自动化学报
ACTA AUTOMATICA SINICA

基于流匹配策略优化的离线强化学习方法

刘腾龙 谢旭辉 黄佳成 郭道洁 兰奕星 徐昕

Flow Matching-based Policy Optimization for Offline Reinforcement Learning

LIU Teng-Long, XIE Xu-Hui, HUANG Jia-Cheng, GUO Dao-Jie, LAN Yi-Xing, XU Xin

在线阅读 View online: https://doi.org/10.16383/j.aas.c250638

基于流匹配策略优化的离线强化学习方法

刘腾龙 [1] 谢旭辉 [1] 黄佳成 [1] 郭道洁 [1] 兰奕星 [1] 徐 昕 [1]

摘 要 离线强化学习旨在利用预先采集的行为数据集优化智能体策略, 其面临的主要挑战是迭代优化的目标策略与产生数据集的行为策略之间存在分布偏移. 现有方法通常采用策略正则化以缓解该问题, 但其难以根据行为数据质量自适应地调整学习过程的约束强度, 并且难以有效建模行为策略中复杂的多峰分布. 针对上述问题, 提出基于流匹配策略优化 (FMPO) 的离线强化学习方法. FMPO 利用流匹配模型对行为策略分布进行建模, 从流匹配策略中选择高优势动作以形成自适应约束, 引导学习策略在行为数据分布邻域内进行优化; 同时, 将流匹配策略生成的动作作为先验输入条件, 利用行为先验促进策略的高效学习. 通过基于流匹配策略的优化机制, FMPO 能够在策略提升与满足数据集分布约束之间实现动态平衡. 实验结果表明, FMPO 在 D4RL 基准任务上取得先进性能, 显著优于现有主流离线强化学习方法.

关键词 离线强化学习; 流匹配; 策略优化; 分布偏移; 策略建模

引用格式 刘腾龙, 谢旭辉, 黄佳成, 郭道洁, 兰奕星, 徐昕. 基于流匹配策略优化的离线强化学习方法. 自动化学报, 2026, **52**(7): 1319−1333

DOI 10.16383/j.aas.c250638 **CSTR** 32138.14.j.aas.c250638

Flow Matching-based Policy Optimization for Offline Reinforcement Learning

LIU Teng-Long[1] XIE Xu-Hui[1] HUANG Jia-Cheng[1] GUO Dao-Jie[1] LAN Yi-Xing[1] XU Xin[1]

Abstract Offline reinforcement learning aims to optimize agent policies using pre-collected behavior datasets. Its central challenge lies in the distribution shift between the iteratively optimized target policy and the behavior policy that generated the dataset. Existing methods typically employ policy regularization to mitigate this issue; however, they struggle to adaptively adjust the strength of the constraints during learning according to the quality of the behavior data, and they have difficulty effectively modeling the complex multimodal distributions of behavior policies. To address these issues, this paper proposes flow matching-based policy optimization (FMPO) for offline reinforcement learning. FMPO leverages a flow matching model to characterize the behavior policy distribution, and selects high-advantage actions from the flow matching policy to form an adaptive constraint that guides the learned policy to be optimized within the neighborhood of the behavior data distribution. Meanwhile, actions generated by the flow matching policy are incorporated as prior conditioning inputs, exploiting the behavioral prior to facilitate efficient policy learning. Through this flow matching-based policy optimization mechanism, FMPO is able to achieve a dynamic balance between policy improvement and adherence to the dataset distribution constraint. Experimental results demonstrate that FMPO achieves strong performance on the D4RL benchmark tasks, significantly outperforming existing mainstream offline reinforcement learning methods.

Keywords offline reinforcement learning; flow matching; policy optimization; distribution shift; policy modeling

Citation Liu Teng-Long, Xie Xu-Hui, Huang Jia-Cheng, Guo Dao-Jie, Lan Yi-Xing, Xu Xin. Flow matching-based policy optimization for offline reinforcement learning. *Acta Automatica Sinica*, 2026, **52**(7): 1319−1333

强化学习 (reinforcement learning, RL)[1−2] 通过与环境在线交互, 以一种试错的方式实现对策略的迭代优化, 近年来已经在大语言模型[3]、计算机游戏[4−5] 等领域取得一定的进展. 然而, 这种通过在线交互反馈的学习方式往往面临样本效率低和安全风险高等问题[6−7]. 在不具备与环境大量在线交互的策略学习条件的高风险场景下, 比如自动驾驶、机器人控制、医疗诊断等, 强化学习方法难以获得较优的求解策略. 为克服这一问题, 学者开始研究直接从任务数据集中学习策略的方法[8]. 在早期阶段, 学者们提出批强化学习[9], 类似机器学习中批量学习的方法. 随着该方向研究热度进一步提升, Levine 等[8] 将此类算法定义为离线强化学习 (offline rein-

收稿日期 2025-11-17 录用日期 2026-05-07
Manuscript received November 17, 2025; accepted May 7, 2026
国家自然科学基金 (T2521006, 62533021, 62403483), 国防科技大学自主科研基金项目 (25-ZZCX-ZXGC-01, ZK25-47) 资助
Supported by National Natural Science Foundation of China (T2521006, 62533021, 62403483) and Innovation Research Foundation of National University of Defense Technology (25-ZZCX-ZXGC-01, ZK25-47)
本文责任编委 罗彪
Recommended by Associate Editor LUO Biao
1. 国防科技大学智能科学学院 长沙 410073
1. College of Intelligence Science and Technology, National University of Defense Technology, Changsha 410073

forcement learning, Offline RL).

离线强化学习旨在从预先收集的固定数据集中直接学习策略, 其中固定的数据集由某个行为策略产生, 这种学习方式能够避免直接与环境在线交互, 从而降低在安全隐患场景中的策略学习风险[10]. 虽然离线强化学习无需直接与环境进行交互, 但其面临新的挑战: 由于仅依赖由行为策略采集的固定离线数据集进行策略学习, 优化得到的目标策略可能偏离数据分布, 从而导致目标策略与行为策略之间存在分布偏移. 在策略评估和策略提升过程中, 这种分布偏移会导致智能体对分布之外的状态–动作对出现值函数过估计的问题, 由此会使学习的策略陷入次优解[10]. 因此, 如何解决分布偏移、缓解值函数的过估计是离线强化学习所面临的主要挑战[11−13].

针对离线强化学习的主要挑战, 一类主流的解决方法是策略正则化. 这类方法主要通过约束目标策略与行为策略之间的距离, 使目标策略接近行为策略, 进一步减少产生分布外动作的可能性, 从而达到缓解分布偏移的目的. TD3 + BC[14] 是一种简单有效的策略约束类方法, 其在 TD3[15] 方法上进行改进, 通过在策略提升的目标函数中引入行为克隆项来约束目标策略与行为策略之间的距离, 有效地缓解了分布偏移问题. 虽然 TD3 + BC 仅进行较小的改进, 但是其在策略学习的性能上却展示出不错的效果. 此外, 还有研究者通过 KL (Kullback-Leibler) 散度、Fisher 散度以及最大均值差异 (maximum mean discrepancy, MMD) 等方法使得目标策略尽可能接近行为策略以缓解分布偏移问题. Jaques 等[16] 通过减少目标策略与行为策略之间的 KL 散度, 使目标策略尽可能接近行为策略. Fisher-BRC[17] 以 Fisher 散度为度量来衡量目标策略与行为策略之间的差异, 通过减少 Fisher 散度以约束目标策略, 防止其过度偏离行为策略. BEAR[18] 通过减少目标策略与行为策略之间的最大均值差异, 使学习的策略逼近行为策略, 从而缓解分布偏移. 虽然这类方法能够缓解分布偏移, 但是其面临着策略表达能力弱和策略约束过度保守的问题, 从而限制了策略提升的空间, 尤其是在行为策略较复杂或是行为策略水平较差的数据集上.

随着生成模型理论和方法的发展, 学者们尝试探索生成模型在离线强化学习中的应用[19−20], 其中一类方法用生成模型来建模产生数据的行为策略或者目标策略[21]. BCQ[22] 和 SPOT[23] 直接使用条件变分自编码器 (variational auto-encoder, VAE) 实现对行为策略的建模, 并约束目标策略与行为策略接近, 减少采样到分布外动作的概率. 但是 VAE 对数据分布的表达能力有限, 尤其是具有复杂多峰的数据分布. 为进一步提升策略分布的表达能力, 得益于扩散模型 (diffusion model) 等生成模型对复杂分布建模的强表达能力, 学者们开始尝试探索扩散模型对策略建模的研究. 与正态分布表达策略相比, 扩散模型能够更好地完成具有多峰分布的复杂数据集的策略建模. Diffusion-QL[24] 方法通过扩散模型[25] 对目标策略进行建模, 并基于 TD3 + BC 算法框架实现策略优化. SfBC[26] 方法和 IDQL[27] 方法将产生数据集的行为策略表示为扩散模型, 直接通过模仿学习的方法实现对扩散模型的策略优化, 在推理阶段通过扩散模型生成多个候选动作, 结合样本内值函数得到最终的动作. 以上方法虽然通过扩散模型在一定程度上提升了策略表达能力, 但是基于扩散模型的方法在训练和推理过程中, 需要通过去噪的形式采样获得动作, 这会导致较低的样本学习效率和推理效率. 为提升训练效率和推理速度, FQL[28] 方法采用流匹配策略实现行为策略的建模, 目标策略通过流匹配策略蒸馏优化得到.

上述离线强化学习方法主要通过约束目标策略接近行为策略, 以此来缓解目标策略与行为策略之间的分布偏移, 但是这类方法的策略约束往往过度保守, 不利于策略性能的进一步提升. 策略约束过度保守会使得目标策略受到行为策略的严重影响, 尤其在数据质量不好的情况下, 这种约束会使策略陷入次优甚至更差的解, 从而使智能体学习性能不佳. 虽然这些方法尝试通过扩散模型、流匹配模型 (flow matching model) 提升策略的表达能力, 但是并没有解决策略约束过度保守的根本问题. 此外, 上述基于扩散模型和流匹配模型的方法一般只是对产生数据集的行为策略进行模仿学习, 不能动态地调整对数据集中不同质量数据的学习程度, 并没有充分发挥扩散模型和流匹配模型对复杂数据分布的强表达能力. 理想情况下, 扩散模型或者流匹配模型在实现对行为的建模时, 应增强对数据集中高质量行为的模仿学习, 削弱数据集中低质量行为的模仿学习. 因此, 一种动态的行为策略学习方法尤为重要, 这将影响目标策略最终的学习性能. 并且, 直接使用扩散模型和流匹配模型作为目标策略, 在测试阶段会遇到推理效率低的问题, 因此如何在保证充分发挥扩散模型和流匹配模型策略表达能力的同时提升智能体的推理效率, 尤为重要.

为解决策略表达能力弱和策略约束过度保守的问题, 本文提出一种简单高效的方法——基于流匹配策略优化的离线强化学习方法. 该方法通过优势函数引导的流匹配策略学习, 强化对数据集中高质

量数据的策略学习, 实现增强行为策略的建模. 该方法基于增强行为策略, 以流匹配策略为条件引导策略优化, 充分利用流匹配策略强大的非线性拟合能力, 为优化策略提供行为先验知识, 同时结合门控函数动态调整策略学习目标, 从而自适应、高效地实现策略的性能提升. 实验结果表明, FMPO (flow matching-based policy optimization) 在 D4RL 基准任务上取得了优异的性能, 显著优于现有主流方法.

1　预备知识

本节介绍强化学习与离线强化学习的基本概念, 分析分布偏移问题及策略约束方法, 并阐述生成模型与流匹配机制在策略分布建模中的作用. 这些理论为后续提出的流匹配策略自适应引导的离线强化学习算法提供了数学与模型基础.

强化学习是一种通过智能体与环境的交互来学习最优策略的机器学习方法. 通常, 强化学习问题可以形式化为马尔可夫决策过程. 其定义为:

$$\mathcal{M} = (\mathcal{S},\ \mathcal{A},\ r,\ P,\ \gamma,\ \rho_0) \tag{1}$$

其中, $\mathcal{S}$ 表示状态空间; $\mathcal{A}$ 表示动作空间; $r\colon \mathcal{S}\times\mathcal{A}\to\mathbf{R}$ 为奖励函数; $P(s'|s,\ a)$ 为状态转移概率函数, s' 为状态 s 的下一状态; $\gamma\in(0,\ 1)$ 为折扣因子; ρ_0 为初始状态分布. 在时间步 k, 智能体观测当前状态 s_k, 通过策略 $\pi(a_k|s_k)$ 选择动作 a_k, 环境根据 P 返回奖励 r_k 并转移至新状态 s_{k+1}. 智能体的目标是学习一个最优策略 π^*, 以最大化期望累计回报:

$$\pi^* = \arg\max_{\pi} \mathrm{E}_{\tau\sim\pi,\ s_0\sim\rho_0}\left[\sum_{k=0}^{\infty}\gamma^k r_k\right] \tag{2}$$

其中, $\tau=(s_0,\ a_0,\ s_1,\ a_1,\ \cdots)$ 表示状态–动作序列.

1.1　Q 函数与策略优化基础

Q 函数 $Q(s,\ a)$ 表示在任意状态 $s\in\mathcal{S}$ 下执行动作 a 后的期望折扣回报, 又称为动作价值函数. 其可以评价在状态 s 下执行动作 a 的优劣:

$$Q^{\pi}(s,\ a) = \mathrm{E}_{\pi}\left[\sum_{k=0}^{\infty}\gamma^k r(s_k,\ a_k)\mid s_0=s,\ a_0=a\right] \tag{3}$$

其满足贝尔曼方程:

$$Q^{\pi}(s,\ a) = \mathrm{E}_{s'\sim P}[r(s,\ a)+\gamma\mathrm{E}_{a'\sim\pi}[Q^{\pi}(s',\ a')]] \tag{4}$$

其中, a' 表示策略 π 在下一状态 s' 产生的下一动作.

1.2　离线强化学习与分布偏移问题

在线强化学习依赖于智能体与环境的频繁交互, 而在实际应用中, 这种交互往往成本高昂或具有风险. 离线强化学习提出在仅利用固定历史数据集 $\mathcal{D}=\{(s,\ a,\ r,\ s')\}$ 的情况下学习策略, 从而避免在线交互. 然而, 由于数据集仅来源于固定的历史行为策略 $\mu(a|s)$, 学习策略 $\pi(a|s)$ 与 $\mu(a|s)$ 之间的分布差异会导致“分布偏移”问题. 该问题会引起值函数在分布外动作上的过估计, 从而不利于策略收敛到更优解.

1.3　生成模型在策略分布建模中的作用

近年来, 生成模型被广泛用于强化学习中的策略建模. 其核心思想是利用生成模型来学习从噪声分布 $p_0(a)$ 到目标动作分布 $p(a|s)$ 的映射. 在此框架下, 变分自编码器[29]、扩散模型[21, 24] 以及流匹配模型[28] 都被用于表达行为策略的复杂分布特征.

本文主要通过流匹配模型实现对行为策略的建模, 流匹配模型通过学习速度场 $v_\theta(x_t,\ t)$, 描述数据从初始分布 $p_0(x)$ 向目标分布 $p_1(x)$ 的连续演化过程:

$$\frac{\mathrm{d}x_t}{\mathrm{d}t} = v_\theta(x_t,\ t) \tag{5}$$

其训练目标为最小化流匹配损失函数:

$$\mathcal{L}_{\text{flow}} = \mathrm{E}_{t,\ x_1,\ \epsilon}\left[\|v_\theta(x_t,\ t)-(x_1-\epsilon)\|_2^2\right] \tag{6}$$

其中, ϵ 为从源分布 (标准高斯) 采样的噪声.

与扩散模型不同, 流匹配模型不依赖噪声调度, 推理效率更高且数值稳定性更好, 因此在复杂策略分布的建模中具有显著优势.

在离线强化学习中, 引入流匹配思想可视为在动作空间中建立“从噪声动作到最优动作”的连续映射. 即将策略生成过程表示为:

$$\frac{\mathrm{d}a_t}{\mathrm{d}t} = v_\theta(a_t,\ s,\ t) \tag{7}$$

其中, 速度场 v_θ 控制动作分布随时间的演化, 使其逐渐趋近高价值区域. 这种方式不仅能通过生成建模约束策略分布, 缓解分布偏移问题, 还可通过流匹配的平滑性引入策略学习的稳定性, 从而提升算法的收敛特性与泛化能力.

2　相关工作

2.1　离线强化学习中的策略正则化方法

策略正则化方法是离线强化学习中用来解决分布偏移问题的一种主流方法, 其旨在通过约束目标策略接近行为策略来缓解分布偏移. 为了促使目标策略接近行为策略, 目前研究主要有基于行为克隆的策略正则化方法和基于策略散度惩罚的正则化方法. 基于行为克隆的方法是在策略优化目标函数中

显式地加入行为克隆损失项, 使学习到的策略动作更接近行为策略产生的动作. TD3 + BC[14] 在策略更新目标中加入行为克隆项, 通过最小化当前策略与行为策略之间的动作分布差异, 限制策略向分布外区域的过度外推. 基于策略散度惩罚的方法通过量化学习策略与行为策略之间的概率分布差异, 利用距离度量 (如 KL 散度[16]、Fisher 散度[17]、MMD[18] 等) 来构建惩罚项, 从而限制目标策略在分布空间中的偏移程度. DMG[30] 通过在数据集邻域内选择高 Q 值动作, 在贝尔曼目标中混合泛化 Q 值与样本内 Q 值以缓解误差累积. BDPO[31] 将行为正则化计算为扩散轨迹中每一步反向转移核的 KL 散度累积和, 并将其作为策略正则化方法中的正则化项, 进而约束目标策略. 然而, 当前策略正则化方法往往过于保守地约束目标策略与行为策略的接近, 忽略了对潜在高价值行为策略的建模. 本研究通过流匹配生成模型实现对高价值行为策略的建模, 从而生成高价值动作, 并通过策略正则化引导目标策略实现高效的策略提升.

2.2 离线强化学习中的生成模型

近年来, 随着生成模型理论与方法的深入研究和发展[32−34], 生成模型在离线强化学习领域得到广泛关注. 早期研究主要使用 VAE 进行行为策略建模. BCQ[22] 与 SPOT[23] 采用 VAE 实现行为策略建模, 通过隐式约束使得目标策略接近行为策略, 利用行为密度约束目标策略. A2PR[29] 通过优势函数增强 VAE, 生成高价值动作实现优势引导的策略正则化. 然而, 这些方法在面对复杂行为策略分布时, 往往面临 VAE 建模能力不足的难点.

扩散模型具有强大的建模能力, 其能够实现对复杂的行为策略分布建模[35−37], 最近在离线强化学习领域得到越来越广泛的关注. Diffusion-QL[24] 将目标策略建模为以状态为条件的扩散模型, 并结合动作价值最大化项实现策略优化. SfBC[26]、IDQL[27] 使用扩散模型对产生数据集的行为策略进行建模, 并结合候选动作得到目标策略. DTQL[38] 提出双策略架构, 通过扩散信任域损失结合扩散策略的表达能力与单步策略的推理效率, 在保持生成质量的同时提升了策略学习效率. Diffusion-DICE[39] 通过结合 DICE 框架的分布修正能力与扩散模型的强表达能力, 利用样本内引导, 在不引入外推误差的情况下实现多模态策略的高效提取. DiffuserLite[40] 提出一种轻量级的规划精炼过程, 通过从粗到细的层次化轨迹生成, 该方法减少了冗余信息建模. EDA[41] 提出一种高效的扩散对齐框架, 通过标量值网络实现扩散密度的直接计算, 仅需极少量的 Q 标签数据即可完成从行为模型到优化策略的高效微调. DIVO[42] 利用值函数条件化的扩散模型生成高质量的分布内样本, 在策略优化阶段动态选择高优势动作, 克服了传统扩散策略对低质量数据不加区分正则化的缺陷, 实现了保守性与探索性之间的平衡. 尽管扩散模型在离线强化学习中展现了强大的生成表达能力, 但是其迭代的前向加噪与反向去噪过程往往带来高昂的计算代价.

针对扩散模型计算效率低的问题, SRPO[43] 通过利用扩散模型的分数函数直接对策略梯度进行正则化, 该方法在保留扩散模型强大表达能力的同时, 减少了训练和推理中耗时的迭代采样过程. Consistency-AC[44] 首次将一致性模型引入离线强化学习, 该工作保留了扩散模型处理多峰分布的高表达能力, 同时通过其特有的自一致性属性, 在无需多步迭代采样的前提下实现了策略推理. 流匹配通过学习从简单先验分布到目标数据分布的确定性常微分方程轨迹, 可以实现高效的推理. FQL[28] 训练一个用于行为克隆的流匹配策略, 并将其与最大化 Q 值结合实现了一步策略优化, 避免了迭代生成中的反向传播问题. EVOR[45] 将流匹配集成到 Q 函数建模中, 实现了价值学习与策略优化的解耦. 而 QIPO[46] 提出能量引导流匹配框架, 基于此设计了 Q 加权迭代策略优化算法, 通过价值导向流匹配策略实现策略改进. 但是这些方法一味地模仿数据集, 不能动态地学习数据集中高质量行为, 并且行为策略无法为目标策略提供高效的引导, 这会导致目标策略收敛到次优甚至更差的解.

3 方法

本节介绍基于流匹配策略优化的离线强化学习方法. 第 3.1 节首先介绍流匹配以及流匹配策略, 作为后续方法的基础; 第 3.2 节介绍优势引导的流匹配策略学习方法, 该方法旨在学习能够生成高优势动作的行为策略; 第 3.3 节介绍基于流匹配的生成式策略优化方法, 其目标是充分利用流匹配策略学习过程中获得的数据集先验知识, 以有效引导策略提升; 算法实现细节及伪代码见第 3.4 节.

基于流匹配策略优化的离线强化学习方法的算法整体框架如图 1 所示, 该图展示了两个核心组成部分:

- 优势引导的流匹配策略学习. 首先从离线数据集中采样离线数据, 并计算对应的优势函数 $A(s, a)$. 基于优势函数构造动作权重 $w(s, a)$, 随后通过加权的流匹配损失对策略进行训练, 从而引导策略倾向于生成具有更高优势的动作.

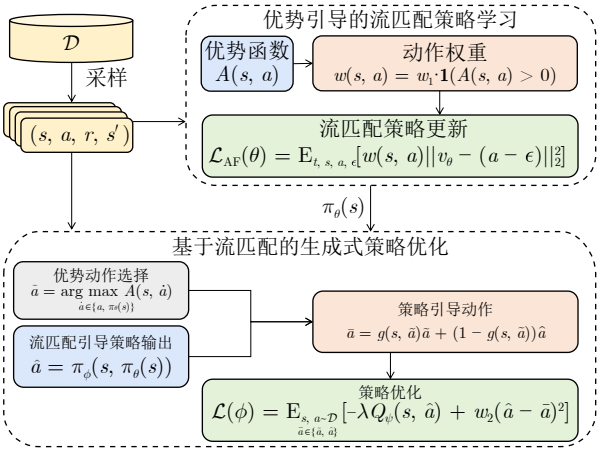


图 1　算法整体框架

Fig. 1　Overall framework of the proposed algorithm

● 基于流匹配的生成式策略优化. 在策略优化阶段, 首先根据优势函数选择优势动作 $\tilde{a}$, 同时由流匹配策略生成候选动作 $\hat{a}$. 随后通过门控函数对两者进行融合得到最终动作 $\bar{a}$, 并结合 Q 函数对策略进行优化.

3.1　流匹配及流匹配策略

在介绍优势引导的流匹配策略学习方法之前, 本节首先介绍流匹配及流匹配策略. 流匹配 (flow matching) 是一种新兴的生成模型, 可用于数据分布的生成建模, 最早由 Lipman 等[47] 提出. 类似于扩散模型, 流匹配旨在从噪声分布出发实现对真实分布的估计. 不同的是, 扩散模型先学习与时间相关的得分函数或噪声, 再由此推导从噪声分布到数据分布的速度场, 从而通过去噪的方式得到数据分布. 而流匹配方法通过定义常微分方程 (ordinary differential equation, ODE) 直接学习时间相关的速度场. 这使得训练目标更为简单——它仅基于噪声样本与真实数据点之间的插值进行学习, 而无需噪声调度 (noise schedule). 因此, 相比于扩散模型, 流匹配在数值上通常更稳定, 并且对超参数的敏感性更低. 给定数据分布 $p(x)$, 流匹配旨在学习一个连续的速度场 (velocity field) $v_\theta(x_t,\ t) : \mathbf{R}^d \times [0,\ 1] \to \mathbf{R}^d$, 该速度场用来描述数据分布如何从一个噪声分布 ($t = 0$) 逐渐"流向"目标真实分布 ($t = 1$), 本文考虑具有线性路径和均匀时间采样的流匹配变体. 该速度场通过最小化以下损失函数实现优化:

$$\mathcal{L}_{\mathrm{flow}}(\theta) = \mathrm{E}_{\substack{t\sim\mathrm{Unif}([0,\ 1]),\\ \epsilon\sim\mathrm{N}(0,\ I),\\ x_1\sim p(x)}}\left[\|v_\theta(x_t,\ t) - (x_1 - \epsilon)\|_2^2\right] \tag{8}$$

其中, x_t 表示流路径, 在这里是 x_1 和 ϵ 之间的线性插值; $x_1 - \epsilon$ 是目标速度, 由于采用线性路径, 从噪声 ϵ ($t = 0$) 到数据 x_1 ($t = 1$) 的理想位移速度就是它们的差值; $\mathrm{N}(0,\ I)$ 表示标准的高斯分布; $\mathrm{Unif}([0,\ 1])$ 表示定义在单位区间上的均匀分布; 速度场 $v_\theta(x_t,\ t)$ 可以被视为从标准正态分布 $\mathrm{N}(0,\ I)$ 到目标数据分布 $p(x)$ 的平均速度场方向, 给定一个边缘速度场 $v_\theta(x_t,\ t)$, 可以通过求解以下常微分方程来生成新的样本: $\frac{\mathrm{d}}{\mathrm{d}t}x_t = v_\theta(x_t,\ t)$.

流匹配策略是一种基于流匹配思想的策略生成方法, 可用于强化学习或模仿学习中的策略分布建模. 该方法最早源于将流匹配的生成建模思想扩展至决策空间的研究, 其核心目标是学习一个能够从随机策略分布流向最优策略分布的连续速度场. 类似于流匹配用于数据生成的方式, 流匹配策略通过定义一个随时间演化的策略分布, 实现从随机动作分布到高价值动作分布的渐进变换. 不同于传统的策略梯度方法直接通过采样和优化期望回报来调整策略参数, 流匹配策略通过学习一个时间相关的条件速度场, 以建模策略分布在时间维度上的连续流动过程. 具体地, 流匹配策略定义了一个基于状态 s、动作 a_t 和时间 t 的速度场 $v_\theta(a_t,\ s,\ t)$, 用于描述策略如何从初始随机分布 (噪声分布) 逐步流向目标分布 (行为策略分布). 这一过程同样可表示为常微分方程形式:

$$\frac{\mathrm{d}}{\mathrm{d}t}a_t = v_\theta(a_t,\ s,\ t) \tag{9}$$

其中 a_t 表示在时刻 t 的动作样本.

为了学习该速度场, 流匹配策略在训练阶段对随机策略样本 ϵ 和行为策略动作样本 a 之间进行插值, 通过最小化流匹配损失来优化策略网络参数:

$$\mathcal{L}_{\mathrm{FMP}}(\theta) = \mathrm{E}_{\substack{t\sim\mathrm{Unif}([0,\ 1]),\\ (s,\ a)\sim\mathcal{D},\\ \epsilon\sim\mathrm{N}(0,\ I)}}\left[\left\|v_\theta(a_t,\ s,\ t) - (a - \epsilon)\right\|_2^2\right] \tag{10}$$

其中, $a_t = (1-t)\epsilon + ta$ 表示动作插值路径. 训练完成后, 给定当前状态 s, 即可通过对上述常微分方程积分生成新的策略样本, 实现从初始随机策略到高性能策略的平滑演化.

3.2　优势引导的流匹配策略学习

此前, 已有多种生成模型被引入离线强化学习领域, 用于对行为策略进行建模与优化, 例如 SfBC、IDQL 和 FQL[28]. 然而, 这些方法在学习效率与策略表达能力之间尚未取得良好的平衡. 具体而言, SfBC 和 IDQL 通过扩散模型对数据集中的行为策略进行学习. 尽管扩散模型具有较强的策略表达能力, 但这类方法需要通过加噪和去噪的方式完成模型训练与采样, 训练过程复杂且计算开销较大.

FQL 则基于流匹配生成模型进行离线强化学习, 但其对策略的约束是基于数据集中所有动作实现的, 即使部分动作质量较低, 也会迫使目标策略向这些动作靠近, 从而可能导致策略陷入次优解, 降低策略提升的效率.

针对上述问题, 本节提出一种优势引导的流匹配策略学习方法, 该方法不同于以往直接通过行为克隆完成流匹配策略学习的做法, 而是将优势函数与流匹配策略学习结合, 通过优势函数筛选出具有较高优势的动作进行行为克隆, 从而得到一个改进的行为策略, 为后续策略提升提供有效引导. 该优势引导的流匹配策略学习方法通过最小化以下损失函数实现优化:

$$\begin{cases}\mathcal{L}_{\mathrm{AF}}(\theta)=\mathrm{E}_{t,\,s,\,a,\,\epsilon}\left[w(s,\,a)||v_\theta-(a-\epsilon)||_2^2\right]\\ w(s,\,a)=w_1\cdot\mathbf{1}(A(s,\,a)>0)\end{cases}\tag{11}$$

其中, $\mathbf{1}(A(s,\,a)>0)$ 是一个指示函数, 当 $A(s,\,a)>0$ 时取 1, 否则取 0; $A(s,\,a)$ 表示优势函数, 用于衡量动作相对于平均水平的质量高低; w_1 为一个超参数; $w(s,\,a)$ 表示动作权重, 其形式多样, 例如可以是通过 Q 值计算得到的玻尔兹曼分布加权 $w(s,\,a)=\exp\big(Q\,(s,\,a)/\beta\big)/\mathrm{E}_{a\sim\mu(a|s)}[\exp\,(Q\,(s,\,a)/\beta)]$ (β 为温度超参数, 用于控制指数加权分布的集中程度), 也可以是通过优势函数得到的非负截断权重 $w(s,\,a)=\max(A(s,\,a),\,0)$.

将优势函数与流匹配策略优化相结合, 不仅能够充分发挥流匹配策略强大的策略表达能力, 而且通过 $\mathbf{1}(A(s,\,a)>0)$ 选择高优势动作, 可以实现对相应速度场的重点学习. 借助该速度场, 流匹配策略可以从噪声样本“流向”高优势动作, 从而得到一个增强的行为策略.

3.3 基于流匹配的生成式策略优化方法

上述基于优势引导的流匹配策略学习方法可以得到一个增强的行为策略, 在此基础之上, 本节提出流匹配策略条件引导的策略优化方法. 一方面, 基于优势引导的流匹配策略相当于改进的行为策略, 能够产生更多具有较高优势的动作; 另一方面, 通过以流匹配策略为条件的策略优化以及高优势选择的策略约束, 能够充分利用流匹配策略基于数据集的先验知识来隐式引导策略, 更好地平衡策略约束与策略提升之间的关系, 在保证策略约束的同时实现高效的策略提升.

从优势引导的流匹配策略和数据集中选择具有较高优势的动作

$$\tilde{a}=\arg\max_{\dot{a}\in\{a,\,\pi_\theta(s)\}}A(s,\,\dot{a})\tag{12}$$

定义 1. 给定状态 s, 以流匹配策略 π_θ 的输出 $\pi_\theta(s)$ 作为条件输入的策略 π_ϕ, 定义为

$$a\sim\pi_\phi(\cdot\mid s,\,\pi_\theta(s))\tag{13}$$

其中 π_ϕ 以 $\pi_\theta(\cdot)$ 在状态 s 下的输出作为引导信号, 用于为策略空间中的学习提供行为策略的先验知识. π_ϕ 可为确定性或随机性形式, 若 π_ϕ 为确定性策略, 上式可简化为 $a=\pi_\phi(s,\,\pi_\theta(s))$. π_ϕ 为流匹配引导策略 (flow matching guidance policy), 其目标是在流匹配策略生成的条件分布上进行学习, 以在目标任务中实现更高效的策略提升.

通过一个门控函数 $g(s,\,\tilde{a})$, 并根据高优势动作 $\tilde{a}$ 和流匹配引导策略 $\pi_\phi(s,\,\pi_\theta(s))$, 确定最终的策略引导动作 $\bar{a}$

$$\begin{cases}\bar{a}=g(s,\,\tilde{a})\tilde{a}+\big(1-g(s,\,\tilde{a})\big)\pi_\phi(s,\,\pi_\theta(s))\\ g(s,\,\tilde{a})=\mathbf{1}(A(s,\,\tilde{a})>0)\end{cases}\tag{14}$$

其中, $\mathbf{1}(A(s,\,\tilde{a})>0)$ 是一个指示函数, 当 $A(s,\,\tilde{a})>0$ 时取 1, 否则取 0. 策略引导动作根据优势函数动态调整, 能够实现自适应策略约束, 并且基于流匹配的生成式策略优化通过最小化以下损失函数实现:

$$\begin{cases}\mathcal{L}(\phi)=\mathrm{E}_{\substack{(s,\,a)\sim\mathcal{D},\\ \bar{a}\in\{\tilde{a},\,\hat{a}\}}}\left[-\lambda Q_\psi(s,\,\hat{a})+w_2(\hat{a}-\bar{a})^2\right]\\ \hat{a}=\pi_\phi(s,\,\pi_\theta(s))\end{cases}\tag{15}$$

其中 $Q_\psi(s,\,\hat{a})$ 为评价器网络对状态–动作对 $(s,\,\hat{a})$ 的评分; $\lambda=\alpha N/\sum_{s_i,\,a_i}Q(s_i,\,a_i)$, α 是一个超参数, N 表示批次大小[14]; $\tilde{a}=\arg\max_{\dot{a}\in\{a,\,\pi_\theta(s)\}}A(s,\,\dot{a})$; w_2 表示超参数. 本文提出的方法能够使目标策略朝着具有较高优势的动作方向进行学习. 同时, 它能够动态地平衡策略提升和策略保守两者之间的关系. 充分发挥流匹配策略的策略表达能力, 为流匹配引导策略提供更多的先验知识以引导其进行高效的策略提升.

3.4 具体实现方法

本文基于经典的离线强化学习算法 TD3 + BC[14], 提出一种基于流匹配的生成式策略优化方法. 该方法在 TD3 + BC 的总体框架下, 引入流匹配策略以实现对策略提升过程的策略引导. 具体而言, Q 值网络、流匹配策略网络、流匹配引导策略以及价值网络的参数分别记为 ψ, θ, ϕ, ω. 其中, Q 值网络包含两个独立的网络, 其参数分别为 ψ_1 与 ψ_2, 对应的目标 Q 值网络参数为 ψ_1' 与 ψ_2'; 目标流匹配引导策略的参数记为 ϕ'.

为了在策略提升与策略保守性之间取得更好的平衡, 本文对实现策略提升的 Q 值最大化项进行了

归一化处理. 由此, 策略提升阶段的优化目标函数可表示为: $\mathcal{L}(\phi)=\mathrm{E}_{s,\,a,\,\bar{a}}[-\lambda Q_{\psi}(s,\,\pi_{\phi}(s,\,\pi_{\theta}(s)))+(\pi_{\phi}(s,\,\pi_{\theta}(s))-\bar{a})^2]$.

在本研究中, 策略优化目标需要根据优势函数进行动态调整, 因此优势函数的精确学习至关重要. 本文通过同时学习 Q 函数与值函数来间接获得优势函数的估计, 从而为后续的策略优化提供指导. 其中, 价值网络 $V_{\omega}(\cdot)$ 的参数 ω 通过最小化以下损失函数进行优化:

$$\mathcal{L}_V(\omega)=\mathrm{E}_{(s,\,a)\sim\mathcal{D}}[(Q_{\psi_i}(s,\,a)-V_{\omega}(s))^2] \tag{16}$$

Q 值网络的参数 ψ 则通过最小化以下时序差分 (temporal-difference, TD) 误差来进行优化:

$$\mathcal{L}_Q(\psi_i)=\mathrm{E}_{\substack{(s,\,a,\,s')\sim\mathcal{D},\\ a'\sim f(s')}}[(r(s,\,a)+\gamma\min_i Q_{\psi_i'}(s',\,a')-Q_{\psi_i}(s,\,a))^2] \tag{17}$$

其中, $Q_{\psi_i'}$ 表示目标 Q 值网络, $i\in\{1,\,2\}$; $f(s')=\mathrm{clip}(\pi_{\phi'}(s',\,\pi_{\theta}(s'))+\epsilon_0,\,-A,\,A)$, $\epsilon_0\sim\mathrm{clip}(\mathrm{N}(0,\,\hat{\sigma}^2),\,-c,\,c)^3$, c 和 $\hat{\sigma}$ 是用于探索的超参数.

优势函数 $A(s,\,a)$ 可通过 Q 函数 $Q_{\psi_i}(s,\,a)$ 与值函数 $V_{\omega}(s)$ 的差值得到:

$$A(s,\,a)=Q_{\psi_i}(s,\,a)-V_{\omega}(s) \tag{18}$$

通过上述方式获得的优势函数, 将被用于优势引导的流匹配策略学习以及基于流匹配的生成式策略优化过程中. 在策略学习阶段, 流匹配策略基于优势函数实现自适应的动态策略约束, 从而有效缓解了过度策略约束导致的保守问题以及策略表达能力不足的问题, 并能够实现高效的策略提升. 算法实现的伪代码见算法 1.

算法 1. 基于流匹配策略优化的离线强化学习方法

输入. 离线数据集 $\mathcal{D}$, 超参数 α, 批次大小 N, 更新率 ρ, 训练步数 T.

输出. 优化后的执行器参数 ϕ.

1. 初始化 Q 值网络参数 ψ_1, ψ_2, 流匹配网络参数 θ, 流匹配引导策略网络参数 ϕ 以及值函数网络 ω;
2. 初始化目标网络: $\psi_1'\leftarrow\psi_1$, $\psi_2'\leftarrow\psi_2$, $\phi'\leftarrow\phi$;
3. **for** $t=1$ **to** T **do**
4. 　从离线数据集采样 $(s,\,a,\,r,\,s')\sim\mathcal{D}$;
5. 　**流匹配策略学习:**
6. 　最小化式 (11) 实现参数 θ 的更新;
7. 　**Q 函数与值函数更新:**
8. 　最小化式 (17) 与式 (16) 实现更新;
9. 　**流匹配引导策略优化:**
10. 　最小化式 (15) 更新参数 ϕ;
11. 　**目标网络更新:**
12. 　$\psi_i'\leftarrow\rho\psi_i+(1-\rho)\psi_i'$, $i=1,\,2$
13. 　$\phi'\leftarrow\rho\phi+(1-\rho)\phi'$
14. **end for**

4　实验结果

4.1　测试基准

为评估 FMPO 的性能, 实验在离线强化学习领域的 D4RL 基准任务上进行. D4RL 提供了丰富的任务和多样化的数据集, 主要包括 HalfCheetah、Hopper、Walker2d 等运动控制任务以及基于 Ant 的迷宫导航任务 (AntMaze). 在数据质量方面, 针对 HalfCheetah、Hopper、Walker2d 三类运动控制任务, D4RL 提供了 medium、medium-replay、medium-expert 等不同质量水平的数据集: medium 为中等表现策略采集的数据, medium-replay 为训练至中等表现过程中积累的回放数据, medium-expert 则由等量的中等表现策略数据与专家级策略数据混合而成.

而对于 Ant 的迷宫导航任务, 根据迷宫的布局, 分为 umaze、medium、large 三种类别; 根据任务设定, 分为固定、diverse、play 三种类型. 其中“固定”类型的数据集表示从固定起始位置到固定目标位置的智能体轨迹数据, 此类数据集有“antmaze-umaze-v2”; “diverse”类型的数据集表示在迷宫中引导智能体到随机的目标位置的轨迹数据, 此类数据集有“antmaze-umaze-diverse-v2、antmaze-medium-diverse-v2、antmaze-large-diverse-v2”; “play”类型的数据集表示从人工选择的初始位置出发, 控制智能体到达人工指定的目标位置而生成的数据, 此类数据集有“antmaze-medium-play-v2、antmaze-large-play-v2”. D4RL 基准已经成为离线强化学习领域公认的用于测试算法性能的数据集, 通过上述基准的设置, 可以更加全面地评估算法在训练结束时的性能以及在不同质量水平数据上的泛化性和鲁棒性.

4.2　对比基线算法

为验证 FMPO 方法的性能, 本节在 D4RL 基准上将其与 8 个基线算法进行对比. 8 个基线算法分别为: TD3 + BC[14]、BCQ[22]、BEAR[18]、CQL[48]、SfBC[26]、Diffusion-QL[24]、DTQL[38]、QIPO[46]. TD3 + BC[14] 在策略更新目标中加入行为克隆项, 通过最小化当前策略与行为策略之间的动作分布差异, 限制策略向分布外区域的过度外推; BCQ[22] 通过 VAE 实现行为策略建模, 通过隐式约束使得目标

策略接近行为策略, 利用行为密度约束目标策略; BEAR[18] 通过 MMD 距离来约束当前策略, 使其生成的动作保持在行为策略的支撑集范围内, 而非单纯地进行分布匹配 (如 KL 散度); CQL[48] 通过在常规 Q 学习的目标函数中引入正则项来压低未见动作的 Q 值, 从而学习到一个保守的价值下界, 有效解决了离线强化学习中因分布偏移导致的价值高估问题; SfBC[26] 使用扩散模型对产生数据集的行为策略进行建模, 并结合候选动作得到目标策略; Diffusion-QL[24] 将目标策略建模为以状态为条件的扩散模型, 并结合动作价值最大化项实现策略优化; DTQL[38] 提出双策略架构, 通过扩散信任域损失结合扩散策略的表达能力与单步策略的推理效率, 在保持生成质量的同时提升了策略学习效率; QIPO[46] 提出能量引导流匹配框架, 基于此设计了 Q 加权迭代策略优化算法, 通过价值导向流匹配策略实现策略改进.

为了保证实验对比的公平性, 对于本文所提出的 FMPO 方法, 每个任务使用 4 个不同的随机种子进行训练, 并报告最后 10 次评估的平均性能及其标准误差. 对比基线算法结果源于它们各自的论文, 对比实验结果在表 1 中展示. 具体的实验参数设置详见表 2.

实验结果表明: FMPO 在 Gym 运动控制的全部 9 个任务上均取得了最优性能. 其中, FMPO 在 halfcheetah-medium-v2 任务中性能尤为显著, 说明该方法在处理中等质量数据时具有很强的能力. 在 AntMaze 导航任务中, FMPO 在 6 个迷宫任务中有 3 项取得最佳效果, 在其他 3 个任务中均接近 SOTA 方法, 也展现出了良好的性能. 其中, FMPO 在 antmaze-umaze-play-v2 任务和更为复杂、更具挑战性的 antmaze-large-diverse-v2 任务中显著超越了所有基线, 这表明 FMPO 在复杂环境中能展现更强大的性能. 实验所使用的硬件配置和软件环境详见表 3 和表 4.

从总体上看, FMPO 在 Gym 任务和 AntMaze 任务中的累计得分均位列第一, 其性能超越了现有的如 Diffusion-QL、DTQL 和 QIPO 等所有基线. 此外, FMPO 在各项任务中取得了高性能得分, 并在多数任务中均将方差控制在较小的范围内, 展示出优秀的算法稳定性.

4.3 算法性能与鲁棒性分析

为深入评估 FMPO 算法的性能, 本节在 D4RL 基准上对 15 个任务进行了实验, 每个任务均使用 4 个不同的随机种子重复运行. 图 2 展示了 4 个算法 (FMPO 与 3 个基线算法) 在 9 个代表性任务上的学习曲线, 3 个基线算法 TD3 + BC、CQL、IQL[49] 的结果基于开源库 CORL[50] 实现, 最终性能如图 3 和图 4 所示. 从最终性能图和学习曲线中不

表 1 FMPO 及基线方法在 D4RL 数据集上的性能表现

Table 1 Performance of FMPO and baseline methods on D4RL datasets

任务类型	TD3 + BC	BCQ	BEAR	CQL	SfBC	Diffusion-QL	DTQL	QIPO	FMPO (Ours)
halfcheetah-medium-v2	48.3	47.0	41.0	44.0	45.9 ± 2.2	51.1 ± 0.5	57.9 ± 0.13	54.16 ± 1.27	**68.70 ± 0.26**
hopper-medium-v2	59.3	56.7	51.9	58.5	57.1 ± 4.1	90.5 ± 4.6	99.6 ± 0.87	94.05 ± 13.27	**100.33 ± 0.25**
walker2d-medium-v2	83.7	72.6	80.9	72.5	77.9 ± 2.5	87.0 ± 0.9	89.4 ± 0.13	87.61 ± 1.46	**90.92 ± 1.82**
halfcheetah-medium-replay-v2	44.6	40.4	29.7	45.5	37.1 ± 1.7	47.8 ± 0.3	50.9 ± 0.11	48.04 ± 0.79	**53.34 ± 3.29**
hopper-medium-replay-v2	60.9	53.3	37.3	95.0	86.2 ± 9.1	101.3 ± 0.6	100.0 ± 0.13	101.25 ± 2.18	**101.37 ± 0.24**
walker2d-medium-replay-v2	81.8	52.1	18.5	77.2	65.1 ± 5.6	95.5 ± 1.5	88.5 ± 2.16	78.57 ± 26.09	**95.94 ± 2.46**
halfcheetah-medium-expert-v2	90.7	89.1	38.9	91.6	92.6 ± 0.5	96.8 ± 0.3	92.7 ± 0.20	94.45 ± 0.49	**97.66 ± 0.66**
hopper-medium-expert-v2	98.0	81.8	17.7	105.4	108.6 ± 2.1	111.1 ± 1.3	109.3 ± 1.49	108.02 ± 5.19	**112.79 ± 0.36**
walker2d-medium-expert-v2	110.1	109.5	95.4	108.8	109.8 ± 0.2	110.1 ± 0.3	110.0 ± 0.07	110.87 ± 1.04	**112.71 ± 0.22**
Gym 上的累计得分	677.4	602.5	411.3	698.5	680.3	791.2	798.3	777.02	**833.76**
antmaze-umaze-play-v2	91.3	0.0	73.0	84.8	92.0 ± 2.1	93.4 ± 3.4	94.8 ± 1.00	93.62 ± 7.05	**97.50 ± 4.33**
antmaze-umaze-diverse-v2	54.6	61.0	61.0	43.3	**85.3 ± 3.6**	66.2 ± 8.6	78.8 ± 1.83	76.12 ± 9.93	82.50 ± 19.20
antmaze-medium-play-v2	0.0	0.0	0.0	65.2	81.3 ± 2.6	76.6 ± 10.8	79.6 ± 1.80	80.00 ± 13.66	**82.50 ± 4.33**
antmaze-medium-diverse-v2	0.0	0.0	8.0	54.0	82.0 ± 3.1	78.6 ± 10.3	82.8 ± 1.71	**86.42 ± 5.44**	80.00 ± 7.07
antmaze-large-play-v2	0.0	6.7	0.0	18.8	**59.3 ± 14.3**	46.6 ± 8.3	52.0 ± 2.23	55.50 ± 29.39	57.50 ± 8.29
antmaze-large-diverse-v2	0.0	2.2	0.0	31.6	45.5 ± 6.6	56.6 ± 7.6	54.0 ± 2.23	32.13 ± 23.16	**62.50 ± 8.29**
AntMaze 上的累计得分	145.9	69.9	142.0	297.7	445.4	418.0	442.0	423.79	**462.50**
所有任务累计得分	823.3	672.4	553.3	996.2	1125.7	1209.2	1240.3	1200.81	**1296.26**

表 2　超参数设置
Table 2　Hyperparameter settings

	参数名称	取值
FMPO	流匹配时间采样分布	Unif([0, 1])
	流匹配模型步数	10
	迭代次数	1e6
	目标网络更新率 ρ	5e − 3
	策略噪声	0.2
	策略噪声截断	(−0.5, 0.5)
	策略更新频率	每 2 步更新一次
	折扣因子	0.99
	执行器学习率	3e − 4
	评价器学习率	3e − 4
网络结构	执行器/评价器隐藏层维度	256
	执行器/评价器层数	3
	激活函数	ReLU
	批次大小	256
	优化器	Adam

表 3　实验硬件配置
Table 3　Experimental hardware configuration

组件	配置
GPU	NVIDIA RTX 3090
CPU	12th Gen Intel(R) Core(TM) i7-12900K

表 4　实验软件环境
Table 4　Experimental software environment

软件	版本
Python	3.9.19
D4RL	1.1
Mujoco	3.1.5
Gym	0.23.1
Mujoco-py	2.1.2.14
PyTorch	1.13.1 + cu11.7

难发现: 1) FMPO 在各项任务中都取得了最优性能, 特别是在 halfcheetah-medium-v2、halfcheetah-medium-replay-v2 和 hopper-medium-v2 中表现尤为突出; 2) FMPO 在学习过程中不仅具有快速收敛的能力, 还具备良好的稳健性, 特别是在 hopper-medium-v2 和 walker2d-medium-expert-v2 任务中.

为了对 FMPO 算法作进一步分析, 本节重点比较了它与基线算法在中位数、均值、IQM (interquartile mean)、Optimality Gap 上的数值, 结果如图 3 所示. 从图中结果来看, FMPO 的中位数、均值以及 IQM 都显著优于其他基线, 说明 FMPO 算法的平均性能优异, 且具有很强的稳健性. 此外, 由 Optimality Gap 指标可知: 相较其他对比方法, FMPO 具有最小的最优性差距, 说明其学习到的策略性能更接近理论上的最优性能.

在算法可靠性评估上, 我们绘制了所有算法在 D4RL 任务上的累计性能分布, 结果如图 4(a) 所示. 其中, 横轴为归一化得分, 纵轴为得分高于该阈值的运行比例. 曲线位置越高, 说明算法在不同任务和随机种子下越能稳定地获得较高得分. 从结果图可以看出, FMPO 的累计性能分布曲线始终位于现有的离线强化学习基线上方, 即 FMPO 算法在各个性能阈值下都保持了更高的运行比例; 相比之下, TD3 + BC、IQL 与 CQL 的分布曲线比较接近而且位于 FMPO 曲线下方, 说明它们的算法总体性能相似且鲁棒性更差. 这表明 FMPO 不仅具有更强的平均性能, 同时在鲁棒性与稳定性方面也更具优势.

由上述实验结果可以得知, FMPO 不仅在性能上达到甚至超过了当前先进水平, 而且在不同任务条件和随机种子下均表现出强大的适应性与鲁棒性.

4.4　基于优势函数门控的行为克隆基线对比

为了进一步分析所提出的 flow matching 机制在策略优化中的作用, 本节构建了一个基于优势函数筛选的简化基线方法. 该方法直接利用优势函数对数据集中的动作进行筛选, 仅对具有正优势的动作样本进行行为克隆学习.

具体而言, 对于离线数据集 $\mathcal{D} = \{(s,\ a,\ r,\ s')\}$, 首先利用评价器网络和价值网络估计对应的优势函数 $A(s,\ a)$. 随后仅当 $A(s,\ a) > 0$ 时, 策略才对数据集动作进行模仿, 其优化目标可以形式化表示为:

$$\mathcal{L}_{\text{gate-BC}}(\theta) = \mathrm{E}_{(s,\ a)\sim\mathcal{D}}\left[\mathbf{1}(A(s,\ a) > 0)\ \|a - \pi_\theta(s)\|^2\right] \tag{19}$$

其中 $\mathbf{1}(\cdot)$ 表示指示函数. 当 $A(s,\ a) > 0$ 时, 策略 π_θ 对数据集中的动作进行模仿, 否则忽略该样本. 该方法的核心思想是利用优势函数筛选出具有较高价值的动作样本, 从而避免策略对低质量行为进行模仿. 从方法角度来看, 该基线可以视为本文方法的一个简化版本, 因为它仅利用优势函数对数据进行筛选, 而没有进一步对高价值动作分布进行建模.

相比之下, 本文提出的 FMPO 方法利用 flow matching 学习高价值动作的生成分布, 并通过生成式策略对优势区域进行建模, 从而能够在策略优化过程中产生新的高价值动作, 而不仅仅局限于模仿数据集中的已有动作. 因此, 该基线能够帮助我们分析 flow matching 模块在整体算法框架中的实际贡献.

实验结果如图 5 所示. 图中展示了 FMPO 与该基线方法在 9 个代表性任务上的学习曲线. 为了

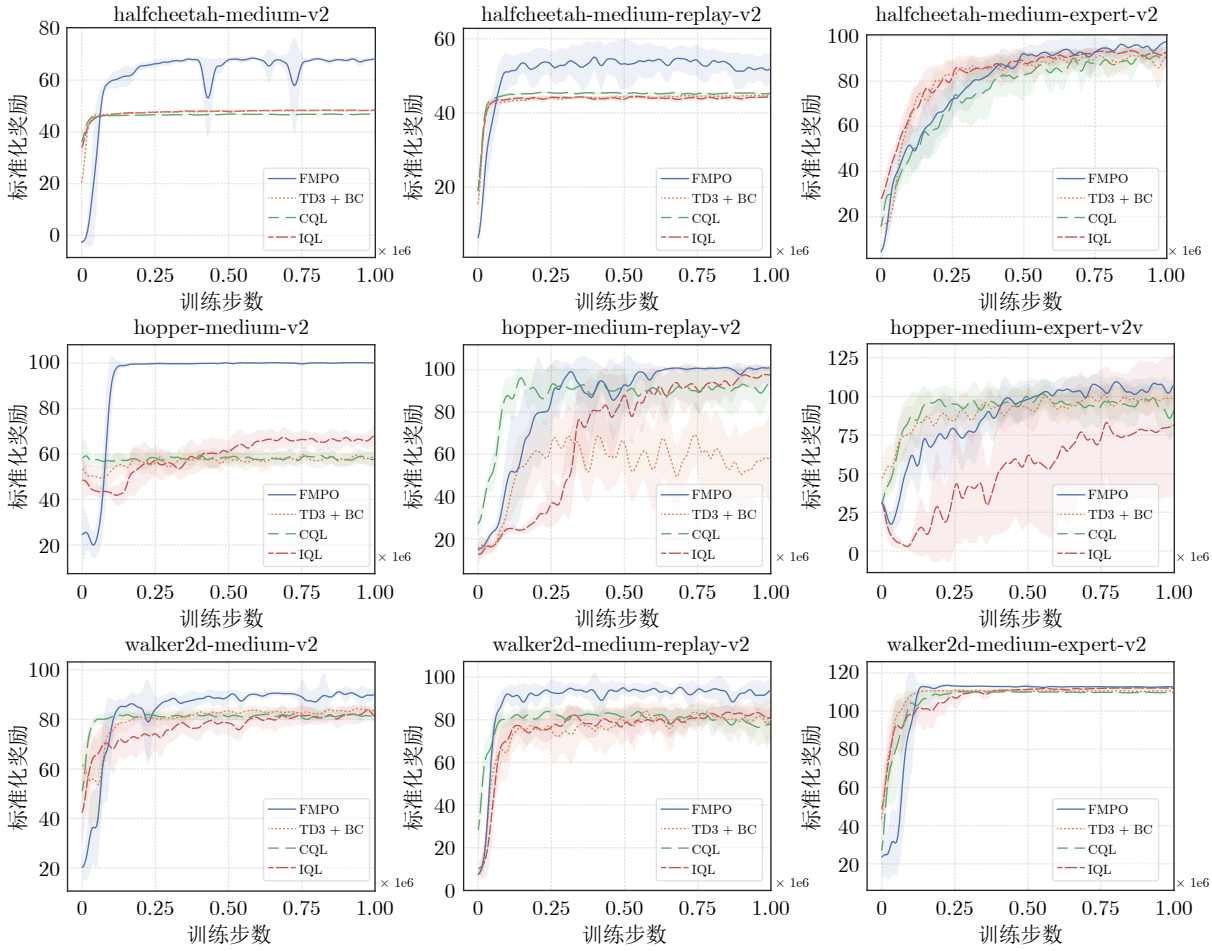


图 2 在 D4RL 基准的 9 个任务上进行的性能比较结果

Fig. 2 Performance comparison results on nine tasks of the D4RL benchmark

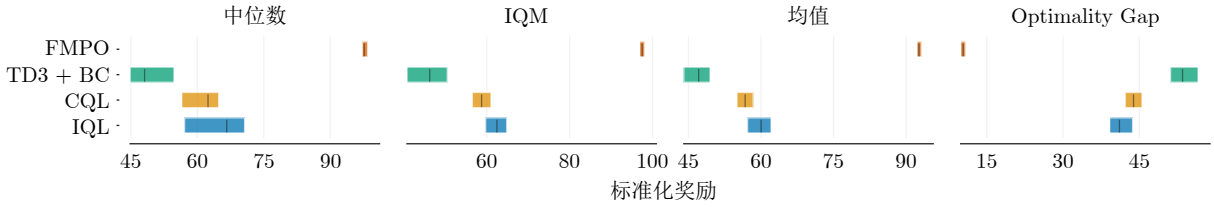


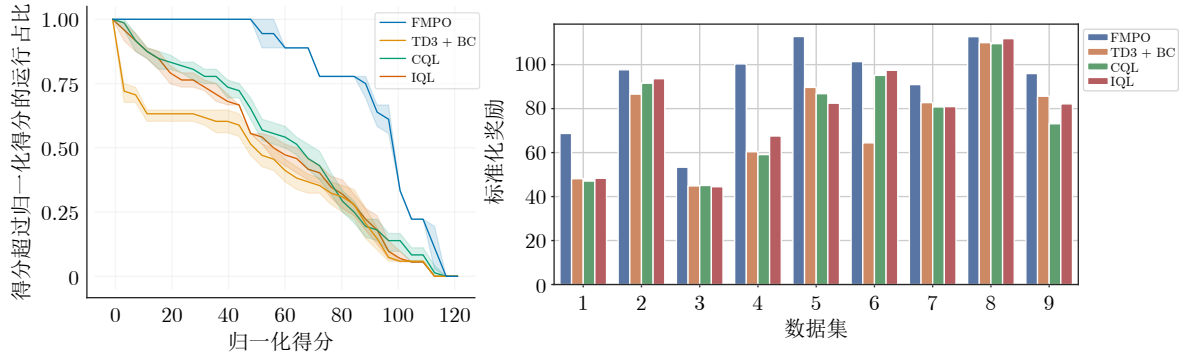
图 3 基于 D4RL 中 15 项任务的可靠性评估 (95% 置信区间)

Fig. 3 Reliability evaluation on 15 tasks in D4RL (95% confidence intervals)

保证实验的公平性, 基线方法除动作选择策略外, 其余训练流程均与 FMPO 保持一致. 实验结果表明, FMPO 在所有任务中均取得了更优的最终性能, 整体优于基于优势函数筛选的简化基线方法. 特别是在 walker2d-medium-v2、halfcheetah-medium-replay-v2 和 halfcheetah-medium-expert-v2 等任务上, FMPO 的性能提升更加明显. 这表明, 通过 flow matching 建模高价值动作分布能够生成新的高价值动作, 从而有效引导策略优化并实现更高效的策略提升.

4.5 超参数敏感实验

为了评估流匹配策略引导权重系数 ω_2 对算法性能的影响, 本文在 D4RL 数据集的 halfcheetah-medium-v2、hopper-medium-v2 和 walker2d-medium-v2 三个任务上进行了超参数敏感性分析. 该实验设置 ω_2 为 0.8、1.0 和 1.5 三种不同取值, 结果如图 6 所示. 在 halfcheetah-medium-v2 任务中, $\omega_2 = 1.5$ 的学习曲线平稳性最好, $\omega_2 = 1.0$ 次之, $\omega_2 = 0.8$ 的曲线波动最大, 但三者的学习曲线收敛后的最终性能相近. 结果表明: 在该任务中, 适当增



(a) 所有任务上的累计性能分布　　(b) 9 个任务上的最终性能比较

(a) Cumulative performance distribution over all tasks　　(b) Final performance comparison on nine tasks

注: 图 4(b) 中横轴 1, 2, 3, 4, 5, 6, 7, 8, 9 分别对应 halfcheetah-medium-v2, halfcheetah-medium-expert-v2, halfcheetah-medium-replay-v2, hopper-medium-v2, hopper-medium-expert-v2, hopper-medium-replay-v2, walker2d-medium-v2, walker2d-medium-replay-v2, walker2d-medium-expert-v2 数据集.

图 4　在 D4RL 基准任务上的性能与鲁棒性分析

Fig. 4　Performance and robustness analysis on D4RL benchmark tasks

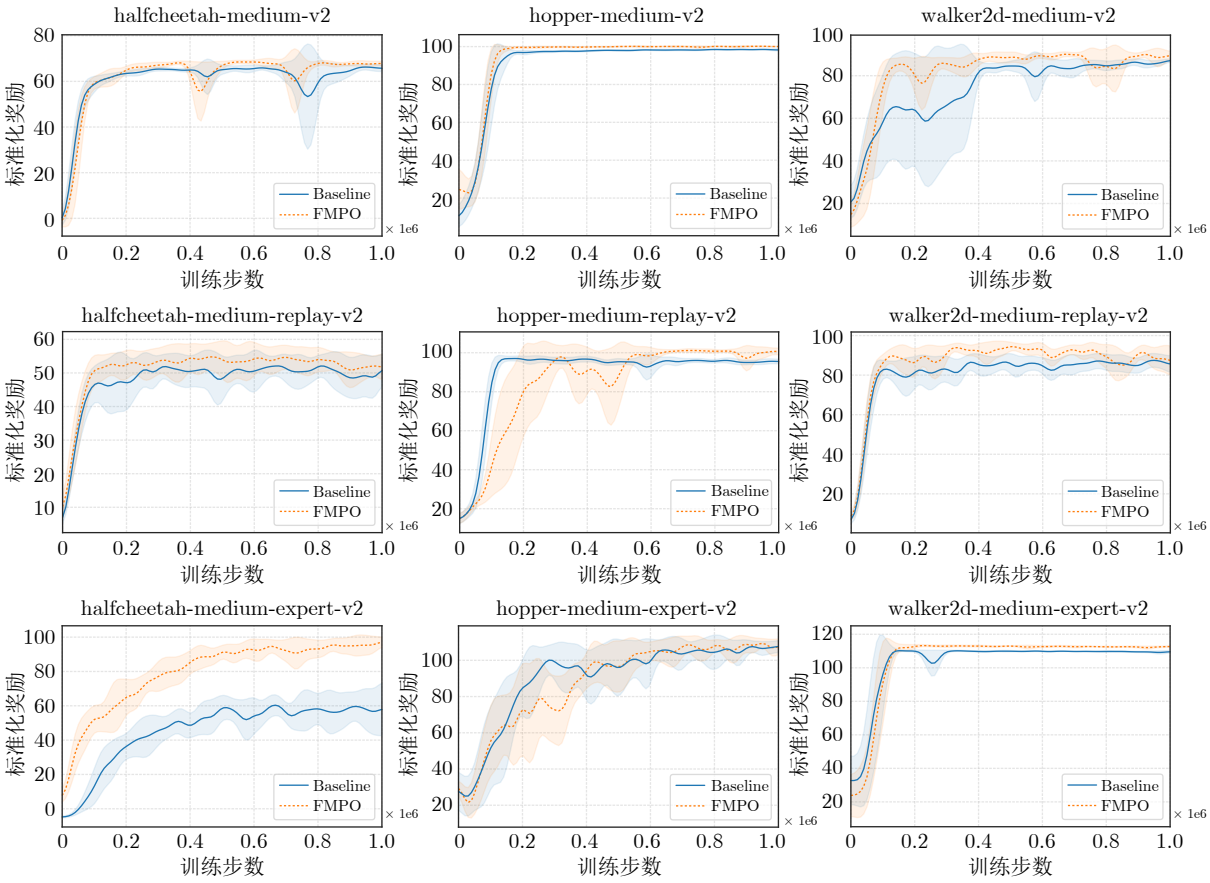


图 5　FMPO 与优势门控行为克隆基线在 9 个任务上的性能对比

Fig. 5　Performance comparison between FMPO and advantage-gated behavior cloning baseline on nine tasks

加高优势动作的权重有助于提升训练过程的稳定性. 在 hopper-medium-v2 任务中, 三条学习曲线都整体表现平稳, 其中 $\omega_2 = 1.0$ 的性能始终保持领先, $\omega_2 = 0.8$ 的性能略差但差距微小, 而 $\omega_2 = 1.5$ 的性能表现相对最差. 结果表明: 过高或者过低的高优势动作权重都会使算法性能降低, 该任务中存在一个适中的最优权重设置. 在 walker2d-medium-v2 任务中, $\omega_2 = 1.5$ 和 $\omega_2 = 0.8$ 的学习曲线均有较大

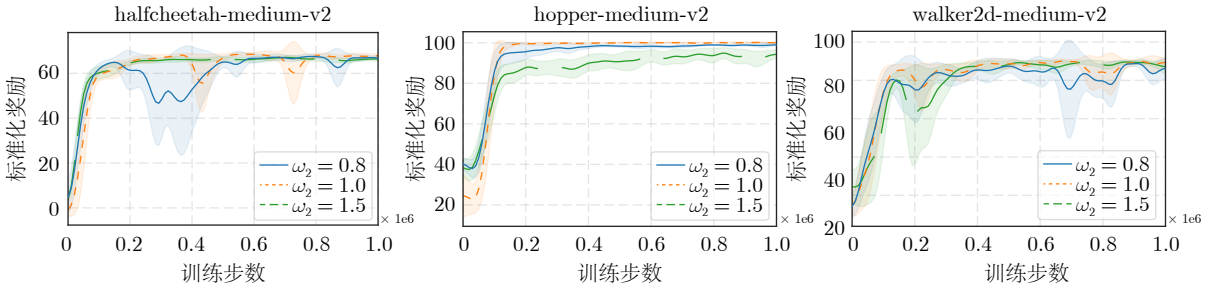


图 6 权重系数 ω_2 的超参数敏感性分析

Fig. 6 Hyperparameter sensitivity analysis of the weighting coefficient ω_2

的波动, $\omega_2 = 1.0$ 的学习曲线最稳定且性能最佳. 结果表明: 过高或者过低的高优势动作的权重都会造成算法的不稳定性, 该任务中存在一个适中的最优权重设置.

综合上述实验结果: $\omega_2 = 1.0$ 的权重参数在不同环境中均能保持良好的性能与高度稳健性. 具体来说, $\omega_2 = 1.0$ 的参数设置能使算法在绝大多数任务中接近或达到绝对最优性能, 且能有效避免参数取值不当可能带来的性能不稳定甚至退化的情况, 在策略提升效率、最终性能和训练稳定性之间实现平衡.

4.6 多目标迷宫可视化实验

为更直观深入地分析不同算法在离线数据条件下的策略学习行为, 本节设计了一个二维连续控制迷宫任务. 该任务在二维空间中进行, 采用连续动作空间. 智能体的观测由其位置和速度组成, 动作 $a \in [-1, 1]^2$ 表示在 x 与 y 方向上施加的线性力. 环境中设置三个目标位置, 分别位于 (1, 1)、(6, 1) 和 (1, 6), 对应奖励分别为 4、2 和 1. 任务目标是从预定义起点导航至回报最高的位置. 在该环境中构建离线数据集 $\mathcal{D} = \{(s_i, a_i, r_i, s'_i)\}_{i=1}^{M}$, 其中 $M = 100\,000$. 数据集分别由指向三个目标的轨迹组成, 到达 (1, 1)、(6, 1) 和 (1, 6) 的轨迹占比分别为 5%、50% 和 45%. 该数据集呈现明显的质量不均衡: 最优目标对应的高回报轨迹仅占少数, 绝大多数轨迹指向次优目标.

为了分析固定策略约束在低质量轨迹数据集上的影响, 本文对 FMPO 与 TD3 + BC 进行了对比实验. 两种方法均训练 500 000 步以保证充分收敛. 图 7(a) ~ 7(b) 展示了训练过程中策略评估时在最后 100 000 步产生的轨迹分布. 结果表明, TD3 + BC 学习得到的策略存在明显局限, 其策略分布主要集中在次优区域, 难以收敛至最优目标位置. 相比之下, FMPO 在相同数据条件下能够有效摆脱次优数据的影响, 并生成收敛到高回报目标的轨迹.

进一步分析发现, 这一差异主要源于 TD3 + BC 中的固定保守策略约束. 该约束要求学习策略在每个状态下均贴近数据集中已有动作, 从而迫使策略同时拟合高质量与低质量行为. 由于行为策略对低质量轨迹赋予更高的数据密度, 学习得到的策略会倾向于重复这些次优行为, 导致生成轨迹呈现出明显的同质化与重复性. 相比之下, FMPO 生成的轨迹表现出更高的价值, 并能够覆盖更广泛的高回报路径. 其关键原因在于 FMPO 能够生成更多高优势候选动作. 通过流匹配引导策略, FMPO 能够更有效地区分高优势动作与原始数据集中的行为动作, 从而显著提升行为策略的表达能力, 并为最终学习到的策略提供更加高效的动作引导. 该实验从策略约束层面对两种方法进行了进一步分析, 表明在低质量离线数据条件下, 过强的固定策略约束可能导致策略陷入次优行为模式, 而 FMPO 通过增强行为建模与高优势动作生成能力, 可以有效缓解这一问题.

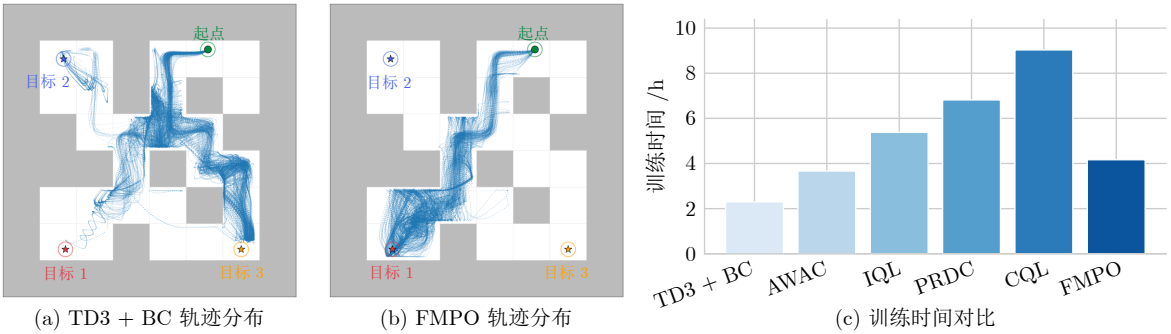
4.7 训练时间对比实验

为了更全面地评估各算法的计算开销, 本文在相同硬件环境下统计并报告了不同方法的训练时间. 所有实验均在相同的计算平台上运行, 并在相同的数据集和训练步数设置下进行, 以保证实验的公平性.

在实现方面, TD3 + BC、AWAC[51]、IQL 和 CQL 方法均采用了 CORL[49] 提供的实现代码[1]. PRDC[52] 则采用其论文作者提供的官方实现代码重新训练. 本文提出的 FMPO 方法在相同实验设置下进行训练与评估.

训练时间统计结果如图 7(c) 所示. 可以看到, 不同离线强化学习方法在训练时间上存在一定差异. 其中, FMPO 的整体训练时间约为 4 h, 明显低于 CQL 和 PRDC 等计算开销较大的离线强化学习方法, 同时仍然能够保持良好的策略性能表现. 这表明, 所提出的方法在保持较强性能的同时, 也具有较好的计算效率.

[1] https://github.com/tinkoff-ai/CORL



(a) TD3 + BC 轨迹分布
(a) TD3 + BC trajectory distribution

(b) FMPO 轨迹分布
(b) FMPO trajectory distribution

(c) 训练时间对比
(c) Training time comparison

图 7　多目标迷宫任务中的策略轨迹分布与计算效率对比

Fig. 7　Comparison of policy trajectory distributions and computational efficiency in multi-goal maze task

5　结束语

强化学习通过智能体与环境在线交互, 以迭代优化的方式学习最优策略, 近年来已经成为解决复杂决策控制任务的重要方式. 然而, 这种在线学习的方式可能带来极大的安全风险和高额的计算代价等问题, 限制了最优策略的学习, 从而影响强化学习在实际问题中的应用. 为了解决这一问题, 研究人员提出离线强化学习, 旨在通过固定的历史数据集实现对策略的学习. 但是其面临新的挑战, 即目标策略与行为策略差异所带来的分布偏移问题. 为了缓解分布偏移问题, 现有方法通过策略正则化约束目标策略与行为策略之间的偏离程度. 但是现有策略约束方法不能动态调整行为策略对目标策略的引导程度, 主要原因有两点: 策略约束缺乏自适应调整约束程度的方式、行为策略没有被很好地建模. 为此, 本文提出基于流匹配策略优化的离线强化学习方法, 通过结合优势函数引导流匹配模型学习高价值行为分布, 并利用生成式策略优化实现从行为策略到目标策略的有效改进, 从而提升离线策略学习性能.

尽管本文方法在离线强化学习任务中取得了良好的性能, 但仍存在一定局限性. 首先, 方法在一定程度上依赖于离线数据集的质量与覆盖度, 当高价值动作样本不足时, 生成模型学习到的动作分布可能受到数据分布限制. 其次, 引入生成模型模块虽然带来了更强的动作建模能力, 但在更高维动作空间或大规模任务中仍可能增加一定计算开销. 未来的研究可以结合数据增强或基于模型的轨迹生成方法, 以提升高价值样本覆盖度, 并探索更加高效的生成模型结构. 此外, 本文实验主要在 D4RL 基准上进行验证, 未来可以进一步在更接近真实场景的离线强化学习基准 (如 NeoRL[53]) 上进行评估, 以验证方法在真实复杂环境中的泛化能力. 同时, 将该方法扩展到多智能体强化学习以及多智能体具身智能场景[54−55], 例如在集中式训练、分布式执行框架下建模多智能体条件动作分布, 也是值得进一步探索的研究方向.

本文提出的流匹配策略自适应引导方法在 D4RL 基准上的实验中展示出优越的性能. 该研究成果为离线强化学习领域, 特别是与生成模型结合的领域提供了新的研究视角, 具有重要的研究意义和价值.

参考文献

1　Liu Quan, Zhai Jian-Wei, Zhang Zong-Chang, Zhong Shan, Zhou Qian, Zhang Peng, et al. A survey on deep reinforcement learning. *Chinese Journal of Computers*, 2018, **41**(1): 1−27
(刘全, 翟建伟, 章宗长, 钟珊, 周倩, 章鹏, 等. 深度强化学习综述. 计算机学报, 2018, **41**(1): 1−27)

2　Sun Chang-Yin, Mu Chao-Xu. Important scientific problems of multi-agent deep reinforcement learning. *Acta Automatica Sinica*, 2020, **46**(7): 1301−1312
(孙长银, 穆朝絮. 多智能体深度强化学习的若干关键科学问题. 自动化学报, 2020, **46**(7): 1301−1312)

3　Guo D Y, Yang D J, Zhang H W, Song J X, Wang P Y, Zhu Q H, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 2025, **645**(8081): 633−638

4　Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, Guez A, et al. Mastering the game of go without human knowledge. *Nature*, 2017, **550**(7676): 354−359

5　Vinyals O, Babuschkin I, Czarnecki W M, Mathieu M, Dudzik A, Chung J, et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 2019, **575**(7782): 350−354

6　Wu Xiao-Guang, Liu Shao-Wei, Yang Lei, Deng Wen-Qiang, Jia Zhe-Heng. A gait control method for biped robot on slope based on deep reinforcement learning. *Acta Automatica Sinica*, 2021, **47**(8): 1976−1987
(吴晓光, 刘绍维, 杨磊, 邓文强, 贾哲恒. 基于深度强化学习的双足机器人斜坡步态控制方法. 自动化学报, 2021, **47**(8): 1976−1987)

7　Zhang Wei-Zhen, He Zhen, Tang Zhang-Fan. Reinforcement learning control design for perching maneuver of unmanned aerial vehicles with wind disturbances. *Journal of Shanghai Jiao Tong University*, 2024, **58**(11): 1753−1761
(张威振, 何真, 汤张帆. 风扰下无人机栖落机动的强化学习控制设计. 上海交通大学学报, 2024, **58**(11): 1753−1761)

8　Levine S, Kumar A, Tucker G, Fu J. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. arXiv preprint arXiv: 2005.01643, 2020.

9　Huang Z H, Xu X, He H B, Tan J, Sun Z P. Parameterized batch reinforcement learning for longitudinal control of autonomous land vehicles. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2019, **49**(4): 730−741

10　Wang Xue-Song, Wang Rong-Rong, Cheng Yu-Hu. A review of offline reinforcement learning based on representation learning. *Acta Automatica Sinica*, 2024, **50**(6): 1104−1128

(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. 自动化学报, 2024, **50**(6): 1104−1128)

11 Chen Peng-Yu, Liu Shi-Rong, Duan Shuai, Duan Jun-Hong, Liu Yang. Gradient loss for offline reinforcement learning. *Acta Automatica Sinica*, 2025, **51**(6): 1218−1232
(陈鹏宇, 刘士荣, 段帅, 端军红, 刘扬. 基于梯度损失的离线强化学习算法. 自动化学报, 2025, **51**(6): 1218−1232)

12 Gu Yang, Cheng Yu-Hu, Wang Xue-Song. Offline reinforcement learning based on prioritized sampling model. *Acta Automatica Sinica*, 2024, **50**(1): 143−153
(顾扬, 程玉虎, 王雪松. 基于优先采样模型的离线强化学习. 自动化学报, 2024, **50**(1): 143−153)

13 Liu T L, Xu X, Xie X H, Lan Y X, Tang T, Ding Z Y. Intrinsic value-aligned policy optimization for offline-to-online reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems*, DOI: 10.1109/TNNLS.2026.3688490

14 Fujimoto S, Gu S S. A minimalist approach to offline reinforcement learning. In: Proceedings of the 35th International Conference on Neural Information Processing Systems. Curran Associates Inc., 2021. Article No. 1540

15 Fujimoto S, van Hoof H, Meger D. Addressing function approximation error in actor-critic methods. In: Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden: PMLR, 2018. 1587−1596

16 Jaques N, Ghandeharioun A, Shen J H, Ferguson C, Lapedriza G, Jones N, et al. Way off-policy batch deep reinforcement learning of implicit human preferences in dialog. arXiv preprint arXiv: 1907.00456, 2019.

17 Kostrikov I, Fergus R, Tompson J, Nachum O. Offline reinforcement learning with fisher divergence critic regularization. In: Proceedings of the 38th International Conference on Machine Learning. PMLR, 2021. 5774−5783

18 Kumar A, Fu J, Tucker G, Levine S. Stabilizing off-policy Q-learning via bootstrapping error reduction. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2019. Article No. 1055

19 Rouxel Q, Donoso C, Chen F, Ivaldi S, Mouret J B. Extremum flow matching for offline goal conditioned reinforcement learning. In: Proceedings of the IEEE-RAS 24th International Conference on Humanoid Robots. Seoul, Korea: IEEE, 2025. 981−988

20 Liu Z Y, Yang Z H, Xu J W, Yang R, Lv J F, Wang B X, et al. ADG: Ambient diffusion-guided dataset recovery for corruption-robust offline reinforcement learning. In: Proceedings of the 39th Conference on Neural Information Processing Systems. 2025. 163045−163079

21 Liu T L, Li J X, Zheng Y N, Niu H Y, Lan Y X, Xu X, et al. Skill expansion and composition in parameter space. In: Proceedings of the 13th International Conference on Learning Representations. Singapore, Singapore: OpenReview.net, 2025.

22 Fujimoto S, Meger D, Precup D. Off-policy deep reinforcement learning without exploration. In: Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA: PMLR, 2019. 2052−2062

23 Wu J L, Wu H X, Qiu Z H, Wang J M, Long M S. Supported policy optimization for offline reinforcement learning. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2022. Article No. 2268

24 Wang Z D, Hunt J J, Zhou M Y. Diffusion policies as an expressive policy class for offline reinforcement learning. In: Proceedings of the 11th International Conference on Learning Representations. Kigali, Rwanda: OpenReview.net, 2023.

25 Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2020. Article No. 574

26 Chen H Y, Lu C, Ying C Y, Su H, Zhu J. Offline reinforcement learning via high-fidelity generative behavior modeling. In: Proceedings of the 11th International Conference on Learning Representations. Kigali, Rwanda: OpenReview.net, 2023.

27 Hansen-Estruch P, Kostrikov I, Janner M, Kuba J G, Levine S. IDQL: Implicit Q-learning as an actor-critic method with diffusion policies. arXiv preprint arXiv: 2304.10573, 2023.

28 Park S, Li Q Y, Levine S. Flow Q-learning. In: Proceedings of the 42nd International Conference on Machine Learning. Vancouver, Canada: PMLR, 2025. 48104−48127

29 Liu T L, Li Y, Lan Y X, Gao H, Pan W, Xu X. Adaptive advantage-guided policy regularization for offline reinforcement learning. In: Proceedings of the 41st International Conference on Machine Learning. Vienna, Austria: PMLR, 2024. Article No. 1268

30 Mao Y X, Wang Q, Qu Y, Jiang Y H, Ji X Y. Doubly mild generalization for offline reinforcement learning. In: Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2024. Article No. 1628

31 Gao C X, Wu C Y, Cao M J, Xiao C J, Yu Y, Zhang Z Z. Behavior-regularized diffusion policy optimization for offline reinforcement learning. In: Proceedings of the 42nd International Conference on Machine Learning. Vancouver, Canada: PMLR, 2025. 18630−18657

32 Zhang H Y, Zhang S Y, Jin J X, Zeng Q X, Qiao Y F, Lu H C, et al. Balancing signal and variance: Adaptive offline RL post-training for VLA flow models. In: Proceedings of the 40th AAAI Conference on Artificial Intelligence. Singapore, Singapore: AAAI, 2026. 18755−18763

33 Lv L, Li Y F, Luo Y, Sun F C, Kong T, Xu J F, et al. Flow-based policy for online reinforcement learning. In: Proceedings of the 39th Conference on Neural Information Processing Systems. San Diego, America: NeurIPS, 2025. 93967−93990

34 Espinosa-Dice N, Zhang Y Y, Chen Y D, Guo B, Oertell O, Swamy G, et al. Scaling offline RL via efficient and expressive shortcut models. In: Proceedings of the 39th Conference on Neural Information Processing Systems. San Diego, America: NeurIPS, 2025. 2376−2414

35 Ki D, Oh J, Shim S W, Lee B J. Prior-guided diffusion planning for offline reinforcement learning. In: Proceedings of the 39th Conference on Neural Information Processing Systems. San Diego, America: NeurIPS, 2025. 59790−59820

36 Hu X M, Li S, Xu Y F, Tang B, Chen L. Enhancing diffusion policies with distribution-matching generator in offline reinforcement learning. In: Proceedings of the 40th AAAI Conference on Artificial Intelligence. Singapore, Singapore: AAAI, 2026. 21894−21902

37 Fang L J J, Liu R X, Zhang J, Wang W J, Jing B Y. Diffusion actor-critic: Formulating constrained policy iteration as diffusion noise regression for offline reinforcement learning. In: Proceedings of the 13th International Conference on Learning Representations. Singapore, Singapore: OpenReview.net, 2025.

38 Chen T Y, Wang Z D, Zhou M Y. Diffusion policies creating a trust region for offline reinforcement learning. In: Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2024. Article No. 1585

39 Mao L Y, Xu H R, Zhan X Y, Zhang W N, Zhang A. Diffusion-DICE: In-sample diffusion guidance for offline reinforcement learning. In: Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2024. Article No. 3136

40 Dong Z B, Hao J Y, Yuan Y F, Ni F, Wang Y T, Li P Y, et al. DiffuserLite: Towards real-time diffusion planning. In: Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2024. Article No. 3895

41 Chen H Y, Zheng K W, Su H, Zhu J. Aligning diffusion behaviors with Q-functions for efficient continuous control. In: Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2024. Article No. 3812

42 Ma Y C, Liu T L, Lan Y X, Yin X, Zhang C X, Zhang X L, et al. Diffusion policies with value-conditional optimization for offline reinforcement learning. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems. Hangzhou, China: IEEE, 2025. 13468−13475

43 Chen H Y, Lu C, Wang Z Y, Su H, Zhu J. Score regularized policy optimization through diffusion behavior. In: Proceedings of the 12th International Conference on Learning Representations. Vienna, Austria: OpenReview.net, 2024.

44 Ding Z H, Jin C. Consistency models as a rich and efficient policy class for reinforcement learning. In: Proceedings of the

12th International Conference on Learning Representations. Vienna, Austria: OpenReview.net, 2024.

45 Espinosa-Dice N, Brantley K, Sun W. Expressive value learning for scalable offline reinforcement learning. arXiv preprint arXiv: 2510.08218, 2025.

46 Zhang S Y, Zhang W T, Gu Q Q. Energy-weighted flow matching for offline reinforcement learning. In: Proceedings of the 13th International Conference on Learning Representations. Singapore, Singapore: OpenReview.net, 2025.

47 Lipman Y, Chen R T Q, Ben-Hamu H, Nickel M, Le M. Flow matching for generative modeling. In: Proceedings of the 11th International Conference on Learning Representations. Kigali, Rwanda: OpenReview.net, 2023.

48 Kumar A, Zhou A, Tucker G, Levine S. Conservative Q-learning for offline reinforcement learning. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2020. Article No. 100

49 Kostrikov I, Nair A, Levine S. Offline reinforcement learning with implicit Q-learning. arXiv preprint arXiv: 2110.06169, 2021.

50 Tarasov D, Nikulin A, Akimov D, Kurenkov V, Kolesnikov S. CORL: Research-oriented deep offline reinforcement learning library. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2023. Article No. 1351

51 Nair A, Gupta A, Dalal M, Levine S. AWAC: Accelerating online reinforcement learning with offline datasets. arXiv preprint arXiv: 2006.09359, 2020.

52 Ran Y H, Li Y C, Zhang F X, Zhang Z Z, Yu Y. Policy regularization with dataset constraint for offline reinforcement learning. In: Proceedings of the 40th International Conference on Machine Learning. Honolulu, USA: PMLR, 2023. 28701−28717

53 Qin R J, Zhang X Y, Gao S Y, Chen X H, Li Z W, Zhang W N, et al. NeoRL: A near real-world benchmark for offline reinforcement learning. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2022. Article No. 1795

54 Yuan L, Zhang Z Q, Li L H, Guan C, Yu Y. A survey of progress on cooperative multi-agent reinforcement learning in open environment. arXiv preprint arXiv: 2312.01058, 2023.

55 Feng Z H, Xue R Q, Yuan L, Yu Y, Ding N, Liu M Q, et al. Multi-agent embodied AI: Advances and future directions. *Science China Information Sciences*, 2026, **69**(5): Article No. 151202

刘腾龙 国防科技大学智能科学学院博士研究生. 主要研究方向为强化学习、生成模型与机器人学习.
E-mail: ltl@nudt.edu.cn
(**LIU Teng-Long** Ph.D. candidate at the College of Intelligence Science and Technology, National University of Defense Technology. His research interests include reinforcement learning, generative models and robot learning.)

谢旭辉 国防科技大学智能科学学院博士研究生. 主要研究方向为多任务强化学习.
E-mail: xiexuhui20@nudt.edu.cn
(**XIE Xu-Hui** Ph.D. candidate at the College of Intelligence Science and Technology, National University of Defense Technology. His main research interest is multi-task reinforcement learning.)

黄佳成 国防科技大学智能科学学院博士研究生. 主要研究方向为模仿学习与 VLA 模型.
E-mail: huangjc@nudt.edu.cn
(**HUANG Jia-Cheng** Ph.D. candidate at the College of Intelligence Science and Technology, National University of Defense Technology. His research interests include imitation learning and VLA model.)

郭道洁 国防科技大学智能科学学院硕士研究生. 主要研究方向为多任务强化学习.
E-mail: guodaojie@nudt.edu.cn
(**GUO Dao-Jie** Master student at the College of Intelligence Science and Technology, National University of Defense Technology. His main research interest is multi-task reinforcement learning.)

兰奕星 国防科技大学智能科学学院讲师. 2023 年获得国防科技大学智能科学学院控制科学与工程专业博士学位. 主要研究方向为强化学习及其在具身智能系统中的应用. 本文通信作者.
E-mail: lanyixing16@nudt.edu.cn
(**LAN Yi-Xing** Lecturer at the College of Intelligence Science and Technology, National University of Defense Technology. He received his Ph.D. degree in control science and engineering from the College of Intelligence Science and Technology, National University of Defense Technology in 2023. His research interests include reinforcement learning and its applications in embodied intelligence systems. Corresponding author of this paper.)

徐 昕 国防科技大学智能科学学院研究员. 2002 年获得国防科技大学机电与自动化学院控制科学与工程专业博士学位. 主要研究方向为智能控制, 强化学习, 机器学习, 机器人和智能车辆.
E-mail: xinxu@nudt.edu.cn
(**XU Xin** Professor at the College of Intelligence Science and Technology, National University of Defense Technology. He received his Ph.D. degree in control science and engineering from the College of Mechatronics and Automation, National University of Defense Technology in 2002. His research interests include intelligent control, reinforcement learning, machine learning, robotics, and autonomous vehicles.)