arXiv:2602.02847v1 [cs.LG] 2 Feb 2026

Caus面向稳健离线强化学习的 Flow Q-Learning g

Mingxuan Li [1] **Junzhe Zhang** [2] **Elias Bareinboim** [1]

摘要

近年来，基于流匹配的高表达能力策略凭借其从离线数据中建模复杂动作分布的能力，已成功应用于强化学习（RL）。这些算法建立在标准策略梯度之上，而标准策略梯度假设数据中不存在未测量的混杂因素。然而，当示范者与学习者的感知能力不匹配时，这一条件在基于像素的演示中未必成立，从而导致离线数据中出现隐式混杂偏差。我们从因果视角研究离线强化学习中混杂观测这一问题，以应对这一挑战。我们提出了一种新颖的因果离线强化学习目标，通过优化策略在混杂偏差可能造成影响时的最坏情况性能。基于这一新目标，我们提出了一种实用的实现方法，从混杂演示中学习高表达能力的流匹配策略，并采用深度判别器评估目标策略与名义行为策略之间的差异。在 25 个基于像素的任务中进行的实验表明，我们提出的抗混杂增强方法所取得的成功率，是未考虑混杂因素的先进离线强化学习方法的 120%。

1. 引言

离线强化学习（RL）为传统在线强化学习提供了一种替代范式，使智能体能够利用先前收集的观测数据有效地学习决策策略，尤其是在底层环境中的主动探索代价高昂、不安全甚至不可行时（Lange 等人，2012；Levine 等人，2020年）。标准的离线强化学习问题可解释为一种受约束的优化问题——智能体在与观测轨迹的状态-动作分布保持接近的同时，力求最大化长期奖励（Levine 等人，2020年）。近年来，越来越多的离线强化学习研究尝试使用表达能力强的策略类，通过深度生成建模显式学习策略的速度场，以表示更复杂的多模态策略分布（Mandlekar 等人，2022；Lipman 等人，2024）。这类表达能力强的策略类包括去噪扩散（Wang 等人，2023年）和流匹配（Park 等人，2025b）。它们能够对数据集中的复杂策略分布进行建模，从而施加准确的行为约束；这对离线强化学习算法至关重要（Tarasov等人，2023年）。

尽管取得了这些进展，上述所有算法都建立在标准的离策略策略梯度框架之上（Sutton，2018年），该框架依赖一个关键假设：示范者的行为策略与学习者的目标策略在状态-动作域上具有相同的支持集（Precup 等人，2000年）。当这些条件不成立时，观测数据中可能会引入隐性混杂偏差，进而给利用离线数据训练强化学习系统带来重大挑战（Bareinboim 等人，2024年；Li 等，2025b）。

示例 1 (混杂的基于像素的观测)。图1展示了一个例子，说明离线强化学习中的隐式混杂偏差所带来的挑战。具体而言，专家动作是根据真实状态观测采样得到的，而最终的基于像素的离线数据集则将这些动作与相应的图像观测配对。机械臂的任务是将彩色立方体移动到指定的半透明位置。然而，图1b中的像素观测是带有遮挡的模糊正面视图，无法完整呈现图1a所示、且构成专家决策依据的真实状态信息。在马尔可夫决策过程假设下利用此类数据集进行学习，会导致性能显著下降；尽管同一方法在示范者所使用的结构化状态观测空间中执行相同任务时效果良好（Park 等人，2025b；Dong 等人，2025）。■

决策中的混杂偏差问题，一直以来都在因果推断（Pearl，2009年）以及近年来的因果强化学习（Bareinboim 等人，2024年）领域中得到研究。大多数方法都会探索关于底层因果机制的额外先验知识；在这些机制下，目标

[1]因果人工智能实验室，哥伦比亚大学，美国纽约州[2]电气工程与计算机科学系，雪城大学，美国纽约州。通信作者：Mingxuan Li <ml@cs.columbia.edu>。

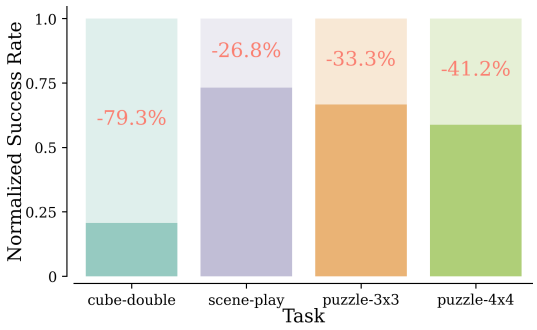
预印本。*2026*年*2*月*4*日。



(a) 状态观测。



*(b)*像素观测。



(c) 性能下降。

图 *1.* **离线数据中的混杂偏差会导致性能损失。** 左中：对于离线强化学习数据集中的基于像素的任务，专家动作依据真实状态向量采样，但在离线学习阶段以像素观测的形式呈现。右：尽管使用了图像增强和强大的神经编码器，最先进的离线强化学习算法在像素观测下的性能仍急剧下降。

策略可由观测数据唯一确定（即具有可识别性）（Pearl & Robins，1995年；Bareinboim 等人，2024年）。当因果知识不可得或不足以确保可识别性时，可以将部分识别方法与标准强化学习算法结合，以推断与观测相容的最坏情况环境中的最优动作（Kallus& Zhou，2018年；Zhang& Bareinboim，2019年；Namkoong et al.，2020年；Zhang& Bareinboim，2024年；Li 等，2025a；b；Hess et al.，2025年）。函数逼近器已用于根据混杂观测评估复杂状态、简单离散动作领域环境中的最坏情况策略回报。然而，如何在具有连续动作的表达能力强的策略类中，利用混杂的离线观测学习鲁棒策略，仍面临重大挑战。[1]

我们通过形式化离线强化学习中的混杂问题，提出一种新颖的离线强化学习目标，并采用流匹配对观测轨迹中复杂的状态-动作分布进行建模，从而克服这一挑战。具体而言，我们的贡献如下：(1) 我们提出一种新颖的离线强化学习目标，使其能够抵御离线数据中隐含的未观测混杂。优化这一目标可在最坏情况环境中得到一种能够选择最优动作的安全策略。(2) 我们提出一种名为 因果流 Q 学习(CFQL) 的算法，利用流匹配对观测到的状态-动作分布建模，从而在表达能力强的策略类上优化新提出的抗混杂目标。我们通过大量实验评估 CFQL。结果表明，在 25 个基于像素的离线强化学习任务中，成功率平均提升了 120%，其中一些任务甚至超过了使用真实状态观测的策略。受篇幅限制，所有证明均见第 C 节。

记号。 我们将始终使用大写字母（V）表示随机变量，小写字母（v）表示其取值，并用花体 $\mathcal{V}$ 表示其定义域。我们使用粗体大写字母（$\boldsymbol{V}$）表示随机变量集合，并令 $|\boldsymbol{V}|$ 表示集合 $\boldsymbol{V}$ 的基数。固定索引 $i, j \in \mathbb{N}$。令 $\boldsymbol{X}_{i:j}$ 表示序列 $\{X_i, X_{i+1}, \ldots, X_j\}$。我们用 $P(\boldsymbol{X})$表示变量 $\boldsymbol{X}$上的概率分布。我们始终将 $P(\boldsymbol{x})$用作概率$P(\boldsymbol{X} = \boldsymbol{x})$的简写； $P(\boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{X} = \boldsymbol{x})$之于 $= P(\boldsymbol{y} \mid \boldsymbol{x})$亦然。最后， $\mathbf{1}_{\boldsymbol{Z}=\boldsymbol{z}}$ 是一个指示函数，若事件$\boldsymbol{Z} = \boldsymbol{z}$ 成立则返回1，否则返回0。

2. 混杂稳健策略学习

我们研究一个序贯决策设定，其中智能体选择一个动作序列，以优化后续奖励。我们假设一个混杂马尔可夫决策过程（CMDP），以显式建模离线强化学习中未观测混杂所带来的挑战（Zhang& Bareinboim，2022；Bennett 等，2021；Kallus& Zhou，2020年；Zhang&Bareinboim，2025年；Li 等，2025a；b）。[2]

定义2.1. 混杂马尔可夫决策过程（CMDP） $\mathcal{M}$ 是由 $\langle \mathcal{S}$、$\mathcal{X}$、$\mathcal{Y}$、$\mathcal{U}$、$\mathcal{F}$、$P\rangle$ 组成的元组，其中（1）$\mathcal{S}$、$\mathcal{X}$、$\mathcal{Y}$ 分别是观测状态、动作和奖励的空间；（2）$\mathcal{U}$ 是未观测外生噪声的空间；（3）$\mathcal{F}$ 是由转移函数f_S：$\mathcal{S} \times \mathcal{X} \times \mathcal{U} \mapsto \mathcal{S}$、行为策略 f_X： $\mathcal{S} \times \mathcal{U} \mapsto \mathcal{X}$和奖励函数 f_Y： $\mathcal{S} \times \mathcal{X} \times \mathcal{U} \mapsto \mathcal{Y}$组成的集合；（4） P是领域 $\mathcal{U}$上的外生分布。

除非另有说明，本文均假设动作域 $\mathcal{X}$为连续的，而状态域 $\mathcal{S}$ 可能复杂且连续；奖励域 $\mathcal{Y}$有界于实区间 $[a, b] \subset \mathbb{R}$内。考虑一个与 CMDP $\mathcal{M}$交互并生成离策略数据的示范智能体。对于每个时间步 $t = 1, \ldots, T$，环境首先从分布 $P(\mathcal{U})$ 中抽取外生噪声 U_t；随后，示范者执行动作 $X_t \leftarrow f_X(S_t, U_t)$，获得后续奖励 $Y_t \leftarrow r_t(S_t, X_t, U_t)$，并转移到下一个状态 $S_{t+1} \leftarrow$

[1]有关离线强化学习与因果决策的更详细讨论，请读者参阅第A节。

[2]参见第C 节，了解因果基础以及将 CMDP 作为马尔可夫决策过程放宽形式的详细讨论。

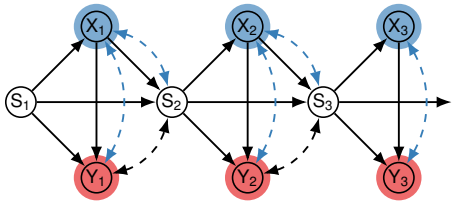


图 2。 **表示混杂马尔可夫决策过程中数据生成机制的因果图。**双向箭头表示专家策略使用但学习者在离线数据集中无法观测到的信息。

$f_S(S_t,\ X_t,\ U_t)$。从学习者的视角来看，示范者的观测轨迹可概括为观测分布 $P(\boldsymbol{X}_{1:T},\ \boldsymbol{S}_{1:T},\ \boldsymbol{Y}_{1:\ T})$。

图 2 展示了一个图形表示（Bareinboim 等，2022），说明混杂马尔可夫决策过程中观测数据的生成过程。更具体地说，实心节点表示观测变量 X_t、S_t、Y_t，箭头表示它们之间的函数关系 f_X、f_S、f_Y。按照惯例，外生变量 U_t 通常不会在图中明确显示；双向箭头 $X_t \longleftrightarrow Y_t$ 和 $X_t \longleftrightarrow S_{t+1}$ 表示存在一个未观测混杂因素（UC）U_t，同时影响动作、状态和奖励。这些双向箭头（以蓝色突出显示）表示离策略数据中动作 X_t、奖励 Y_t 和状态 S_{t+1} 之间存在未观测混杂因素，违反了无未观测混杂因素条件（NUC）（Robbins，1985；Bareinboim 等，2024）。正如上文所述，此类违反会给离线强化学习带来挑战。

CMDP $\mathcal{M}$ 中的策略 π 是一个决策规则 $\pi(x_t \mid s_t)$，它将状态映射为动作域$\mathcal{X}$上的一个分布。干预 do(π) 是一种操作，它用策略 π替换 CMDP $\mathcal{M}$ 中的行为策略 f_X。令 $\mathcal{M}_\pi$ 为干预 do(π) 所诱导的子模型。干预分布 $P_\pi(\boldsymbol{X}_{1:T},\ \boldsymbol{S}_{1:T},\ \boldsymbol{Y}_{1:\ T})$ 定义为 $\mathcal{M}_\pi$中观测变量上的分布，

$$P_\pi(\bar{\boldsymbol{x}}_{1:T}, \bar{\boldsymbol{s}}_{1:T}, \bar{\boldsymbol{y}}_{1:T}) = P(s_1)\prod_{t=1}^{T}\Big(\pi(x_t \mid s_t)\mathcal{T}(s_t, x_t, s_{t+1})\mathcal{R}(s_t, x_t, y_t)\Big) \quad (1)$$

其中，转移分布 $\mathcal{T}$和奖励分布 $\mathcal{R}$ 由以下内容给出，对于 $t = 1, \ldots, T$，

$$\mathcal{T}(s_t, x_t, s_{t+1}) = \int_{\mathcal{U}} \mathbf{1}_{s_{t+1}=f_S(s_t,x_t,u_t)} P(u_t), \quad (2)$$

$$\mathcal{R}(s_t, x_t, y_t) = \int_{\mathcal{U}} \mathbf{1}_{y_t=f_Y(s_t,x_t,u_t)} P(u_t). \quad (3)$$

为方便起见，我们将奖励函数 $\mathcal{R}(s, x)$写为期望值 $\sum_y y\mathcal{R}(s, x, y)$。固定折扣因子$\gamma \in [0, 1]$。智能体的一个常见目标是优化其累积回报 $R_t = \sum_{i=0}^{\infty} \gamma^i Y_{t+i}$。

离线强化学习。 在离线强化学习中，当不存在会在动作与后续结果之间引入虚假相关的未观测混杂因素时，可以从离线数据中识别转移分布 $\mathcal{T}$和奖励函数$\mathcal{R}$ 的参数化形式。这意味着候选策略的期望回报可以从离线数据的采样过程中估计出来。随后，通过最大化所估计的回报（在适当的行为正则化下），即可获得最优策略。

例如，令状态-动作价值函数 $Q_\phi(s,\ x) = \mathbb{E}_\pi[R_t \mid S_t = s, X_t = x]$ 表示目标策略 policy π 在给定状态 s 和动作 x以及参数 ϕ时的期望回报。令 θ 表示 π 的参数。带行为正则化的基本离线强化学习目标定义如下（Wu 等人，2019年；Fujimoto 和 Gu，2021；Tarasov等人，2023年）：

$$\mathcal{L}(\theta) = \mathbb{E}_{\substack{x,s\sim\mathcal{D},\\ x'\sim\pi_\theta}}\Big[-Q_\phi(s, x') - \alpha \log \pi(x|s)\Big], \quad (4)$$

其中 $\mathcal{D}$ 是离线观测；α 是控制行为克隆（BC）正则项强度的一些系数。Q值评论家 Q_ϕ 通过离线数据集上的经验 Bellman 损失进行训练。目标策略通过对式（4）中的目标应用基于梯度的优化来学习。当数据满足 NUC条件时，在标准凸性假设下，可以保证目标策略得到改进并收敛至最优解（Zhang 等人，2020年）。

因果强化学习。然而，NUC条件在现实应用中并非总是成立。当示范者和学习者的感知能力不同时，可能会产生混杂偏差，从而给离线学习算法带来挑战。

示例 2（混杂像素观测（续））。再次考虑图 1 中的示例，该示例展示了离线强化学习数据中存在的隐式混杂偏差。图 1c 展示了基于像素训练的流策略的回报，并与在相同任务上使用结构化状态观测训练的策略进行比较（Park 等人，2025b；Dong 等人，2025年）。我们看到各项任务的性能均呈现一致的下降趋势，其中一些甚至下降了近 80%。基于像素的观测下的所有成功率，均以同一算法在结构化状态观测下的性能为基准进行归一化，并在 OGBench（Park 等人，2025a）中每个类别的全部 5 项任务上取平均。■

近年来，因果推断领域（Kallus &Zhou，2018年；Zhang & Bareinboim，2019年；Joshi 等人，2024b；Namkoong 等人，2020年；Li 等，2025b）和安全强化学习领域（Kumar 等人，2020年；Nakamoto 等人，2023年）不断涌现相关研究，旨在应对策略学习中的数据偏差和分布偏移挑战。与我们的设定最为接近的是，Zhang &Bareinboim（2025年）解决了这些挑战

通过推导出一种新颖的因果贝尔曼方程，为混杂观测下的价值函数提供下界，从而应对离策略评估中的混杂偏差。具体而言，策略 π 的价值函数 $Q_\pi(s,\ x)$ 由函数 $Q_\pi(s,\ x)$ 给出的下界约束，

$$\underline{Q_\pi}(s,x) = \mu(\neg x \mid s)\Big(a + \gamma \min_{s'} \underline{V_\pi}(s')\Big) + \mu(x \mid s)\Big(\widetilde{\mathcal{R}}(s,x) + \gamma \sum_{s',x'} \widetilde{\mathcal{T}}(s,x,s')\,\underline{V_\pi}(s')\Big) \tag{5}$$

其中，状态价值下界 $V_\pi(s) = \sum_x \pi(x \mid s) Q_\pi(\{v6, x)$ ，且 $P(\neg x \mid s) = 1 - P(x \mid s)$；$\mathcal{T}$和 $\mathcal{R}$ 分别是根据观测分布计算得到的名义转移分布和奖励函数，即

$$\widetilde{\mathcal{T}}(s,x,s') = P\left(S_{t+1} = s' \mid S_t = s, X_t = x\right), \tag{6}$$

$$\widetilde{\mathcal{R}}(s,x) = \mathbb{E}\left[Y_t \mid S_t = s, X_t = x\right]. \tag{7}$$

随后，学习者可以通过在动作空间上迭代优化上述价值函数下界，从受混杂因素影响的离线数据中获得有效策略（Li 等，2025b）。然而，现有的因果贝尔曼方程仅适用于离散动作，因此无法处理复杂的多模态连续动作分布。本文余下部分将致力于解决这一挑战。

3. 因果流 Q 学习

本节将介绍一种因果离线强化学习目标，并在此基础上提出因果流 Q 学习。

我们注意到，对于任意策略 π，状态价值函数的下界为 $V_\pi(s) = \sum_x \pi(x \mid s) \underline{Q_\pi}(s,\ x)$，其中$Q_\pi(s,\ x)$ 由式（5）给出。经过一些简化，可以推导出闭式解，如下所示。

定理 3.1。 对于奖励信号为 $Y_t \in [a,b] \subseteq \mathbb{R}$ 的 *CMDP* 环境 $\mathcal{M}$，固定策略 π。对于任意状态 $s \in \mathcal{S}$，状态价值函数 $V_\pi(s) \geq V_\pi(s)$ 的下界 $\underline{V_\pi}(s)$ 如下所示，

$$\underline{V_\pi}(s) = \mathbb{E}_{\substack{x \sim p(\cdot|s) \\ x', x^* \sim \pi(\cdot|s)}} \Big[\mathbf{1}_{x \neq x'} \Big(a + \gamma \min_{s^*} \underline{Q_\pi}(s^*, x^*)\Big) + \mathbf{1}_{x = x'} \Big(\widetilde{\mathcal{R}}(s,x) + \gamma \sum_{s'} \widetilde{\mathcal{T}}(s,x,s')\,\underline{Q_\pi}(s', x^*)\Big)\Big] \tag{8}$$

图3展示了说明此更新步骤的备份示意图。与标准贝尔曼最优方程（Bellman，1966）一样，公式（8）使用当前的最优价值函数估计值递归更新状态价值函数。

另一方面，公式（8）明确考虑了混杂观测的离策略特性：当行为策略采取与目标策略相同的动作 $x_t = x$ 时，更新遵循标准贝尔曼方程，并使用下一个采样状态 s_t；当采样动作 $x_t \neq x$ 与目标动作不同时，我们的算法则改为使用

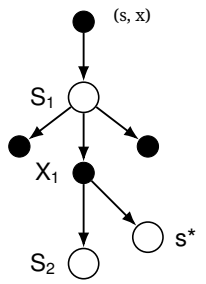


图 3。 因果策略梯度的备份示意图。

与下一个最坏情况或最好情况状态 s^* 相关联的价值函数，分别对应下界和上界的估计。

定理 3.1 给出了策略期望回报的下界 π，该策略针对一个与混杂观测相容的最坏情况 CMDP 实例进行优化 $\mathcal{M}$ 。最大化定理 3.1 中的下界，可在底层环境中得到具有性能保证的安全策略。在因果离线强化学习目标中，该下界用作式（4）中的 Q 损失，并可通过函数近似器计算。具体而言，令 θ 表示目标策略的参数 π，并令 $\mathcal{D}$ 表示离线观测的有限集合。于是，可根据观测数据估计因果离线强化学习目标，如下所示：

$$\mathcal{L}(\theta) = \mathbb{E}_{\substack{(s,x,y,s') \sim \mathcal{D} \\ x', x^* \sim \pi_\theta(\cdot|s)}} \Big[D(s,x,x')\Big(y + \gamma \underline{Q_{\pi_\theta}}(s', x^*)\Big) + (1 - D(s,x,x'))\Big(a + \gamma \min_{s^*} \underline{Q_{\pi_\theta}}(s^*, x^*)\Big)\Big] \tag{9}$$

在上述各个量中，$D(s,\ x,\ x')$ 是一个判别器，用于近似状态 s 处的指示函数 $\mathbf{1}_{x=x'}$。它以状态—动作对 $s,\ x$ 为输入；如果策略分布 $\pi_\theta(x|s)$ 与离线数据一致，则返回 1。 x^* 是一个重载记号，通常表示从 π_θ 中采样得到的下一步动作。我们将在下文进一步讨论训练细节。

我们将式（9）中的抗混杂 Q 值目标纳入离线强化学习目标，该目标基于 Park 等人提出的 Flow Q-Learning（FQL）。(2025b)。请注意，所提出的因果离线强化学习目标并不局限于 FQL，也可以纳入其他基于梯度的策略学习方法，以减轻混杂偏差的影响。例如，我们在第 G 节展示了一种采用价值流（Dong 等人，2025年）的替代实现。

与 FQL 类似，我们提出的增强方法称为因果FQL（CFQL），其重点是在训练期间学习两个策略：(1) 一个行为克隆（BC）策略 μ_ω，用于模仿名义行为策略分布 $P(X_t \mid S_t)$，该分布来自观测数据；(2) 一个目标策略 π_θ，用于决定学习者的动作。具体而言，BC 策略 $\mu_\omega(s,\ z)$ 是一种连续归一化流（Lipman 等人，2022），以状态 s 和独立噪声 z 作为输入。它会将

噪声 z 转换为从名义行为分布 $P(X_t \mid S_t)$ 中抽取的动作 x，这一过程遵循多步去噪过程。另一方面，目标策略 $\pi_\theta(s,z)$ 是一个单步生成模型，可将噪声 z 直接映射为动作域中的一个样本 x。

现在我们描述所提出算法因果FQL（算法1）的细节。CFQL包含四个主要组成部分：评论家网络的集成 Q_{ϕ_i}、行为克隆流策略 μ_ω、单步目标策略 π_θ，以及判别器 D_ψ，用于区分从行为克隆策略和目标策略中抽取的动作。更具体地说，步骤4-9训练评论家Q网络的集成 Q_{ϕ_i}，$i = 1, 2, \ldots\ldots, N$。每个网络 Q_{ϕ_i} 均基于混杂观测数据训练，遵循对Q集成平均值进行单步贝尔曼更新（步骤8）的标准最小化$\bar{Q}_\phi(s', x') = \frac{1}{N}\sum_i Q_{\phi_i}(s', x')$。

从步骤10至15，该算法训练速度场$v_\omega(t,s,x^t)$，即行为克隆流策略 μ_ω的速度场。它将独立噪声 x^0 到观测动作x^1 的变换建模为以恒定速度运动的自由粒子轨迹。速度场 $v_\omega(t,s,x^t)$是一个神经网络，记录轨迹在任意时间步 t处的梯度。训练遵循标准流匹配过程（Lipman 等，2022；2024年）。步骤16至20训练因果策略梯度式(9)中的判别器网络 D_ψ，以区分从行为克隆流 μ_ω 和目标策略 π_θ中抽取的样本。具体而言，该算法获取从 $\mu_\omega(s,z)$和 $\pi_\theta(s,z)$中抽取的样本，给定观测状态s 和独立噪声 z（步骤17-19）。行为克隆流策略的样本由已训练的速度场按照标准欧拉方法（算法2）计算得到。判别器 $\mathcal{D}_\psi$的训练目标随后由下式给出，

$$\mathcal{L}_{\text{Disc}}(\psi) = \mathbb{E}_{s\sim\mathcal{D}, z\sim\mathcal{N}(0,I_d)}\big[\log D_\psi(s, \mu_\omega(s,z)) + \log(1 - \mathcal{D}_\psi(s, \pi_\theta(s,z)))\big] \tag{10}$$

最小化上述判别器损失等价于解决一个二元分类问题：类别 1 表示动作 x 来自 BC 流策略 μ_ω，而类别 0 表示在给定当前状态 s 的情况下，动作是从单步目标策略 π_θ 中采样得到的。

最后，CFQL 训练一个单步目标策略 π_θ，在步骤 21–24 中优化与混杂观测相容的最坏情况回报。这里的关键挑战是，如何可靠地估计最坏情况价值函数 $\min_{s^*} Q_{\pi_\theta}(s^*, x^*)$，即公式（9）中的函数，同时避免过度悲观。为应对这一挑战，我们不在整个状态空间或整个数据批次上取最小值（Li 等，2025b），而是提出学习多个 Q 网络的集成，并在学得的集成中取最小值，以模拟最坏情况。

算法 1因果流 Q 学习（因果FQL）

1: **输入：** 离线数据集 $\mathcal{D}$。
2: **当**尚未收敛时 **执行**
3: 采样批次 $\{(s, x, y, s')\} \sim \mathcal{D}$
▽训练评论器 Q_{ϕ_i}
4: **对于** $i = 1, \ldots, N$ **执行**
5: $z \sim \mathcal{N}(0, I_d)$
6: $x' \leftarrow \pi_\theta(s', z)$,
7: 更新 ϕ_i 以最小化：
$\mathbb{E}\big[(Q_{\phi_i}(s,x) - y - \gamma\bar{Q}_\phi(s',x'))^2\big]$
8: **循环结束**
▽训练行为克隆流策略 μ_ω
9: $x^0 \sim \mathcal{N}(0, I_d)$
10: $x^1 \leftarrow x$
11: $t \sim$ 均匀分布(0,1)
12: $x^t \leftarrow (1-t)x^0 + x^1$
13: 更新 ω 以最小化 $\mathbb{E}\big[\|v_\omega(t,s,x^t) - (x^1 - x^0)\|^2\big]$
▽训练判别器 D_ψ
14: $z \sim \mathcal{N}(0, I_d)$
15: $x \leftarrow \mu_\omega(s,z)$
16: $x' \leftarrow \pi_\theta(s, z)$
17: 更新 ψ 以最小化：
$\mathbb{E}\big[$对数$(1 - \mathcal{D}_\psi(s, x))$ + 对数 $D_\psi(s, x')\big]$
▽训练单步目标策略 π_θ
18: $z \sim \mathcal{N}(0, I_d)$
19: $x \leftarrow \pi_\theta(s, z)$
20: 更新 θ 以最小化：
$\mathbb{E}\big[-\underline{Q_*}(s, x) + \alpha\|x - \mu_\omega(s, z)\|^2\big]$
21: **结束循环**
22: **返回**单步流策略 π_θ。

算法 2 行为克隆流策略 $\mu_\omega(s,z)$

输入： 速度场 $v_\omega(t, s, z)$，时间步 M。
对于 $t = 0, 1, \ldots, M-1$ **执行**3:
$z \leftarrow z + v_\omega(t/M, s, z)/M$4: **结束循环**5: **返回**采样动作 z

利用判别器 D_ψ，可以近似计算因果策略梯度中的价值函数评估，如下所示：

$$\underline{Q_*}(s,x) = D_\psi(s,x)\bar{Q}_\phi(s,x) + (1 - D_\psi(s,x)) \min_i Q_{\phi_i}(s,x) \tag{11}$$

通过优化以下目标函数来训练单步目标策略：

$$\mathcal{L}_\pi(\theta) = \mathbb{E}_{s\sim\mathcal{D}, x\sim\pi_\theta}\big[-\underline{Q_*}(s,x)\big] + \alpha\mathcal{L}_{\text{Distill}}(\theta) \tag{12}$$

其中，α 是用于控制目标策略正则项强度的超参数；$\mathcal{L}_{\text{Distill}}(\theta)$是蒸馏

表 *1.* **聚合离线评估结果**。我们报告了不同任务组的聚合平均成功率和平均偏差。Causal-FQL 在所有任务组上均持续优于所有基线方法。与原始 FQL 相比，混杂鲁棒的 Causal-FQL 平均实现了 1.2× 的成功率提升。

	高斯策略			流策略			
	IQL	ReBRAC	FBRAC	IQN	IFQL	FQL	Causal-FQL
视觉立方体单人游戏单任务（5 个任务）	64 ± 17	$\mathbf{82 \pm 7}$	64 ± 10	58 ± 7	44 ± 9	65 ± 12	$\mathbf{81 \pm 6}$
visual-cube-double-play-singletask（5 个任务）	$\mathbf{11 \pm 6}$	1 ± 1	2 ± 1	1 ± 0	2 ± 2	6 ± 1	$\mathbf{11 \pm 3}$
视觉场景游戏单任务（5 个任务）	26 ± 5	28 ± 5	11 ± 1	$\mathbf{41 \pm 6}$	21 ± 2	$\mathbf{41 \pm 4}$	$\mathbf{43 \pm 1}$
visual-puzzle-3x3-play-singletask（5 个任务）	2 ± 3	20 ± 1	1 ± 0	19 ± 1	21 ± 0	20 ± 1	$\mathbf{26 \pm 2}$
visual-puzzle-4x4-play-singletask（5 个任务）	0 ± 0	5 ± 1	0 ± 0	$\mathbf{13 \pm 3}$	4 ± 7	11 ± 3	$\mathbf{13 \pm 2}$
归一化均值	0.72	0.95	0.55	0.92	0.65	1.0	**1.21**

损失，其定义遵循（Park 等人，2025b），即：

$$\mathcal{L}_{\text{Distill}}(\omega) = \mathbb{E}_{\substack{s\sim\mathcal{D},\\ z\sim\mathcal{N}(0,I_d)}} \left[\|\pi_\theta(s,z) - \mu_\omega(s,z)\|_2^2\right] \quad (13)$$

所提出的 CFQL 与原始 FQL 类似，只是额外加入了判别器训练步骤和正则化，以改进最坏情况价值函数估计。这意味着可以使用标准自动微分框架（Baydin 等人，2018年）一致地执行基于梯度的策略优化。

4. 实验

本节中，我们将在 25 项 OGBench（Park 等人，2025a）视觉任务上评估新提出的因果流 Q 学习。我们旨在回答以下问题：

Q1：在基于像素的任务上，因果FQL 如何优于原始 FQL？**Q2：**在离线到在线的基于像素任务中，因果 FQL 的样本效率如何？**Q3：**哪些超参数对因果FQL 至关重要？

实验设置。 我们在近期提出的 OGBench（Park 等人，2025a）中的离线强化学习任务上评估我们的方法。由于我们考虑解决基于像素的离线强化学习任务中的混杂偏差（见第4节），以往仅支持结构化状态观测的基准测试（如 D4RL（Fu 等人，2020年）不在我们的考虑范围内。OGBench 最初被提出作为一个目标条件强化学习基准，因此我们使用其-单任务 变体，以兼容标准离线强化学习框架。更具体地说，我们选择操作任务，因为其多模态特性适合展示流策略。对于所有任务，我们向智能体提供 $64 \times 64 \times 3$ RGB 图像观测（例如图1b），并采用帧堆叠和随机裁剪作为图像增强。为确保公平比较，我们为所有方法使用规模相近的价值网络/策略网络/演员网络，并在所有任务中使用 IMPALA（Espeholt 等人，2018年）的图像编码器。完整的设置详情请参见第F节。



图 *4*。**基于像素的 OGBench 任务**。

基线方法。 我们使用以下代表不同策略提取策略的算法，作为离线强化学习任务的基线。由于视觉任务具有较高的计算需求，我们主要采用那些在以往工作中（Dong 等人，2025年；Park 等人，2025b）表现出稳定性能的基线算法。我们选择 IQL（Kostrikov 等人，2021）和 ReBRAC（Tarasov 等人，2023）作为使用高斯策略学习标量 Q 值这一类算法的代表。我们选择 IQN（Dab- ney 等人，2018）作为分布式 Q 值与流策略基线，并选择 FBRAC、IFQL 和 FQL（Park 等人，2025b）作为标量 Q 值评论器与流策略基线。FBRAC 和 IFQL 分别是将流策略应用于 Diffusion-QL（DQL）（Wang 等人，2023）和 IDQL（Hansen-Estruch 等人，2023）原始算法的变体。

对于离线到在线实验，我们主要采用 ReBRAC 和 FQL 作为基线方法，因为它们在离线实验中的表现相较于其他基线方法更强。

评估。 对于离线实验，我们对所有基线方法和 `Causal-FQL` 提出的方法均训练 500K 步，对 4 个随机种子的结果取平均，并报告均值和标准差。表格中，我们突出显示每项任务上达到最佳性能 95% 以内的结果，遵循（Park 等人，2025b）的做法。对于离线到在线实验，我们在离线阶段采用相同的设置，并将训练扩展至在线阶段，额外进行 500K 步。实现和实验细节见第E 和 F节。

4.1. 离线性能

从表 1 的汇总结果，Causal-FQL （CFQL）在所有任务组中都取得了最佳或接近最佳的表现，超过不了解因果关系的原始 FQL

表 2. **完整的离线评估结果**。 Causal-FQL 在所有定义域上均优于原始FQL，并在 25 个定义域中的 19 个上取得最佳或接近最佳的结果。由于计算需求较高，我们报告了 4 个随机种子的平均结果，并将各定义域中达到最佳性能 95% 以内的数值加粗；有可用的既有研究结果时则予以采用（Park 等人，2025b；Dong 等人，2025年）。

	高斯策略		流策略				
	IQL	ReBRAC	FBRAC	IQN	IFQL	FQL	Causal-FQL
visual-cube-single-play-singletask-task1-v0 (*) 70 ± 12		**83 ± 6**	55 ± 8	5 ± 4	49 ± 7	81 ± 12	**85 ± 6**
visual-cube-single-play-singletask-task2-v0	59 ± 13	**85 ± 6**	70 ± 11	58 ± 9	45 ± 11	58 ± 12	**81 ± 5**
visual-cube-single-play-singletask-task3-v0	64 ± 35	**88 ± 4**	81 ± 4	83 ± 11	67 ± 6	73 ± 16	**84 ± 5**
视觉立方体单人游戏单任务任务4-v0	68 ± 12	**78 ± 9**	56 ± 19	**78 ± 3**	34 ± 13	63 ± 8	**74 ± 5**
视觉立方体单人游戏单任务任务5-v0	59 ± 14	**76 ± 9**	59 ± 10	64 ± 9	27 ± 10	49 ± 12	**79 ± 7**
视觉双立方体单任务任务1-v0 (*) 34 ± 23		4 ± 4	6 ± 2	4 ± 1	8 ± 6	23 ± 4	**43 ± 11**
视觉立方体双人游戏单任务任务2-v0	**2 ± 1**	0 ± 0	2 ± 2	0 ± 0	0 ± 0	0 ± 0	**2 ± 1**
视觉立方体双人游戏单任务任务3-v0	**7 ± 4**	2 ± 2	2 ± 1	0 ± 0	1 ± 1	4 ± 2	4 ± 2
视觉立方体双人游戏单任务任务4-v0	**1 ± 1**	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0	**1 ± 1**
视觉立方体双人游戏单任务任务5-v0	**11 ± 2**	0 ± 0	0 ± 0	1 ± 1	2 ± 1	4 ± 1	4 ± 2
视觉场景游戏单任务任务1-v0 (*)	**97 ± 2**	**98 ± 4**	46 ± 4	**95 ± 2**	86 ± 10	**98 ± 3**	**100 ± 0**
视觉场景游戏单任务任务2-v0	21 ± 16	30 ± 15	0 ± 0	79 ± 15	0 ± 0	86 ± 8	**92 ± 3**
视觉场景游戏单任务任务3-v0	12 ± 9	10 ± 7	10 ± 3	**31 ± 14**	19 ± 2	22 ± 6	22 ± 1
视觉场景游戏单任务任务4-v0	1 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0	1 ± 1	**2 ± 1**
视觉场景游戏单任务任务5-v0	**0 ± 0**	**0 ± 0**	**0 ± 0**	**0 ± 0**	**0 ± 0**	**0 ± 0**	**0 ± 0**
视觉3×3拼图游戏单任务任务1-v0 (*)	7 ± 15	88 ± 4	7 ± 2	84 ± 1	**100 ± 0**	94 ± 1	**98 ± 1**
3×3视觉拼图单任务任务2-v0	0 ± 0	**12 ± 1**	0 ± 0	6 ± 2	0 ± 0	0 ± 0	1 ± 0
3×3视觉拼图单任务任务3-v0	0 ± 0	1 ± 1	0 ± 0	1 ± 0	2 ± 1	0 ± 0	**3 ± 2**
3×3视觉拼图单任务任务4-v0	1 ± 1	0 ± 1	0 ± 0	3 ± 0	1 ± 0	5 ± 4	**8 ± 3**
3×3视觉拼图单任务任务5-v0	0 ± 0	0 ± 0	0 ± 0	1 ± 0	0 ± 0	1 ± 2	**19 ± 4**
4×4视觉拼图单任务任务1-v0 (*)	0 ± 0	26 ± 6	0 ± 0	21 ± 3	8 ± 15	33 ± 6	**37 ± 4**
4×4视觉拼图单任务任务2-v0	0 ± 0	0 ± 0	1 ± 1	**14 ± 2**	1 ± 1	1 ± 2	4 ± 1
4×4视觉拼图单任务任务3-v0	0 ± 0	0 ± 0	0 ± 0	9 ± 3	9 ± 15	16 ± 6	**18 ± 4**
4×4视觉拼图单任务任务4-v0	0 ± 0	1 ± 1	0 ± 0	**12 ± 4**	2 ± 2	2 ± 1	**6 ± 2**
4×4视觉拼图单任务任务5-v0	0 ± 0	0 ± 0	1 ± 1	**11 ± 4**	0 ± 0	1 ± 1	1 ± 1

20%。值得注意的是，在单立方体 任务组中，CFQL 的表现可与先前最先进的高斯基线方法（ReBRAC）相媲美，而原始 FQL 则远远落后。第 4 节展示了各任务组中每项任务的完整离线结果。CFQL 在 25 项任务中的 19 项上取得了最佳或接近最佳的表现。尤其是在 visual-cube-double-play-singletask-task1 中，CFQL 的表现几乎是原始 FQL 的两倍；而在 visual-puzzle-3x3-play-singletask-task5 中，只有我们的方法取得了明显高于零的成功率。我们发现，CFQL 的表现至少不逊于 FQL，这在实证上验证了抗混杂目标函数是最优 Q 值的一个有信息量的下界。此外，也有一些任务中 FQL 和 CFQL 的表现都不佳。我们推测，这些任务面临的最大挑战可能并非混杂偏差，而是离线强化学习的其他方面，例如表征学习，或仅仅是由于离线数据的状态-动作空间覆盖有限。

4.2. 离线到在线性能

我们还研究了Causal-FQL的离线到在线微调性能。总体而言，Causal-FQL相较原始FQL取得了出色的微调性能。它平均比先前的最先进水平FQL的提升了8%微调性能，表明Causal-FQL在离线任务和离线到在线任务中均具有广泛适用性。在4×4视觉拼图上，Causal-FQL也是文献中首个完全解决任务1的算法。图5 展示了训练期间的完整评估曲线。

Causal-FQL 与以往离线强化学习工作的一个显著区别在于，我们发现有必要在在线阶段使用不同的目标函数。回想一下，图 1 中，离线数据集中存在混杂偏差，因为数据收集策略（专家）在结构化状态观测空间中运行，而学习者的策略则在像素观测空间中运行。专家采取的动作依赖于像素空间中可能无法完全观测到的信息，从而产生混杂偏差。然而，在在线微调时，数据收集策略就是学习者自身的策略，不存在观测空间不匹配。因此，原则上，我们应在在线微调期间切换回原始 FQL 目标函数。

然而，在实现中，标准做法是用在线经验逐步替换回放缓冲区中的离线轨迹，从而使训练不会因突然的分布偏移而受到影响。因此，在线阶段提供给智能体的实际数据是一种混合

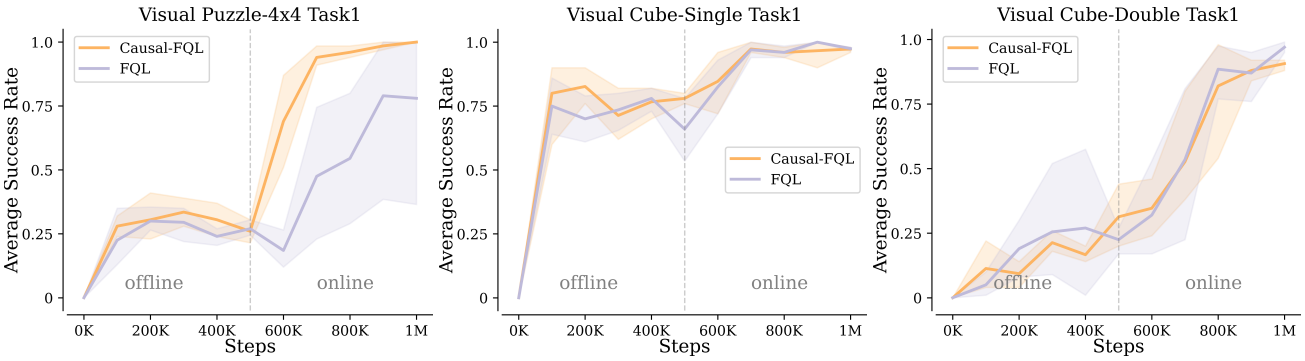


图*5*。 **离线到在线评估**。 Causal-FQL 在从离线评估中未能接近最优解决的任务类型中选出的三个代表性任务上，收敛至接近最优的成功率。所有结果均取4个随机种子的平均值。

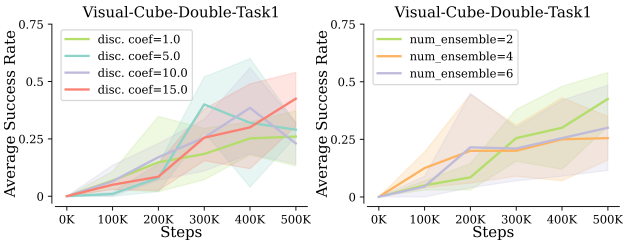


图*6*。 **经过适当调优的判别器损失系数和评论器集成数量能够提升性能**。

有混杂数据与无混杂数据。目标函数的选择需针对具体任务进行调优。在visual-cube-single和visual-cube-double中，我们发现在在线阶段完全切换回FQL目标函数效果更好；而在4×4视觉拼图中，我们发现从在线和离线回放缓冲区进行均衡采样，并对批次的每一半使用相应的Causal-FQL/FQL目标函数效果更好。我们相信，对Causal-FQL在线微调期间所使用的目标函数进行更精细的控制，可以进一步提升性能和样本效率。

4.3. 超参数调优

本节研究如何调优混杂鲁棒 Q 值目标（公式（11））的两个基本组成部分：判别器和 Q 值集成。鉴于任务难度适中，我们主要在`visual-cube-double`的任务1上调优超参数。对于每个任务组，我们进一步调优任务1的初始判别器损失系数。下面介绍主要发现。详细超参数见第F节。

判别器损失系数。我们对判别器损失系数进行扫描，范围为 $\{1, 5, 10, 15\}$。第4.3 节展示了不同系数下的完整成功率曲线。我们注意到，系数过小时，该方法与原始 FQL 接近，并且通常会在离线训练阶段后期导致收敛不稳定。在其他环境中也观察到了类似模式。

Q网络集成的数量。 我们使用数值 $\{2, 4, 6\}$ 改变 Q 值集成的数量，同时将所有其他参数设为先前调优中确定的最优值。我们的研究结果表明，总体而言，较少的集成能够取得更好的结果。这可能是因为从数量更多的集成中取最小值的倾向会导致结果过度悲观。

5. 结论

我们研究了基于像素的离线强化学习任务中隐式混杂偏差的存在，并提出 Causal-FQL，这是一种新颖的基于流的离线智能体，能够抵御此类混杂因素。从理论上讲，我们为因果策略梯度建立了一种新颖的目标函数，以优化目标策略在最坏情况环境中的表现。在实践中，我们提出了一种简单而高性能的抗混杂离线强化学习目标函数近似方法，该方法利用判别器和评论家网络的深度集成。大量实验表明，在离线任务以及从离线过渡到在线设置的高难度视觉控制任务中，Causal-FQL 的表现均优于以往最先进的方法。Causal-FQL 是首个在 OGBench 视觉基准测试中完全解决某些任务的方法。最后，我们强调在强化学习问题中进行有原则的因果分析至关重要；尽管以往离线强化学习文献中的观测空间不匹配看似无关紧要，我们证明它们会对视觉任务造成重大障碍。未来的工作将着重探索因果策略梯度的其他实现方式，包括采用不同的评论家网络和表达能力更强的策略类。

影响声明

本文探讨了这样一种鲁棒策略学习的理论与算法框架：学习者利用表达能力强的策略类，从通过被动观察专家示范者而获得的混杂离线数据中进行学习。当专家用于确定动作价值的输入变量未知，且与学习者使用的观测状态不同时，就会产生隐式混杂偏差。示例1展示了一种情形：专家能够访问底层的结构化状态，而学习者则试图从专家自然轨迹的视频记录中进行学习。

我们的框架可应用于多个领域，包括自动驾驶汽车开发、工业机器人技术和慢性病管理等。对于安全关键领域而言，我们的方法尤其能够降低从涉及未观测混杂因素的示范数据中进行离线强化学习（RL）训练所带来的潜在风险。随着黑箱人工智能系统日益普及，而我们对其长期社会影响的理解仍然有限，我们认为，构建可信赖的人工智能决策系统正变得愈发重要。

参考文献

Agrawalla, B. K., Nauman, M., Agrawal, K., and Kumar, A. floq：通过流匹配训练评论器，以扩展基于价值的强化学习的计算规模。载于 *NeurIPS 2025* 研讨会：第二届强化学习实验研究者与理论研究者对齐研讨会，2025年。网址：https://openreview.net/forum?id=WwQoSHGCXg。

Balke, A. 和 Pearl, J. 不完全依从研究中的处理效应界限。《美国统计学会杂志》，92(439)：1172–1176，1997年9月。

Ball, P. J., Smith, L. M., Kostrikov, I. 和 Levine, S. 利用离线数据进行高效在线强化学习。载于 Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S. 和 Scarlett, J.（编），国际机器学习会议（*ICML 2023*），*2023*年*7*月*23*日至*29*日，美国夏威夷檀香山，第 202 卷，机器学习研究论文集，第 1577–1594 页。PMLR，2023年。网址：https://proceedings.mlr.press/v202/ball23a.html。

Bareinboim, E., Correa, J. D., Ibeling, D. 和 Icard, T.论 *Pearl* 因果层级与因果推断的基础，第 507–556 页。美国计算机协会，美国纽约州纽约市，第 1 版，2022。国际标准书号 9781450395861。网址：https://doi.org/10.1145/3501714.3501743。

Bareinboim, E., Zhang, J. 和 Lee, S. 因果强化学习导论。技术报告 R-65，因果人工智能实验室，哥伦比亚大学，2024年12月。Baydin, A. G., Pearlmutter, B. A., Radul, A. A. 和 Siskind, J. M. 机器学习中的自动微分：综述。《机器学习研究期刊》，18(1):1– 43，2018年。网址：http://jmlr.org/papers/v18/17-468.html。Bellman, R. 动态规划。《科学》，153 (3731):34–37，1966年。doi: 10.1126/science.153.3731. 34。网址：https://www.science.org/doi/abs/10.1126/science.153.3731.34。

Bennett, A. 和 Kallus, N. 近端强化学习：部分可观测马尔可夫决策过程中的高效离策略评估。运筹学研究，72(3):1071–1086，2024年。doi: 10.1287/OPRE.2021.0781。URLhttps://doi.org/10.1287/opre.2021.0781。

Bennett, A.、Kallus, N.、Li, L. 和 Mousavi, A. 存在潜在混杂因素的无限时域强化学习中的离策略评估。载于人工智能与统计国际会议，第 1999–2007 页。PMLR，2021年。

Black, K.、Brown, N.、Driess, D.、Esmail, A.、Equi, M.、Finn, C.、Fusai, N.、Groom, L.、Hausman, K.、Ichter, B.、Jakubczak, S.、Jones, T.、Ke, L.、Levine, S.、Li-Bell, A.、Mothukuri, M.、Nair, S.、Pertsch, K.、Shi, L. X.、Tanner, J.、Vuong, Q.、Walling, A.、Wang, H. 和 Zhilinsky, U. π_0：用于通用机器人控制的视觉-语言-动作流模型，2026年。URLhttps://arxiv.org/abs/2410.24164。

Chen, L., Lu, K., Rajeswaran, A., Lee, K., Grover, A., Laskin, M., Abbeel, P., Srinivas, A. 和 Mordatch, I. 决策 Transformer：通过序列建模实现强化学习。载于 Ranzato, M., Beygelzimer, A., Dauphin, Y. N., Liang, P. 和 Vaughan, J. W.（编），《神经信息处理系统进展》第*34*卷：神经信息处理系统大会 *2021* 年会，*NeurIPS 2021*，*2021* 年 *12* 月 *6—14* 日，线上举行，第 15084–15097 页，2021 年。网址：https://proceedings.neurips.cc/paper/2021/hash/7f489f642a0ddb10272b5c31057f0663-Abstract.html。

Correa, J. 与 Bareinboim, E. 软干预的一般可迁移性：完备性结果。载于 Larochelle, H.、Ranzato, M.、Hadsell, R.、Balcan, M. 与 Lin, H.（编），《神经信息处理系统进展》，第33卷，第10902–10912页。Curran Associates, Inc.，2020年。网址 https://proceedings.neurips.cc/paper_files/paper/2020/file/

7b497aa1b2a83ec63d1777a88676b0c2-论文 pdf.

Dabney, W., Ostrovski, G., Silver, D., and Munos, R. 用于分布式强化学习的隐式分位数网络。载于 Dy, J. 和 Krause, A.（编），第 *35* 届国际机器学习会议论文集，第 80 卷，机器学习研究论文集，第 1096–1105 页。PMLR，2018年7月10日至15日。网址：https://proceedings.mlr.press/v80/dabney18a.html。

Deng, S., Yan, M., Wei, S., Ma, H., Yang, Y., Chen, J., Z hang, Z., Yang, T., Zhang, X., Cui, H., Zhang, Z., and Wang, H. GraspVLA：一种在十亿规模合成动作数据上预训练的抓取基础模型。载于 第 *9* 届机器人学习年会，2025年。网址：https://openreview.net/forum?id=zEC8TOXDkH。

Dong, P.、Zheng, C.、Finn, C.、Sadigh, D. 和 Eysenbach, B. 价值流，2025年。网址 https://arxiv.org/abs/2510.07650。

Espeholt, L.、Soyer, H.、Munos, R.、Simonyan, K.、Mnih, V.、Ward, T.、Doron, Y.、Firoiu, V.、Harley, T.、Dunning, I.、Legg, S. 和 Kavukcuoglu, K. IMPALA：采用重要性加权的 actor-学习者架构实现可扩展的分布式深度强化学习。在 Dy, J. G. 和 Krause, A.（编）编，第 *35* 届国际机器学习会议，*ICML 2018*，*Stockholmsm¨assan*，瑞典斯德哥尔摩，*2018*年7月*10*日至*15*日，机器学习研究论文集第 *80* 卷，第 1406–1415 页。PMLR，2018年。网址 http://proceedings.mlr.press/v80/espeholt18a.html。

Fu, J.、Kumar, A.、Nachum, O.、Tucker, G. 和 Levine, S. D4RL：深度数据驱动强化学习数据集。arXiv，2020年。

Fujimoto, S. 和 Gu, S. 离线强化学习的极简方法。载于 Beygelzimer, A.、Dauphin, Y.、Liang, P. 和 Vaughan, J. W.（编），神经信息处理系统进展，2021年。网址：https://openreview.net/forum?id=Q32U7dzWXpc。

Garg, D.、Hejna, J.、Geist, M. 和 Ermon, S. 极限 Q 学习：无熵的最大熵强化学习，2023年。网址：https://arxiv.org/abs/2301.02328。

Ghugare, R. 和 Eysenbach, B. 归一化流是适用于强化学习的强大模型，2025年。网址：https://arxiv.org/abs/2505.23527。

Guo, H., Cai, Q., Zhang, Y., Yang, Z. 和 Wang, Z. 针对部分可观测马尔可夫决策过程的可证明高效离线强化学习。载于 Chaudhuri, K., Jegelka, S., Song, L., Szepesv´ari, C., Niu, G., 和Sabato, S.（编），国际机器学习大会，*ICML 2022*，*2022*年7月*17*日至*23*日，美国马里兰州巴尔的摩，第162卷，机器学习研究论文集，第8016–8038页。PMLR，2022。网址：https://proceedings.mlr.press/v162/guo22a.html。

Hansen-Estruch, P., Kostrikov, I., Janner, M., Kuba, J. G., and Levine, S. IDQL：将隐式 Q 学习作为一种采用扩散策略的演员-评论器方法，2023年。URL https://arxiv.org/abs/2304.10573。

Hess, K., Frauen, D., Melnychuk, V., and Feuerriegel, S. 在未观测混杂因素下高效且精确的离策略学习，2025年。URLhttps://arxiv.org/abs/2502.13022。

Imbens, G. W. and Angrist, J. D. 局部平均处理效应的识别与估计。《计量经济学》，62 (2):467–475, 1994.

Intelligence, P., Black, K., Brown, N., Darpinian, J.,Dha-balia, K., Driess, D., Esmail, A., Equi, M., Finn,C.,Fusai, N.,Galliker, M.Y., Ghosh,D., Groom, L., Haus-man,K., Ichter, B., Jakubczak, S.,Jones,T., Ke,L.,LeBlanc, D.,Levine, S.,Li-Bell, A., Mothukuri,M.,Nair, S., Pertsch, K., Ren, A.Z., Shi, L.X., Smith, L.,Springenberg, J.T., Stachowicz, K., Tanner, J., Vuong,Q., Walke, H., Walling,A., Wang,H., Yu, L., andZhilinsky, U. $\pi_{0.5}$：一种具备开放世界泛化能力的视觉-语言-动作模型，2025年。URLhttps://arxiv.org/abs/2504.16054。Janner, M., Li, Q., and Levine, S. 将离线强化学习视为单一大型序列建模问题。载于 Ranzato, M., Beygelzimer, A., Dauphin, Y. N., Liang, P., and Vaughan, J. W.（编），神经信息处理系统进展*34*：神经信息处理系统大会年会，*NeurIPS 2021*，*2021*年*12*月*6*日至*14*日，线上举行，第1273–1286页，2021年。URL https://proceedings.neurips.cc/paper/2021/hash/099fe6b0b444c23836c4a5d07346082b-Abstract.

超文本标记语言。

Jiang, Y., Cheng, S., Ding, Y., Gao, F., 和 Qi, B. AsyncVLA：面向视觉-语言-动作模型的异步流匹配，2025年。网址https://arxiv.org/abs/2511.14148。

Joshi, S.、Zhang, J. 和 Bareinboim, E.《部分可识别性下的安全策略学习：一种因果方法》。载于第三十八届 *AAAI* 人工智能大会、第三十六届人工智能创新应用大会和第十四届人工智能教育进展研讨会论文集，AAAI'24/IAAI'24/EAAI'24。AAAI 出版社，

2024年a。国际标准书号 978-1-57735-887-9。doi: 10.1609/aaai. v38i12.29198。网址 https://doi.org/10.1609/aaai.v38i12.29198。

Joshi, S.、Zhang, J. 和 Bareinboim, E.《迈向部分可识别条件下的安全策略学习：一种因果方法》。载于 Wooldridge, M. J.、Dy, J. G. 和 Natarajan, S.（编），第三十八届 *AAAI* 人工智能大会（*AAAI 2024*）、第三十六届创新人工智能应用大会（*IAAI 2024*）、第十四届教育领域人工智能进展研讨会（*EAAI 2014*），*2024*年*2*月*20*日至*27*日，加拿大温哥华，第 13004–13012 页。AAAI 出版社，2024b。doi：10.1609/AAAI.V38I12.29198。网址：https://doi.org/10.1609/aaai.v38i12.29198。

Kallus, N. 和 Zhou, A.《抗混杂的策略改进》。载于 Bengio, S.、Wallach, H. M.、Larochelle, H.、Grauman, K.、Cesa-Bianchi, N. 和 Garnett, R.（编），神经信息处理系统进展 *31*：*2018*年神经信息处理系统年度会议，*NeurIPS 2018*，*2018*年*12*月*3*日至*8*日，蒙特·利尔，加拿大，第 9289–9299 页，2018年。网址：
h
t
t
p
s://
proceedings.neurips.cc/paper/2018/hash/3a09a524440d44d7f19870070a5ad42f-Abstrac
html.

Kallus, N. 和 Zhou, A. 无限时域强化学习中的抗混杂策略评估。神经信息处理系统进展，33：22293–22304，2020年。

Kostrikov, I.、Nair, A. 和 Levine, S. 使用隐式 Q 学习进行离线强化学习，2021年。

Kumar, A.、Zhou, A.、Tucker, G. 和 Levine, S. 离线强化学习的保守 Q 学习。载于 Larochelle, H.、Ranzato, M.、Hadsell, R.、Balcan, M. 和 Lin, H.（编），神经信息处理系统进展 *33*：神经信息处理系统大会 *2020* 年会，*NeurIPS 2020*，*2020*年*12*月*6*日至*12*日，线上，2020年。网址：
h
t
t
p
s://
proceedings.neurips.cc/paper/2020/hash/0d2b2061826a5df3221116a5085a6052-Abstrac
html. Lange, S., Gabel, T., 和 Riedmiller, M. 批量强化学习。载于强化学习：现状综述，第 45–73 页。施普林格，2012年。Levine, S., Kumar, A., Tucker, G., 和 Fu, J. 离线强化学习：教程、综述及开放问题展望。*arXiv* 预印本 *arXiv:2005.01643*，2020年。

Li, M., Zhang, J., 和 Bareinboim, E. 因果对齐课程学习。载于第十二届国际学习表征会议，2024年。网址：https://openreview.net/forum?id=hp4yOjhwTs。Li, M., Zhang, J., 和 Bareinboim, E. 基于混杂离线数据的自动奖励塑形。载于第四十二届国际机器学习会议，2025a。网址：https://openreview.net/forum?id=Hu7hUjEMiW。

Li, M., Zhang, J. 和 Bareinboim, E. 混杂因素下的鲁棒深度强化学习：一种因果方法。载于 第三十九届神经信息处理系统年会，2025b。网址：https://openreview.net/forum?id=9fUr5iFU9j。

Lillemark, H. J., Huang, B., Zhan, F., Du, Y. 和 Keller, T. A. 流等变世界模型：部分可观测动态环境中的记忆，2026。网址：https://arxiv.org/abs/2601.01075。

Lipman, Y., Chen, R. T., Ben-Hamu, H., Nickel, M. 和 Le, M. 用于生成建模的流匹配。*arXiv* 预印本 *arXiv:2210.02747*，2022。

Lipman, Y., Havasi, M., Holderrieth, P., Shaul, N., Le, M., Karrer, B., Chen, R. T., Lopez-Paz, D., Ben-Hamu, H. 和 Gat, I. 流匹配指南与代码。*arXiv* 预印本 *arXiv:2412.06264*，2024年。

Liu, T., Zhao, S., 和 Rhinehart, N. 迈向基础激光雷达世界模型：高效的潜在流匹配。发表于 第三十九届神经信息处理系统大会，2025年。网址 https://openreview.net/forum?id=OyX9cC9WaV。

Liu, X., Gong, C., 和 Liu, Q. 流动既直且快：学习使用整流流生成和迁移数据。*arXiv* 预印本 *arXiv:2209.03003*，2022。

Lv, L., Li, Y., Luo, Y., Sun, F., Kong, T., Xu, J., 和 Ma, X. 面向在线强化学习的基于流的策略。发表于 第三十九届神经信息处理系统大会，2025年。网址 https://openreview.net/forum?id=CANUXhPoyn。

Mandlekar,A., Xu,D., Wong, J.,Nasiriany,S., Wang, C.,Kulkarni, R., Fei-Fei, L., Savarese, S., Zhu,Y., andMart´ın-Mart´ın, R. 机器人操作中，离线人类示范数据学习有哪些关键因素？载于机器人学习会议，第 1678–1690 页。PMLR，2022。Manski, C. F. 治疗效应的非参数界限。《美国经济评论》，80：319–323，1989。网址：
h
t
tps://api.semanticscholar.org/CorpusID:153656329。

Miao, R., Qi, Z., and Zhang, X. 非参数模型下回合式部分可观测马尔可夫决策过程的离策略评估。载于 Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A.（编），神经信息处理系统进展第 *35* 卷：*2022* 年神经信息处理系统大会年会，*NeurIPS 2022*，美国路易斯安那州新奥尔良，*2022* 年 *11* 月 *28* 日至 *12* 月 *9* 日，2022。网址：http://papers.nips.cc/paper_files/paper/2022/hash/03dfa2a7755635f756b160e9f4c6b789-Abstract.html。

Nair, A., Gupta, A., Dalal, M. 和 Levine, S. Awac：利用离线数据集加速在线强化学习，2021年。网址 https://arxiv.org/abs/2006.09359。

Nakamoto, M., Zhai, S., Singh, A., Mark, M. S., Ma, Y., Finn, C., Kumar, A. 和 Levine, S. Cal-ql：用于高效在线微调的校准离线强化学习预训练。载于 Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M. 和 Levine, S.（编），神经信息处理系统进展 *36*：*2023* 年神经信息处理系统大会年会，*NeurIPS 2023*，美国新奥尔良，*2023* 年 *12* 月 *10* 日至 *16* 日，2023年。网址 http://papers.nips.cc/paper_files/paper/2023/hash/c44a04289beaf0a7d968a94066a1d696-Abstract.html。

Namkoong, H., Keramati, R., Yadlowsky, S. 和 Brunskill, E. 未观测混杂下序贯决策的离策略策略评估。神经信息处理系统进展，33:18819–18831，2020年。Nandiraju, G., Ju, Y., Xu, C. 和 Wang, H. HDFlow：面向长时域机器人装配的分层扩散流规划。载于*NeurIPS 2025* 具身世界模型决策研讨会，2025年。网址：https://openreview.net/forum?id=NDGP58PuQX。Narvekar, S., Peng, B., Leonetti, M., Sinapov, J., Taylor, M. E. 和 Stone, P. 强化学习领域的课程学习：框架与综述。《机器学习研究期刊》，21:181:1–181:50，2020年。网址：http://jmlr.org/papers/v21/20-212.html。Nguyen, K., Le, A. T., Pham, T., Huber, M., Peters, J. 和 Vu, M. N. Flowmp：通过条件流匹配学习用于机器人规划的运动场，2025年。网址：https://arxiv.org/abs/2503.06135。Park, S., Frans, K., Eysenbach, B. 和 Levine, S. OGBench：离线目标条件强化学习基准测试。载于第十三届国际学习表征会议论文集，2025a年。网址：https://openreview.net/forum?id=M992mjgKzI。

Park, S.、Li, Q. 和 Levine, S. Flow Q-Learning。在第四十二届国际机器学习会议，2025b。网址：https://openreview.net/forum?id=KVf2SFL1pi。

Pearl, J. 因果关系：模型、推理与推断。剑桥大学出版社，第 2 版，2009 年。doi：10.1017/CBO9780511803161。

Pearl, J. 和 Robins, J. 基于含隐变量因果模型的序贯计划概率评估。在第十一届人工智能不确定性会议论文集，UAI'95，第 444–453 页，美国加利福尼亚州旧金山，1995 年。Morgan Kaufmann 出版公司。国际标准书号 1558603859。

Precup, D.、Sutton, R. S. 和 Singh, S. P. 用于离策略策略评估的资格迹。载于 第十七届国际机器学习会议，第 759–766 页，2000年。Robbins, H. 关于实验序贯设计的若干方面。载于*Herbert Robbins* 论文选集，第 169–177 页。施普林格，1985年。Rohbeck, M.、Bunne, C.、Brouwer, E. D.、Huetter, J.- C.、Biton, A.、Chen, K. Y.、Regev, A. 和 Lopez, R. 运用流匹配对跨时间和条件的复杂系统动力学进行建模。载于 第十三届国际学习表征会议，2025年。网址：https://openreview.net/forum?id=hwHoBmOTrV。Shi, C.、Uehara, M.、Huang, J. 和 Jiang, N. 一种针对存在混杂因素的部分可观测马尔可夫决策过程进行离策略评估的极小极大学习方法。载于 Chaudhuri, K.、Jegelka, S.、Song, L.、Szepesv´ari, C.、Niu, G. 和 Sabato, S.（编），国际机器学习会议，*ICML 2022*，*2022*年*7*月*17*日至*23*日，美国马里兰州巴尔的摩，第 162 卷，收录于 机器学习研究论文集，第 20057–20094 页。PMLR，2022。网址：https://proceedings.mlr.press/v162/shi22f.html。Song, Y.、Zhou, Y.、Sekhari, A.、Bagnell, D.、Krishnamurthy, A. 和 Sun, W. 混合强化学习：同时使用离线和在线数据可提高强化学习的效率。载于 第十一届国际学习表征会议，*ICLR 2023*，卢旺达基加利，*2023*年*5*月*1*日至*5*日。OpenReview.net，2023年。网址：https://openreview.net/forum?id=yyBis80iUuU。Sutton, R. S. 强化学习：导论。*ABradford Book*，2018年。

Tarasov, D.、Kurenkov, V.、Nikulin, A. 和 Kolesnikov, S. 重新审视离线强化学习的极简方法。载于第三十七届神经信息处理系统会议，2023年。网址：https://openreview.net/forum?id=vqGWslLeEw。

Tchetgen,E. J.T.,Ying,A., Cui, Y.,Shi, X.,and Miao, W. 近端因果学习导论，2020年。

Tian, J. 与 Pearl, J. 因果概率：界限与识别。数学与人工智能年刊，28(1-4):287–313，2000年。doi: 10.1023/A:1018912507879。URLhttps://doi.org/10.1023/A:1018912507879。

Tiofack, F. N., Hellard, T. L., Schramm, F., Perrin-Gilbert, N. 和 Carpentier, J. 引导流策略：在离线强化学习中从高价值动作中学习，2025年。网址 https://arxiv.org/abs/2512.03973。Tong, A., FATRAS, K., Malkin, N., Huguet, G., Zhang, Y., Rector-Brooks, J., Wolf, G. 和 Bengio, Y. 通过小批量最优传输改进并泛化基于流的生成模型。 机器学习研究汇刊，2024年。ISSN 2835-8856。网址 https://openreview.net/forum?id=CD9Snc73AW。专家认证。Wang, Z., Hunt, J. J. 和 Zhou, M. 将扩散策略作为一种适用于离线强化学习的表达能力强的策略类。在 第十一届学习表征国际会议，2023年。网址 https://openreview.net/forum?id=AHvFDPi-FA。Wu, Y., Tucker, G. 和 Nachum, O. 行为正则化离线强化学习，2019年。网址 https://arxiv.org/abs/1911.11361。Yang, H., Yu, C., Sun, P. 和 Chen, S. 基于不完美示范数据的混合策略优化。在 第*37*届神经信息处理系统国际会议论文集，NIPS '23，美国纽约州雷德胡克，2023年。Curran Associates Inc. Zhang, J. 和 Bareinboim, E. 动态治疗方案中的近最优强化学习。在 神经信息处理系统进展，第13401–13411页，2019年。Zhang, J. 和 Bareinboim, E. 人类能否置身事外？在 因果学习与推理会议，第1010–1025页。PMLR，2022年。Zhang, J. 和 Bareinboim, E. 用于稳健离策略评估中处理混杂因素的资格迹。技术报告 R-105，哥伦比亚大学因果人工智能实验室，2024年5月。

Zhang, J. 和 Bareinboim, E. 用于稳健离策略评估中处理混杂因素的资格迹：一种因果方法。在第*41*届人工智能不确定性会议 *(UAI)*，2025年。

Zhang, K.、Koppel, A.、Zhu, H. 与 Basar, T. 策略梯度方法全局收敛至（几乎）局部最优策略。*SIAM* 控制与优化期刊，58(6):3586–3612，2020年。

Zhang, Y.、Yu, S.、Zhang, T.、Guang, M.、Hui, H.、Long, K.、Wang, Y.、Yu, C. 和 Ding, W.，SAC Flow：通过速度重参数化序列建模实现基于流的策略的样本高效强化学习，2026年。网址 https://arxiv.org/abs/2509.25756。

A. 相关工作

离策略因果强化学习。当无未观测混杂假设不成立时，需要先识别奖励和转移分布，再评估策略价值，或者对可能的策略价值进行定界。关于直接从混杂数据中识别策略价值，已有丰富的研究文献（Shi 等，2022；Miao 等，2022；Guo 等，2022；Bennett & Kallus，2024）。但这些研究通常会引入其他关键的学习假设，例如近端因果推断文献脉络中桥函数的存在性（Tchetgen 等，2020）。另一方面，在没有额外假设的情况下，可以利用定界方法涵盖策略价值的全部可能范围。Manski 的奠基性工作（Manski，1989）在带有上下文信息的单阶段处理模型（即上下文赌博机模型）中，利用观测数据为不可识别设定下的因果效应构造了最早的界限。随后，这些界限被扩展到工具变量设定（Balke & Pearl，1997；Imbens & Angrist，1994）、反事实因果概率的部分识别（Tian & Pearl，2000），以及自动构造奖励塑造函数（Li 等，2025a）。本工作受到近期两项研究的启发：通过定界对策略价值进行部分识别（Zhang & Bareinboim，2025），以及在离散动作空间中开展混杂稳健离策略学习（Li 等，2025b）。

离线强化学习。离线强化学习（Levine 等人，2020年）研究如何从由其他行为策略收集的静态数据集中学习最优策略，同时使其与数据集中的行为状态-动作分布保持紧密一致。文献中研究了多种用于施加这一约束的策略，包括但不限于保守性（Kumar 等人，2020年；Nakamoto 等人，2023年）、生成建模（Chen 等人，2021年；Janner 等人，2021年）、分布外检测（Garg 等人，2023年）以及行为正则化（Fujimoto 和 Gu，2021年；Tarasov 等，2023年；Park 等人，2025b；Dong 等人，2025年；Agrawalla 等人，2025年）。离线训练后，可以加入在线微调阶段，进一步提升智能体的性能。为避免突发的分布偏移导致遗忘（Nakamoto 等人，2023年），可以采用均衡采样（Song 等人，2023年；Yang 等人，2023年）、在回放缓冲区中保留离线数据（Park 等人，2025b；Nair 等人，2021年；Ball 等人，2023年），或对 Q 值施加惩罚（Nakamoto 等人，2023年）。

强化学习中的流匹配。得益于流匹配强大的多模态分布建模能力，以及相较于基于扩散的方法中求解偏微分方程更容易实现，流匹配（Lipman 等，2022；Liu 等人，2022；Tong 等人，2024）近年来在强化学习领域受到广泛关注。它在强化学习中的应用涵盖规划（Nandiraju 等人，2025；Nguyen 等人，2025）、世界模型学习（Liu 等人，2025；Rohbeck 等人，2025；Lillemark 等人，2026）、在线学习（Lv 等人，2025）、离线学习（Park 等人，2025b；Dong 等人，2025；Agrawalla 等人，2025；Tiofack 等人，2025；Zhang 等人，2026；Ghugare 和 Eysenbach，2025），以及多模态视觉-语言-动作模型（VLAs）（Black 等人，2026；Intelligence 等人，2025；Deng 等人，2025；Jiang 等人，2025）。

B. 局限性

所提出的 Causal-FQL 主要有两类局限性。第一类局限性与 FQL（Park 等人，2025b）有关：1) 训练过程中需要对常微分方程进行数值求解，可能导致训练速度较慢；2) FQL 仅采用了最简单形式的流匹配。结合近期的发展成果，例如整流流（Liu 等人，2022）和最优传输条件流匹配（Tong 等人，2024年），应能带来更快的收敛、更稳定的训练以及更好的性能；3) FQL 没有内置的探索机制，这可能会限制其在线微调性能。第二类局限性源于对因果离线强化学习目标的近似。随着我们向 FQL 流程中引入更多组件，可调超参数的组合空间也变得比以往更难调整。鉴于基于像素的离线强化学习任务计算开销较大，在开展更大规模的应用之前，下一步将是设计一种更简单的抗混杂离线强化学习方法。

C. 扩展预备知识与证明

因果基础。我们简要回顾因果推断的基础，以便于后续讨论。我们的 CMDP 定义与结构因果模型的定义密切相关（Pearl，2009）。

定义 C.1。一个 SCM $\mathcal{M}$ 是一个元组 $\langle \boldsymbol{U}, \boldsymbol{V}, \mathscr{F}, P \rangle$，其中

- $\boldsymbol{U}$ 是一个外生变量集合；
- $\boldsymbol{V}$ 是一个内生变量集合；

- $\mathscr{F}$ 是一组函数，使得每个 $f_V \in \mathscr{F}$；
- 外生分布 $P(\boldsymbol{U})$。

f_V 根据系统中其他变量的组合作为参数来决定内生变量 $V \in \boldsymbol{V}$ 的取值。也就是说，$V \leftarrow f_V(\boldsymbol{PA}_V, \boldsymbol{U}_V)$，$\boldsymbol{PA}_V \subseteq \boldsymbol{V}$，$\boldsymbol{U}_V \subseteq \boldsymbol{U}$。外生变量 $\boldsymbol{U}$ 的取值从 $P(\boldsymbol{U})$ 中抽取。

变量子集 $\boldsymbol{X} \subseteq \boldsymbol{V}$ 上的策略 π 是一系列决策规则 $\{\pi(X|\boldsymbol{S}_X)\}_{X\in\boldsymbol{X}}$，其中每个 $\pi(X|\boldsymbol{S}_X)$ 都是一个概率分布，将一组协变量 $\boldsymbol{S}_X \subseteq \boldsymbol{V}$ 的定义域映射到动作 X 的定义域。变量 $\boldsymbol{X}$ 上的策略 π 干预记为 do(π)，是将每个 $X \in \boldsymbol{X}$ 的取值设定为由策略 $X \sim \pi(X|\boldsymbol{S}_X)$ 决定的值的操作（Correa 与 Bareinboim，2020年），以替代通常决定其取值的函数 $f_{\boldsymbol{X}} = \{f_X : \forall X \in \boldsymbol{X}\}$。对于 SCM $\mathcal{M}$，令 $\mathcal{M}_\pi$ 为由干预 do(π) 诱导的 M 的子模型。对于集合 $\boldsymbol{Y} \subseteq \boldsymbol{V}$，干预分布 $P_\pi(\boldsymbol{Y})$ 定义为子模型 $\mathcal{M}_\pi$ 中 $\boldsymbol{Y}$ 上的分布，即 $P_{\mathcal{M}}(\boldsymbol{Y};\pi) \triangleq P_{\mathcal{M}_\pi}(\boldsymbol{Y})$；当显而易见时，限制条件 $\mathcal{M}$ 不予显式说明。

每个结构因果模型 $\mathcal{M}$ 还对应一个因果图 $\mathcal{G}$（例如图 2），它是一个有向无环图（DAG），其中节点表示内生变量 $\boldsymbol{V}$，箭头表示每个结构函数 $f_V \in \mathscr{F}$ 的参数 $\text{PA}_V, \boldsymbol{U}_V$。按照惯例，外生变量 $\boldsymbol{U}$ 通常不会显式显示。然而，双向箭头 $V_i \leftrightarrow V_j$ 表示存在未观测混杂因素（UC），$U_{i,j} \in \boldsymbol{U}$ 同时影响 V_i, V_j（Bareinboim 等，2022）。我们将使用标准的图论家族缩写来表示图关系，例如父节点（pa）、子节点（ch）、后代节点（de）和祖先节点（an）。例如，$\mathcal{G}$ 中 $\boldsymbol{X}$ 的父节点集合记作 $\text{pa}(\boldsymbol{X})_\mathcal{G} = \cup_{X\in\boldsymbol{X}}\text{pa}(X)_\mathcal{G}$。大写形式 Pa、Ch、De、An 也包括该参数本身，例如 $\text{Pa}(\boldsymbol{X})_\mathcal{G} = \text{pa}(\boldsymbol{X})_\mathcal{G} \cup \boldsymbol{X}$。在 $\mathcal{G}$ 中，从节点 X 到节点 Y 的一条路径，是指一组边构成的序列，其中任何特定节点都不会出现多次。若 DAG $\mathcal{G}$ 中的两组节点 $\boldsymbol{X}, \boldsymbol{Y}$ 被第三组节点 $\boldsymbol{Z}$ d-分离，则记作 $(\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}|\boldsymbol{Z})_\mathcal{G}$；这表示从 $\boldsymbol{X}$ 中的节点到 $\boldsymbol{Y}$ 中的节点的每条边路径，都会被 $\boldsymbol{Z}$ 中的节点“阻断”。阻断判据遵循 Pearl（2009，定义 1.2.3）。关于结构因果模型的更多详情，请参阅 Pearl（2009）和 Bareinboim 等人（2022）。

从 MDP 到 CMDP。 在 MDP 中，所有状态变量对智能体都是可观测的。从示范者的视角转换到学习者的视角时，这一假设被隐含地预先采纳了。然而，正如我们在正文中所展示的，这通常未必成立。我们认为，强化学习的其他方面也存在同样的问题（Bareinboim 等人， 2024年），例如离策略学习（Joshi 等人， 2024a；Zhang&Bareinboim，2019年； Li 等， 2025b）、课程学习（Narvekar 等人， 2020年；Li 等， 2024年）以及奖励塑形（Li 等， 2025a）。这促使我们提出 CMDP，将其作为一种恰当的建模方式，用于刻画混杂偏差，同时保持马尔可夫性质，以实现高效的策略学习。

与 MDP 的图示相比，CMDP 仅在每个时间步内增加了三条双向箭头。根据我们上文讨论的因果图定义，这表明当前时间步的三个变量——动作 X_t、奖励 Y_t 和下一状态 S_{t+1}——可能受到相同潜在混杂变量的影响。在基于像素的离线强化学习任务中，这些变量可能对应于被遮挡的关节角度、末端执行器的精确角速度或物体的精确位置；这些信息都是示范者的策略所依赖的，但学习者无法通过像素观测完全获取。

定理 3.1 的证明 现在我们可以给出定理中因果策略梯度目标的详细证明。 3.1。**定理 C.2** (定理 3.1 的重述)。 对于 *CMDP* 环境 $\mathcal{M}$，其奖励信号为 $Y_t \in [a,b] \subseteq \mathbb{R}$，固定一个策略 π。对于任意状态 $V_\pi(s) \geq V_\pi(s)$，状态价值函数 $s \in \mathcal{S}$ 的下界 $\underline{V_\pi}(s)$ 如下所示：

$$\underline{V_\pi}(s) = \mathbb{E}_{\substack{x\sim p(\cdot|s)\\ x',x^*\sim\pi(\cdot|s)}}\left[\mathbf{1}_{x\neq x'}\left(a + \gamma \min_{s^*} \underline{Q_\pi}(s^*, x^*)\right) + \mathbf{1}_{x=x'}\left(\widetilde{\mathcal{R}}(s,x) + \gamma \sum_{s'} \widetilde{\mathcal{T}}(s,x,s')\, \underline{Q_\pi}(s', x^*)\right)\right] \tag{14}$$

证明。由于 $\underline{V_\pi}(s) = \sum_x \pi(x \mid s)\underline{Q_\pi}(s, x)$，由公式（5）可知，状态价值函数的下界为

$$\underline{V_\pi}(s) = \sum_x \pi(x \mid s)\mu(\neg x \mid s)\left(a + \gamma \min_{s'} \underline{V_\pi}(s')\right) \tag{15}$$

$$+ \sum_x \pi(x \mid s)\mu(x \mid s)\left(\widetilde{\mathcal{R}}(s,x) + \gamma \sum_{s',x'} \widetilde{\mathcal{T}}(s,x,s')\, \underline{V_\pi}(s')\right) \tag{16}$$

沿用（Zhang &Bareinboim，2024年，定理1）中的简化步骤，

$$\underline{V_\pi}(s) = \sum_x \mu(x \mid s)\pi(\neg x \mid s)\left(a + \gamma \min_{s'} \underline{V_\pi}(s')\right) \tag{17}$$

$$+ \sum_x \mu(x \mid s)\pi(x \mid s)\left(\widetilde{\mathcal{R}}(s,x) + \gamma \sum_{s',x'} \widetilde{\mathcal{T}}(s,x,s')\,\underline{V_\pi}(s')\right) \tag{18}$$

上式还可以进一步写为

$$\underline{V_\pi}(s) = \sum_{x,x'} \mu(x \mid s)\pi(x' \mid s)\mathbf{1}_{x \neq x'}\left(a + \gamma \min_{s'} \underline{V_\pi}(s')\right) \tag{19}$$

$$+ \sum_{x,x'} \mu(x \mid s)\pi(x' \mid s)\mathbf{1}_{x = x'}\left(\widetilde{\mathcal{R}}(s,x) + \gamma \sum_{s',x'} \widetilde{\mathcal{T}}(s,x,s')\,\underline{V_\pi}(s')\right) \tag{20}$$

最后，将对策略乘积 $\mu(x \mid s)\pi(x' \mid s)$ 的求和替换为期望，即可证明该结论。 □

D. 补充结果

除操作任务外，我们还在视觉运动任务（例如`visual-antmaze`系列）上测试了 Causal-FQL。其结果不如操作任务上的结果理想，这符合预期。我们推测，这是因为 FQL 和 Causal-FQL 直接从策略中采样动作，而没有考虑 Q 值的噪声分布；相比之下，高斯策略提取结合拒绝采样在处理单峰奖励分布方面表现更好。因此，BRAC 等较早的高斯 SOTA 方法在运动任务中表现良好。作为对我们这一假设的间接证明，在第D节中，ReBRAC（Tarasov等人，2023年）也优于其自身基于流的对应方法 FBRAC（Park 等人，2025b），而 IFQL（采用拒绝采样的 FQL）也优于 FQL。

表*3.* **运动离线评估结果**。 Causal-FQL 和原始FQL 的表现都不及基于高斯的方法。我们报告了4个随机种子的平均结果，并按照先前的工作（Park 等人，2025b），将各领域中达到最佳表现95%以内的数值加粗。

	高斯策略		流策略				
	IQL	ReBRAC	FBRAC	IQN	IFQL	FQL	Causal-FQL
视觉蚂蚁迷宫-中等-导航-单任务-任务1-v0 (*)	**78 ± 9**	54 ± 15	27 ± 3	62 ± 7	**81 ± 3**	32 ± 3	40 ± 15
视觉蚂蚁迷宫-中等-导航-单任务-任务2-v0	90 ± 3	**96 ± 1**	42 ± 4	88 ± 2	87 ± 1	60 ± 2	61 ± 2
视觉蚂蚁迷宫-中等-导航-单任务-任务3-v0	80 ± 6	**97 ± 1**	32 ± 4	64 ± 5	92 ± 1	35 ± 8	45 ± 3
视觉蚂蚁迷宫-中等-导航-单任务-任务4-v0	**89 ± 4**	**93 ± 2**	23 ± 2	71 ± 6	84 ± 3	35 ± 2	39 ± 2
视觉蚂蚁迷宫-中等-导航-单任务-任务5-v0	84 ± 2	**97 ± 1**	25 ± 4	84 ± 1	89 ± 2	29 ± 7	29 ± 6
视觉蚂蚁迷宫-传送-导航-单任务-任务1-v0 (*)	5 ± 2	2 ± 0	1 ± 1	2 ± 1	**7 ± 4**	2 ± 1	4 ± 2
视觉蚂蚁迷宫-传送-导航-单任务-任务2-v0	10 ± 2	10 ± 3	6 ± 5	7 ± 3	**13 ± 3**	6 ± 1	7 ± 4
视觉蚂蚁迷宫-传送-导航-单任务-task3-v0	7 ± 7	4 ± 1	10 ± 4	6 ± 4	8 ± 9	9 ± 4	**14 ± 5**
视觉蚂蚁迷宫-传送-导航-单任务-task4-v0	4 ± 6	4 ± 0	10 ± 2	4 ± 2	**18 ± 2**	9 ± 1	11 ± 3
视觉蚂蚁迷宫-传送-导航-单任务-task5-v0	2 ± 1	2 ± 1	2 ± 1	2 ± 1	**4 ± 2**	1 ± 1	**4 ± 2**

E. 实现细节

`Causal-FQL` 基于 FQL 的代码库构建（Park 等人，2025b）[3]。`Causal-FQL` 继承了 FQL 的大部分架构设计，仅针对因果离线强化学习目标进行了最小限度的修改。首先，我们添加了一个动作判别器和一种抗混杂 Q 加权方案：构建一个额外的动作判别器网络（其规模与 Q 评论家相同），在演员损失中使用该网络对流动作与单步目标流动作进行分类，根据判别器逻辑值计算事实权重，并按照公式（11）混合集成 Q 值；此外，还添加了一个可选用指数衰减的判别器损失项，并设置了一个在线学习阶段的开关，用于禁用判别器加权，改为使用集成上的普通平均 Q 值进行训练。由于计算资源有限，我们无法对`Causal-FQL` 在视觉任务上的完整超参数空间进行全面调优。不过，我们确实观察到，在所有视觉任务中，使用默认 FQL 超参数的性能都稳定优于 FQL。

[3] https://github.com/seohongpark/fql

F. 实验细节

任务与环境。 我们使用 OGBench（Park 等人，2025a）中的基于像素的任务进行评估。我们在 5 个环境中的 25 个视觉任务上进行评估。下面列出了我们使用的数据集，每个数据集定义了 5 个奖励最大化任务，索引从 -任务1 到 -任务5。在正文中，我们评估了涉及多个步骤子任务的操作任务，其奖励范围介于子任务数量的负值与 0 之间。在第 D 节中，我们评估了运动任务，其奖励为 -1 或 0，表示是否达到目标。

- 视觉立方体单人游戏单任务
- 视觉立方体双人游戏单任务
- 视觉场景游戏单任务
- visual-puzzle-3x3-play-singletask
- visual-puzzle-4x4-play-singletask

方法与超参数。我们采用 FQL（Park 等人，2025b）和价值流（Dong 等人，2025年）中的基线实现。[4] 下面列出我们使用的基线方法。我们采用 FQL 和价值流代码库中推荐的默认超参数。有关复现结果的详细命令，请参阅脚注中的链接。

- **IQL** (Kostrikov 等人，2021）。隐式 Q 学习使用期望分位数回归来表示 Q 值，随后据此选择最优动作。我们使用代码库中推荐的默认逆温度 α。
- **ReBRAC** (Tarasov 等人，2023年）。ReBRAC 是一种基于高斯策略的离线强化学习方法，它使用 TD3 学习价值，并采用行为正则化。
- **FBRAC** (Park 等人，2025b）。这是 BRAC 的一种变体，改用流策略。它不使用单步流策略提取，因此需要进行时间反向传播。
- **IQN** (Dabney 等人，2018年）。IQN 是一种分布式强化学习基线，通过分位数值表示 Q 值分布。动作选择基于对采样动作的 Q 值取 argmax。
- **IFQL** (Park 等人，2025b）。隐式扩散 Q（IDQL，(Hansen-Estruch 等人，2023年）的一种变体，采用流策略和拒绝采样进行动作选择，该方法最早由 (Park 等人，2025b）提出。
- **FQL** (Park 等人，2025b）。FQL 使用单步流策略来最大化评论家集成所学习的 Q 值。它使用行为克隆流策略进行行为正则化。

G. 因果价值流

这里介绍因果价值流的实现细节，以及将新提出的因果离线强化学习目标应用于价值流（Dong 等人，2025年）的初步离线评估结果，以此展示我们方法的通用性。我们采用 Causal-FQL 中的判别器系数和集成数量设置，并将价值流中的其他参数保持为默认值。与 Causal-FQL 一样，我们额外添加了一个动作判别器网络，其规模与 Q 评论家网络相同。计算公式（11）时，我们使用通过欧拉方法和训练好的回报向量场集成采样得到的值。随后，我们对平均 Q 值重新加权，并将最小 Q 值作为最坏情况值加入。价值流的其他部分保持不变。

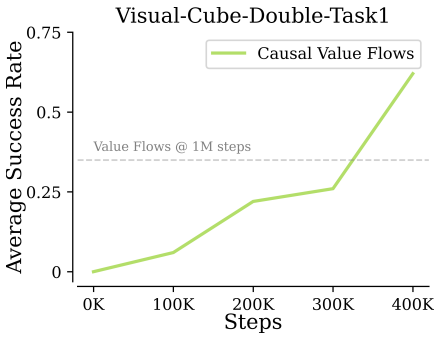


图 7. **因果价值流优于基础版价值流**。

无需进一步调参，在 visual-cube-double-play 的任务 1 上，增强型因果价值流仅需 40 万步离线训练，成功率便能超过 0.6，**其性能几乎是基础价值流的两倍**；而基础价值流经过 100 万步离线训练后，成功率也只有 0.32。

[4] https://github.com/chongyi-zheng/value-flows