.
[34] Lu C, Ball P, Parker-Holder J, et al. Revisiting Design Choices in Offline Model Based Reinforcement Learning[C]. International Conference on Learning Representations, 2022.
[35] Rigter M, Lacerda B, Hawes N. RAMBO-RL: Robust Adversarial Model-Based Offline Reinforcement Learning[C]. Neural Information Processing Systems, 2022: 16082-16097.
[36] Kim B, Min O. Model-based Offline Reinforcement Learning with Count-based Conservatism[C]. International Conference on Machine Learning, 2023.
[37] Hong M, Qi Z, Xu Y. Model-based Reinforcement Learning for Confounded POMDPs[C]. Proceedings of the 41st International Conference on Machine Learning, 2024: 18668--18710.
[38] Chen L, Lu K, Rajeswaran A, et al. Decision Transformer: Reinforcement Learning via Sequence Modeling[C]. Neural Information Processing Systems, 2021: 15084-15097.
[39] Janner M, Li Q, Levine S. Offline Reinforcement Learning as One Big Sequence Modeling Problem[C]. Neural Information Processing Systems, 2021: 1273-1286.
[40] Wang K, Zhao H, Luo X, et al. Bootstrapped Transformer for Offline Reinforcement Learning[C]. Neural Information Processing Systems, 2022: 34748-34761.
[41] Xu M, Shen Y, Zhang S, et al. Prompting Decision Transformer for Few-Shot Policy Generalization[C]. Proceedings of the 39th International Conference on Machine Learning, 2022: 24631--24645.
[42] Yamagata T, Khalil A, Santos-Rodriguez R. Q-learning Decision Transformer: Leveraging Dynamic Programming for Conditional Sequence Modelling in Offline RL[C]. Proceedings of the 40th International Conference on Machine Learning, 2023: 38989--39007.
[43] Zhu T, Qiu Y, Zhou H, et al. Towards Long-delayed Sparsity: Learning a Better Transformer through R eward Redistribution[C]. Proceedings of the Thirty-Second International Joint Conference on Art ificial Intelligence, IJCAI-23, 2023: 4693-4701.
[44] Huang S, Hu J, Chen H, et al. In-Context Decision Transformer: Reinforcement Learning via Hierarchical Chain-of-Thought[C]. Proceedings of the 41st International Conference on Machine Learning, 2024: 19871--19885.
[45] 王天久, 刘全, 乌兰. 基于不确定性权重的保守 Q 学习离线强化学习算法[J]. 计算机科学, 2024, 51(09): 265-272.
[46] 侯永宏, 丁旺, 任懿等. 基于优质样本筛选的离线强化学习算法 [J]. 模式识别与人工智能, 2024, Volume(Issue): 1022-1032.
[47] 顾扬, 程玉虎, 王雪松. 基于优先采样模型的离线强化学习[J]. 自动化学报, 2024, 50(01): 143-153.
[48] García J, Fern, Fernández. A Comprehensive Survey on Safe Reinforcement Learning[J]. Journal of Machine Learning Research, 16(42): 1437-1480.
[49] Hans A, Schneegaß D, Schäfer A M, et al. Safe exploration for reinforcement

learning[C]. European Symposium on Artificial Neural Networks, 2008: 143-148.
[50] Gu S, Yang L, Du Y, et al. A Review of Safe Reinforcement Learning: Methods, Theories, and Applications[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(12): 11216-11235.
[51] Zhao W, He T, Chen R, et al. State-wise Safe Reinforcement Learning: A Survey[C]. Proceedings of the Thirty-Second International Joint Conference on Art ificial Intelligence, IJCAI-23, 2023: 6814-6822.
[52] Yang D, Gao Y, Hu B, et al. GWPF: Communication-efficient federated learning with Gradient-Wise Parameter Freezing[J]. Computer Networks, 2024, 255: 110886.
[53] Sun S, Xue W, Wang R, et al. DeepScalper: A Risk-Aware Reinforcement Learning Framework to Capture Fleeting Intraday Trading Opportunities[C]. Proceedings of the 31st ACM International Conference on Information & Knowledge Management, 2022: 1858–1867.
[54] Sun S, Wang X, Xue W, et al. Mastering Stock Markets with Efficient Mixture of Diversified Trading Experts[C]. Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2023: 2109–2119.
[55] Han J, Jentzen A, E W. Solving high-dimensional partial differential equations using deep learning[J]. Proceedings of the National Academy of Sciences, 2018, 115(34): 8505-8510.
[56] Silver D, Schrittwieser J, Simonyan K, et al. Mastering the game of Go without human knowledge[J]. Nature, 2017, 550(7676): 354-359.
[57] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models[J]. Advances in neural information processing systems, 2020, 33: 6840-6851.
[58] Karras T, Aittala M, Aila T, et al. Elucidating the design space of diffusion-based generative models[J]. Advances in neural information processing systems, 2022, 35: 26565-26577.
[59] Zhang R, Yu T, Shen Y, et al. Text-Based Interactive Recommendation via Offline Reinforcement Learning[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 36(10), 2022: 11694-11702.
[60] 刘全, 颜洁, 乌兰. 扩散模型期望最大化的离线强化学习方法[J]. 软件学报: 1-15.
[61] Fujimoto S, Hoof H, Meger D. Addressing Function Approximation Error in Actor-Critic Methods[C]. Proceedings of the 35th International Conference on Machine Learning, 2018: 1587--1596.
[62] Fujimoto S, Gu S S. A minimalist approach to offline reinforcement learning[J]. Advances in neural information processing systems, 2021, 34: 20132-20145.
[63] Lu C, Ball P, Teh Y W, et al. Synthetic experience replay[J]. Advances in Neural Information Processing Systems, 2023, 36: 46323-46344.
[64] Fu J, Kumar A, Nachum O, et al. D4rl: Datasets for deep data-driven reinforcement learning[J]. arXiv preprint arXiv:2004.07219, 2020.
[65] Todorov E, Erez T, Tassa Y. Mujoco: A physics engine for model-based control[C]. 2012 IEEE/RSJ international conference on intelligent robots and systems, 2012: 5026-5033.
[66] Towers M, Kwiatkowski A, Terry J, et al. Gymnasium: A standard interface for

reinforcement learning environments[J]. arXiv preprint arXiv:2407.17032, 2024.
[67] Awac: Accelerating online reinforcement learning with offline datasets[EB/OL]. arXiv preprint arXiv:2006.09359, 2020.
[68] Dohare S, Hernandez-Garcia J F, Lan Q, et al. Loss of plasticity in deep continual learning[J]. Nature, 2024, 632: 768 - 774.
[69] Prudencio R F, Máximo M R O A, Colombini E L. A Survey on Offline Reinforcement Learning: Taxonomy, Review, and Open Problems[J]. IEEE Trans. Neural Networks Learn. Syst., 2024, 35(8): 10237-10257.
[70] Singh Y, Kumar R, Kabdal S, et al. YouTube Video Summarizer using NLP: A Review[J]. Int. J. Perform. Eng., 2023, 19(12): 817-823.
[71] Nan J, Deng W, Zhang R, et al. A Long-Term Actor Network for Human-Like Car-Following Trajectory Planning Guided by Offline Sample-Based Deep Inverse Reinforcement Learning[J]. IEEE Trans Autom. Sci. Eng., 2024, 21(4): 7094-7106.
[72] Fang X, Zhang Q, Gao Y, et al. Offline Reinforcement Learning for Autonomous Driving with Real World Driving Data[C]. 2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC), 2022: 3417-3422.
[73] Gao C, Huang K, Chen J, et al. Alleviating Matthew Effect of Offline Reinforcement Learning in Interactive Recommendation[C]. Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2023: 238-248.
[74] Xiao T, Wang D. A General Offline Reinforcement Learning Framework for Interactive Recommendation[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021: 4512-4520.
[75] Xiong W, Zhong H, Shi C, et al. Nearly Minimax Optimal Offline Reinforcement Learning with Linear Function Approximation: Single-Agent MDP and Markov Game[C]. In Proceedings of the International Conference on Learning Representations (ICLR), 2023.
[76] Cui Q, Du S S. Provably Efficient Offline Multi-agent Reinforcement Learning via Strategy-wise Bonus[C]. Proceedings of the 36th International Conference on Neural Information Processing Systems, 2022.
[77] Lee A X, Devin C, Springenberg J T, et al. How to Spend Your Robot Time: Bridging Kickstarting and Offline Reinforcement Learning for Vision-based Robotic Manipulation[C]. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2022: 2468-2475.
[78] Jin J, Graves D, Haigh C, et al. Offline Learning of Counterfactual Predictions for Real-World Robotic Reinforcement Learning[C]. International Conference on Robotics and Automation (ICRA), 2022: 3616-3623.
[79] Zhou G, Ke L, Srinivasa S S, et al. Real World Offline Reinforcement Learning with Realistic Data Source[C]. Offline RL Workshop NeurIPS, 2023: 7176-7183.
[80] Siegel N Y, Springenberg J T, Berkenkamp F, et al. Keep Doing What Worked: Behavior Modelling Priors for Offline Reinforcement Learning[C]. In Proceedings of the International Conference on Learning Representations (ICLR), 2020.
[81] Le H M, Voloshin C, Yue Y. Batch Policy Learning under Constraints[C]. Proceedings of the 36th International Conference on Machine Learning, 2019: 3703-

3712.
[82] Lee J, Paduraru C, Mankowitz D J, et al. COptiDICE: Offline Constrained Reinforcement Learning via Stationary Distribution Correction Estimation[C]. In Proceedings of the International Conference on Learning Representations, 2022.
[83] Lin Q, Tang B, Wu Z, et al. Safe Offline Reinforcement Learning with Real-Time Budget Constraints[C]. Proceedings of the 40th International Conference on Machine Learning, 2023: 21127-21152.
[84] Choi J J, Castañeda F, Tomlin C J, et al. Reinforcement Learning for Safety-Critical Control under Model Uncertainty, using Control Lyapunov Functions and Control Barrier Functions[C]. Robotics: Science and Systems, 2020.
[85] Xu H, Zhan X, Zhu X. Constraints Penalized Q-learning for Safe Offline Reinforcement Learning[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022: 8753-8760.
[86] Wang Z, Hunt J J, Zhou M. Diffusion Policies as an Expressive Policy Class for Offline Reinforcement Learning[C], 2023.
[87] Jhingran S, Goyal M K, Rakesh N. DQLC: A Novel Algorithm to Enhance Performance of Applications in Cloud Environment[J]. Int. J. Perform. Eng., 2023, 19(12): 771-778.
[88] Liu Z, Guo Z, Lin H, et al. Datasets and Benchmarks for Offline Safe Reinforcement Learning[J]. Journal of Data-centric Machine Learning Research, 2024, (12): 1-29.
[89] Gronauer S. Bullet-Safety-Gym A Framework for Constrained Reinforcement Learning[R]. Department of Electrical and Computer Engineering, Technical University of Munich, 2022.
[90] Ji J, Zhang B, Zhou J, et al. Safety Gymnasium: A Unified Safe Reinforcement Learning Benchmark[C]. Proceedings of the 37th International Conference on Neural Information Processing Systems, 2023.
[91] Stooke A, Achiam J, Abbeel P. Responsive safety in reinforcement learning by pid lagrangian methods[C]. International Conference on Machine Learning, 2020: 9133-9143.
[92] Bain M, Sammut C. A Framework for Behavioural Cloning[C]. Machine Intelligence 15, Intelligent Agents 1995: 103-129.

攻读学位期间参与的科研项目

[1] 基于 5G 网络的智能交互展示系统及应用研究课题，项目编号：2023cyyd04

攻读学位期间发表的学术论文目录

研究生期间发表的期刊：

[1] Xiaohan Yang , et al. Safe Policy Improvement via Diffusion Model for Offline Reinforcement Learning[J]，Journal of Internet Technology, 2025(JCR Q4 中科院四区)(第一作者)(已录用)

致谢

在完成这篇硕士论文之际，我心中充满了感激之情。回首这段硕士研究生的学习与研究历程，心中充满了感慨与感恩。在这段充满挑战与收获的旅程中，我收获了知识、成长与珍贵的情谊，也遇到了许多给予我帮助与支持的人。此刻，我怀着无比诚挚的心情，向他们表达最深切的感谢。

首先，我要衷心感谢我的导师李钧老师。导师不仅在学术上给予了我悉心的指导，更在生活中给予了我无微不至的关怀。从论文的选题、研究方法的确定，到论文的撰写与修改，导师都倾注了大量的心血。他严谨的治学态度、渊博的学识以及对学术的执着追求，让我深受启发，也让我明白了做学问的真正意义。在遇到困难和挫折时，导师总是耐心地倾听我的困惑，给予我鼓励和建议，让我重新找回信心和方向。他的教诲如同一盏明灯，照亮了我前行的道路，让我在学术的海洋中不断探索与成长。我将永远铭记导师的恩情，并将这份感恩化作前行的动力，在未来的学术道路上继续努力。

其次，我还要感谢 PRiSE 研究组的王老师，卢老师，唐老师等各位老师。在与他们相处的这段时光里，我收获了诸多宝贵的经验与回忆。老师们学识渊博、治学严谨，他们总是耐心地为我解答各种学术难题，引导我在研究的道路上不断探索前行。他们的悉心指导，让我在专业领域有了长足的进步，也让我对科研工作和科研生活有了更深刻的理解。

我有幸结识了许多志同道合的朋友和同门师兄弟姐妹。他们如同一束束温暖的光，照亮了我这段旅程，让我在学术探索的道路上不再孤单。感谢我的同门师兄弟姐妹们。我们一同在实验室里度过无数个日夜，互相讨论学术问题，分享研究心得。在实验遇到瓶颈时，我们一起头脑风暴，寻找解决问题的思路；在论文撰写过程中，我们互相审阅、提出建议，共同进步。你们的陪伴和支持，让我感受到团队的力量和温暖。特别感谢刘江师兄，在我最迷茫的时候，是你耐心地倾听我的烦恼，给予我鼓励和建议，让我重新振作起来；感谢孙梦婷同学在实验过程中与我并肩作战，共同探讨问题、分享经验；感谢陈宝柱师弟、

徐恺师弟、杜浩杰师弟和李国俊师弟在数据收集和实验验证中提供的帮助。

感谢我的朋友们。在我忙碌的学术生活中，你们是我放松和充电的港湾。我们一起在校园里漫步，分享生活的点滴；在周末的聚餐中，畅谈理想与未来。你们的乐观和积极感染着我，让我在面对压力时能够保持良好的心态。感谢你们在我需要帮助时伸出援手，无论是学业上的困惑还是生活中的琐事，你们总是第一时间给予我支持。你们的友谊是我在这段旅程中最宝贵的财富，我会永远珍惜。

我要特别感谢我的父母。他们始终如一地支持我的学业，在我迷茫时给予鼓励，在我疲惫时给予温暖。父母的理解和支持是我能够专心致志完成学业的坚强后盾。记得在论文写作最紧张的阶段，父亲每天都会关心我的进度，母亲则时常提醒我要注意劳逸结合。他们的关爱让我在繁忙的科研工作中始终保持积极乐观的心态。

最后，我要衷心感谢参与本论文评审的各位老师。您在百忙之中抽出时间，对我的论文进行审阅和指导，提出了许多宝贵的意见和建议。这些意见不仅帮助我进一步完善了论文的结构和内容，更让我在学术研究的严谨性和规范性上有了更深的认识。感谢您为我提供了宝贵的学习和成长机会，让我在学术道路上不断进步。我将铭记您的教诲，继续努力，不辜负您的期望。再次向您表示最衷心的感谢！

道阻且长，快马加鞭。

安徽工程大学学位论文原创性声明

本人郑重声明：我恪守学术道德，崇尚严谨学风。所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已明确注明和引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品及成果的内容。论文为本人亲自撰写，我对所写的内容负责，并完全意识到本声明的法律后果由本人承担。

学位论文作者签名：杨小强

日期：2025 年 5 月 9 日