arXiv:2602.21072v1 [cs.LG] 24 Feb 2026

Duke
UNIVERSITY

面向动力学偏移离线强化学习的局部动力学感知域适配

Zhangjie Xia[*‡] Yu Yang[†‡] Pan Xu[§]

2026年2月25日

摘要

动力学偏移离线强化学习（RL）旨在利用有限的目标数据和大量在不同转移动力学下收集的源数据，为目标域学习策略。现有方法通常在整个状态空间范围内全局处理动力学失配，或通过逐点数据过滤来处理；这些方法可能会遗漏局部的跨域相似性，或带来高昂的计算成本。我们提出局部化动力学感知域适应（LoDADA），利用局部动力学失配，以更好地复用源数据。LoDADA对源数据集和目标数据集中的状态转移进行聚类，并通过域判别估计簇级动力学差异。保留差异较小的簇中的源转移，并过滤掉差异较大的簇中的源转移。由此得到一种细粒度且可扩展的数据选择策略，既避免了过于粗略的全局假设，也避免了代价高昂的逐样本过滤。我们提供了理论洞见，并在具有多种全局和局部动力学偏移的环境中进行了广泛实验。结果表明，通过更好地利用局部分布不匹配，LoDADA始终优于最先进的动力学偏移离线强化学习方法。

1 引言

强化学习（RL）（Kaelbling 等人，1996；Sutton 等人，1998）的目标是通过与环境交互并接收奖励反馈，学习一种用于最优决策的策略。然而，在自动驾驶（Kiran 等人，2021）和医疗保健（Coronato 等人，2020）等现实世界应用中，与环境（即目标域）直接交互可能十分危险或耗时。因此，实践者通常先在结构相似的源域中收集经验，再迁移学习得到的策略。一个关键挑战在于，由于转移动力学不匹配，在源域中训练的策略在部署到目标域时可能表现不佳。该问题通常被称为离动力学强化学习（Eysenbach 等人，2021；He 等人，2025年）。本文研究一种实际设定，即动力学偏移离线强化学习（Liu 等人，2022；Wang 等人，2024），其中智能体无法与

[*]纽约大学；邮箱：zx1357@nyu.edu
[†]杜克大学；邮箱：yu.yang@duke.edu
[‡]共同贡献[§]杜克大学；邮箱：
pan.xu@duke.edu

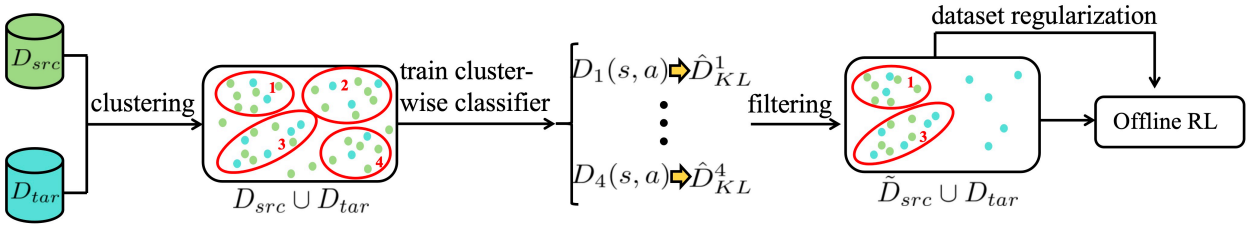


图1：**我们提出的框架概览**。首先，我们在混合源域与目标域数据集上进行 K-means 聚类。然后，估计每个簇中源域与目标域动力学之间的局部KL散度。我们使用逐簇分类器，有选择地保留源域转移，供下游离线强化学习算法使用。此外，我们还引入数据集正则化项，以确保策略与目标数据集保持一致。

环境在线交互，而必须转而从有限目标数据集以及在不同动力学下收集的更大规模源数据集中学习。

人们对动力学偏移离线强化学习的兴趣日益浓厚，相关先前工作大致可分为三类：(i) 使用动力学感知正则化修改源数据中奖励的数据增强方法（Liu 等，2022；Lyu 等，2024a；Wang 等，2024；Guo 等，2024）；(ii) 根据源转移与目标域转移的相似性来筛选源转移的数据过滤方法（Xu 等，2023；Wen 等人，2024；Lyu 等，2025）；以及近期出现的 (iii) 利用源数据学习目标动力学并生成合成目标样本的数据生成方法（Guo 等，2026）。然而，大多数方法都在整个状态-动作空间范围内全局处理源域与目标域动力学不匹配问题，这可能导致在域之间一致的区域过于保守，而在差异显著的区域又过于乐观。例如，Liu 等人（2022）等方法在整个数据集上针对 (s, a, s') 元组训练单个域分类器，而在目标数据有限时，这一做法颇具挑战。一个值得注意的例外是基于最优传输的逐点数据过滤（Lyu 等，2025），但其计算成本会随数据集大小增长，限制了其实用性 Dvurechensky 等人（2018）。因此，现有方法可能造成数据利用效率低下，并且往往无法利用细粒度局部相似性，而这些相似性本可帮助实现从源域到目标域更稳健、高效的迁移。

我们提出的方法 **Lo**局部**D**动力学感知**AD**域自适应（**Adaptation**，**LoDADA**）以一种新颖的方式应对这些挑战。不同于依赖全局或逐点假设的以往方法，LoDADA 通过 K-means 对源域和目标域的下一状态进行聚类，以识别状态空间中两个数据集对齐良好的局部区域。我们舍弃远离簇中心的源样本，保留与目标域更匹配的样本。在每个簇内，我们训练一个分类器来区分源样本和目标样本，并利用其输出估计源域与目标域转移动力学之间的局部 KL 散度。估计的 KL 散度较大的簇表明源域与目标域状态转移不一致，因此被视为可靠性较低。LoDADA 不会舍弃整个簇，而是保留所有簇中的样本，以维持状态空间覆盖度，同时优先采用估计的 KL 散度较小的簇中的状态转移。在策略优化阶段，我们进一步引入一个正则化项，用于惩罚学习得到的策略与拟合目标数据集而成的行为策略之间的偏差，从而鼓励学习得到的策略保持接近目标域的行为。

通过根据簇级 KL 估计选择性地保留源转移，LoDADA 减少了动力学失配引起的偏差，并能更有效地利用源数据。这种基于聚类的框架为在局部动力学偏移下平衡选择性过滤与跨域适应提供了一种有原则的方法。过滤步骤之后，我们使用隐式 Q 学习（IQL）优化策略

（Kostrikov 等人，2021）作为骨干方法，并加入一个正则化项，鼓励学习得到的策略接近目标域的行为策略。我们的主要贡献总结如下：

- 我们指出，现有异动力学离线强化学习方法存在一个关键局限：它们要么在全局层面、要么在逐点层面处理动力学失配，这可能导致数据低效和性能下降。为解决这一问题，我们提出 LoDADA，通过估计局部源目标动力学差异进行按簇过滤，并保留与目标动力学最一致的源转移。
- 我们进行了理论分析，为 LoDADA 学得的策略建立了回报下界，从而在动力学失配的情况下为目标域中的性能提供保证。
- 我们在离线设定下，于四种 MuJoCo 环境中对所提方法进行了实证评估，涵盖不同类型和程度的异动力学偏移。结果表明，在重力和摩擦偏移下，我们的方法平均较基线方法提升 19.9%，在形态偏移下提升 32.7%，在局部扰动下提升 29.3%。此外，我们还展示了 LoDADA 在 AntMaze 的高难度导航任务和 Adroit 的操作任务中的优异表现。

2 相关工作

强化学习中的领域适应。强化学习中的领域适应侧重于在具有不同转移动力学的域之间适应策略。先前工作包括策略与价值函数迁移，即复用源域知识以加速目标域中的学习（Abel 等人，2018；Xu 等人，2023），以及鲁棒马尔可夫决策过程，即在预定义的不确定性集合内针对最坏情况动力学优化策略（Liu 和 Xu，2024；Liu 等人，2024b；He 等人，2025年；Gu 等人，2025年）。然而，这些方法通常需要精心指定的不确定性集合，并且可能无法覆盖真实的目标动力学。更重要的是，由于它们僵化地假设完全无法获取目标数据，因此往往过于保守。更新近的工作探索了利用由动力学失配产生的奖励增广进行在线强化学习领域适应（Eysenbach 等人，2021；Guo 等人，2024）以及表示偏差（Lyu 等人，2024a）。相比之下，我们的方法通过基于聚类的框架，在纯离线设定中解决领域适应问题，从而在显著的动力学偏移下更灵活、高效地复用源数据。

动力学偏移离线强化学习。动力学偏移离线强化学习旨在利用丰富的源数据集和有限目标数据集，提升目标域中的策略学习效果；在该目标域中，在线交互不可行。以往研究通过重要性加权对齐源分布与目标分布（Nachum 等人，2019；Lee 等人，2021）、进行奖励增广（Liu 等人，2022）、使用对抗性领域适应减少状态动作动力学的不匹配（Barde 等人，2020；Xu 等人，2021），或根据源数据与目标域的相似度过滤源数据（Liu 等人，2020；Wen 等人，2024；Lyu 等，2025）来解决这一问题。另一类研究采用基于模型的方法，在目标域中学习动力学模型，从而在动力学不匹配的情况下以更有原则的方式重用源转移（Yu 等人，2020；Guo 等人，2026）。这些方法往往难以应对严重的分布变化：在覆盖范围有限时，重要性加权会遭遇较高方差；对抗性适应在实践中可能不稳定；启发式过滤可能降低样本效率；而基于模型的方法则容易受到模型偏差的影响。相比之下，我们的方法通过保留有用的源经验，同时降低不可靠簇的权重，在稳健性与覆盖范围之间取得平衡。

基于聚类的强化学习方法 基于聚类的强化学习方法通过将相似样本分组成簇，利用状态-动作空间中的结构，从而降低方差并提升泛化能力。近期方法将聚类用于表征学习，例如学习能够捕捉环境动态的潜在嵌入，然后对其进行聚类，以发现可复用行为或中间目标（Machado 等人，2017；Eysenbach 等人，2019）。这些方法通常依赖启发式相似度度量，

而这些度量可能无法准确反映潜在的转移动力学。我们的方法利用聚类不仅对数据进行结构化，还将分布对齐整合到各个簇中，确保分组后的样本既具有一致性，又与目标域的动力学相关。

符号说明： $\mathbb{H}(X)$ 是随机变量 X 的熵。 $z = s \oplus a$ 是状态 s和动作 a的向量拼接，可视为 $(s,\ a)$ 的恒等表示。

3 预备知识

强化学习问题通常表述为一个马尔可夫决策过程（MDP），由元组 $\mathcal{M} =(\mathcal{S},\ \mathcal{A},\ P,\ r,\ \gamma,\ \rho_0)$ 指定，其中 $\mathcal{S}$ 是状态空间， $\mathcal{A}$ 是动作空间， $P(s'|s,\ a)$ 是转移核， $r(s,\ a)$ 是给定状态 - 动作对 $(s,\ a)$ 的奖励函数， $\gamma \in [0, 1)$ 是折扣因子， ρ_0 是初始状态分布。

在动力学偏移离线设定中，我们考虑两个马尔可夫决策过程：源域 $\mathcal{M}_{\text{src}} = (\mathcal{S}, \mathcal{A}, P_{\text{src}}, r, \gamma, \rho_0)$ 和目标域 $\mathcal{M}_{\text{tar}} = (\mathcal{S}, \mathcal{A}, P_{\text{tar}}, r, \gamma, \rho_0)$。两个域仅在转移动力学上有所不同，即 $P_{\text{src}}(s'|s, a) \neq P_{\text{tar}}(s'|s, a)$，而共享相同的奖励函数 $r_{\text{src}}(s, a) = r_{\text{tar}}(s, a)$。在离线设定中，智能体无法与环境交互，必须从静态数据集中学习（Levine 等，2020）。假设给定一个源数据集 $D_{\text{src}} = \{(s, a, s', r)_{\text{src}}\}$ 和一个规模较小的目标数据集$D_{\text{tar}} = \{(s, a, s', r)_{\text{tar}}\}$。目标是学习策略 π，通过利用混合数据集 $D_{\text{mix}} = D_{\text{src}} \cup D_{\text{tar}}$，最大化其在目标域中的期望回报 $\max_\pi \mathbb{E}_{\pi, P_{\text{tar}}}\left[\sum_{t=0}^{\infty} \gamma^t r_{\text{tar}}(s_t, a_t)\right]$。通常有 $|D_{\text{tar}}| \ll |D_{\text{src}}|$。

4 方法

本节分析最优目标策略与基于筛选后的源数据集学习得到的策略之间的目标域性能差距。该分析促使我们采用一种基于聚类的流程，估计局部 KL 散度以筛选源转移，并将其融入策略优化。最后，我们总结了异动力学离线强化学习的完整算法。所有证明均见A 节。

4.1 理论洞见

˜异动力学设定下数据筛选的目标，是从 D_{src} 源数据集˜中构建筛选后的源数据集 D_{src}，使其与目标数据集 D_{src} 之间的动力学差异 D_{tar} 最小化。遵循Lyu 等人 (2024a的工作，我们有如下命题。

命题 4.1。 对于任意 (s, a)，记其表示为 z，并假设 $s'_{\text{src}} \sim P_{\mathcal{M}_{\text{src}}}(\cdot|z)$ $s'_{\text{tar}} \sim P_{\mathcal{M}_{\text{tar}}}(\cdot|z)$。那么，测量源域与目标域之间的表示偏差，等价于测量两个域之间的动力学失配，即 $D_{\text{KL}}\big(P(z|s'_{\text{src}}) \| P(z|s'_{\text{tar}})\big) = D_{\text{KL}}\big(P(s'_{\text{src}}|z) \| P(s'_{\text{tar}}|z)\big) + \mathbb{H}(s'_{\text{src}}) - \mathbb{H}(s'_{\text{tar}})$。

定理 4.1 与 Lyu 等人 （2024a）的定理 4.3 之间的区别在于 KL 散度的方向。我们推导的是从源表示分布到目标表示分布的 KL 散度（而非反向）。这一选择符合我们的目标，即刻画从筛选后的源数据中学习到的策略与目标最优策略之间可能相差多远（定理 4.2）。由于 KL 散度具有非对称性，我们在 A 节 重现了证明，以确保完整性。

由于对于固定的源域/目标域，熵项是常数，因此最小化动力学偏差等价于最小化表示偏差。具体而言，令 $\mathbb{H}(s'_{\text{src}}) - \mathbb{H}(s'_{\text{tar}}) = B$ 为常数。

如果表示偏差的上界为 ϵ，即 $D_{\mathrm{KL}}(P(z|s'_{\mathrm{src}})\|P(z|s'_{\mathrm{tar}})) \leq \epsilon$，那么动力学偏差 $D_{\mathrm{KL}}(P(s'_{\mathrm{src}}|z)\|P(s'_{\mathrm{tar}}|z))$ 的上界为 $\epsilon - B$。

以下假设来自 Eysenbach 等人 (2021)，量化了最优目标策略 π^* 在源动力学下评估时可能遭受的性能差距。

假设 1. 令 $\pi^* = \arg\max_\pi \mathbb{E}_p\left[\sum \gamma^t r(s_t, a_t)\right]$ 为目标域中的奖励最大化策略。则源域与目标域中的期望奖励至多相差 $2R_{\max}\sqrt{1-e^{B-\epsilon}}$:

$$\left|\mathbb{E}_{\pi^*,P_{src}}\left[\sum_t \gamma^t r(s_t,a_t)\right] - \mathbb{E}_{\pi^*,P_{tar}}\left[\sum_t \gamma^t r(s_t,a_t)\right]\right| \leq 2R_{\max}\sqrt{1-e^{B-\epsilon}},$$

其中 $B = \mathbb{H}(s'_{src}) - \mathbb{H}(s'_{tar})$。

现在，我们给出目标域中由筛选后的源数据集学习得到的策略与最优目标策略之间性能差距的上界。

定理 *4.2*（离线性能界）。令 $\hat{\pi}^*_{\tilde{D}_{src}}$ 为在筛选后的~源数据集 D_{src}中最大化奖励的策略，$\pi^*_{D_{tar}}$ 为在未过滤的目标数据集 D_{tar}中最大化奖励的策略，π^* 为在目标域中最大化奖励的策略。假设 π^* 满足假设 *1*，并令 C 为无动力学偏移时的离线强化学习性能界，即 $\mathbb{E}_{P_{tar},\pi^*}\left[\sum \gamma^t r(s_t,a_t)\right] - \mathbb{E}_{P_{tar},\pi^*_{D_{tar}}}\left[\sum \gamma^t r(s_t,a_t)\right] \leq C$。则 $\hat{\pi}^*_{\tilde{D}_{src}}$ 在目标域上获得近最优奖励:

$$\mathbb{E}_{P_{tar},\hat{\pi}^*_{\tilde{D}_{src}}}\left[\sum_t \gamma^t r(s_t,a_t)\right] \geq \mathbb{E}_{P_{tar},\pi^*}\left[\sum_t \gamma^t r(s_t,a_t)\right] - \underbrace{C}_{\textit{offline algorithm error}} - \frac{4R_{\max}}{1-\gamma}\underbrace{\sqrt{1-e^{B-\epsilon}}}_{\textit{dynamics deviation}} - \frac{2R_{\max}}{1-\gamma}\underbrace{\mathbb{E}_{\rho^{\hat{\pi}^*_{\tilde{D}_{src}}}_{\mathcal{M}_{src}},P_{src}}\left[D_{\mathrm{TV}}\left(\hat{\pi}^*_{\tilde{D}_{src}}(\cdot|s')\|\pi^*_{D_{tar}}(\cdot|s')\right)\right]}_{\textit{filtering deviation}}.$$

注释4.3。 与 Wen 等人 (2024) 和 Lyu 等人 (2025年) 给出的性能界相比，我们的定理具有两项主要优势：(1) 针对真正的最优目标策略提供保证： 该界将学习得到的策略在目标域中的回报直接与最优目标策略进行比较，从而确保保证是相对于目标环境中可实现的最佳表现给出的。(2) 无需依赖源行为策略： 以往研究通常与源行为策略进行比较，这实际上偏向于模仿源数据集。相比之下，我们的界以在目标数据集上训练的策略为参照来衡量策略偏差，这更能体现过滤的目标。

4.2 基于聚类的 KL 估计

定理 4.2 为基于聚类的过滤策略提供了依据。具体而言，该界表明，有效的过滤应同时降低 动力学偏差 和 过滤偏差。因此，为了降低策略~之间的差异 $\hat{\pi}^*_{\tilde{D}_{\mathrm{src}}}$ 与 $\pi^*_{D_{\mathrm{tar}}}$，筛选后的源数据集 D_{src} 应与目标数据集 D_{tar} 相匹配，衡量依据为动力学失配的 KL 散度。

我们也可以将动力学失配等价地视为表征层面的差异。具体而言，我们提出了一种用于估计表征散度 $D_{\mathrm{KL}}\left(P(z|s'_{\mathrm{src}})\|P(z|s'_{\mathrm{tar}})\right)$ 的基于聚类的估计器。该

方法的核心思想是将相似的下一状态聚类，使得在每个簇内，源域与目标域状态-动作对之间的差异主要反映<sty le id='1'>局部化动力学失配。

给定两个数据集 $D_{\text{src}} = \{(s_i, a_i, s_i')\}_{i=1}^{N_{\text{src}}}$和 $D_{\text{tar}} = \{(s_j, a_j, s_j')\}_{j=1}^{N_{\text{tar}}}$，我们首先将它们合并，并在下一状态 s'上运行 K-means。由此得到 K 个互不相交的簇 $\{\mathcal{N}^n\}_{n=1}^{K}$，其中每个簇 $\mathcal{N}^n$ 都包含下一状态相似的源域和目标域状态转移。然后，我们将 $\mathcal{N}^n$ 中的每个 s' 与其对应的 (s, a)配对，形成 $\mathcal{N}^n = \{(z_i, s_i')\}_{i=1}^{N_0^n+N_1^n}$，其中 N_0^n和 N_1^n分别是 $\mathcal{N}^n$中源样本和目标域样本的数量。我们在每个簇内分配二元标签：目标域样本 z_i^{tar}标记为类别 1，源样本 z_i^{src}标记为类别 0。目标是使用以下交叉熵损失训练一个簇特定分类器$D_n(z) = P(y = 1|z)$

$$L = -\mathbb{E}_{z\sim P(z|s'_{\text{tar}})}[\log D_n(z)] - \mathbb{E}_{z\sim P(z|s'_{\text{src}})}[\log(1 - D_n(z))]. \tag{4.1}$$

根据贝叶斯法则，分类器的输出满足

$$D_n(z) = \frac{P(z|y = 1)P(y = 1)}{P(z|y = 1)P(y = 1) + P(z|y = 0)P(y = 0)}. \tag{4.2}$$

在每个簇内，下一状态 s' 在构造上是相似的。因此，我们假设该表示z 主要反映源域和目标域的标签信息，即 $P(z|y = 1) \approx P(z|s'_{\text{tar}})$ 以及$P(z|y = 0) \approx P(z|s'_{\text{src}})$。定义密度比为 $w(z) = P(z|s'_{\text{tar}})/P(z|s'_{\text{src}})$。将此近似代入公式（4.2），可得

$$D_n(z) = \frac{Aw_n(z)}{Aw_n(z) + (1 - A)} \implies \hat{w}_n(z) = \frac{(1 - A)D_n(z)}{A(1 - D_n(z))},$$

其中，$A = P(y = 1) = N_1^n/(N_1^n + N_0^n)$，且 $1 - A = P(y = 0) = N_0^n/(N_1^n + N_0^n)$。由于 KL 散度可以写为 $D_{\text{KL}}(P(z|s'_{\text{src}})\|P(z|s'_{\text{tar}})) = \mathbb{E}_{z\sim P(z|s'_{\text{src}})}[\log 1/w(z)]$,，因此我们可以使用源样本对其进行估计：

$$\hat{D}_{\text{KL}}^n = \frac{1}{N_0^n}\sum_{i=1}^{N_0^n} \log \frac{1}{\hat{w}_n(z_i^{\text{src}})}, \tag{4.3}$$

ˆ其中 D_{KL}^n 是使用簇特定分类器和 $\log 1/\hat{}w_n(z_i^{\text{src}}) =: d_i$ 得到的 KL 散度估计簇；是 KL 散度的点估计。

4.3 策略优化

现在我们介绍所提出方法中的策略优化模块。在离线强化学习中，一个关键挑战是在部署时防止出现分布外（OOD）动作。优势加权方法（Peng 等人，2019）通过根据估计的优势对行为克隆进行加权，鼓励策略模仿值更高的数据集动作，从而缓解这一问题。然而，在异动力学设置中，直接模仿混合数据集，尤其是模仿在不匹配动力学下采集的源转移，可能会降低目标域的性能。为解决这一问题，我们（i）使用 KL 估计值筛选源数据集；（ii）使用（归一化的）逐点 KL 估计值自适应地重新加权学习。直观而言，估计的 KL 散度较小的源转移与目标动力学更一致，因此在优化过程中应具有更大的影响力。我们采用以下价值函数目标：

$$\mathcal{L}_Q = \mathbb{E}_{D_{tar}}\left[\left(Q_\theta - \mathcal{T}Q_\theta\right)^2\right] + \mathbb{E}_{(s,a,s',d)\sim\tilde{D}_{src}}\left[\exp\left(-\alpha\cdot\hat{d}_i\right)\left(Q_\theta - \mathcal{T}Q_\theta\right)^2\right], \tag{4.4}$$

其中， α 是一个超参数，$\hat{d}_i = \frac{d_i - \text{最大值}_i d_i}{\text{最大值}_i d_i - \text{最小值}_i d_i}$ 表示 $i \in \{1, \ldots, |\tilde{D}_{src}|\}$ 的归一化逐点KL估计值。经过筛选后，保留的源样本预计比被丢弃的样本更符合目标域。由于目标数据集较小（且可能含有噪声），适当提高筛选后源数据的聚合权重，可以利用数量充足的高质量转移来提升数据效率。

除了基于KL的重加权外，我们还引入了策略正则化项，以鼓励学习得到的策略保持在目标数据集的支持集附近。遵循 Lyu 等人 (2025年) 的做法，我们训练一个条件变分自编码器（条件变分自编码器），记为 $\hat{\pi}_b^{\text{tar}}(a|s)$，以对目标行为策略进行建模。随后，策略按照以下方式进行优化：

$$\mathcal{L}_\pi^{\text{reg}} = \mathcal{L}_\pi - \lambda \mathbb{E}_{s \sim \tilde{D}_{\text{src}} \cup D_{\text{tar}}} \left[\log \hat{\pi}_b^{\text{tar}}\big(\pi(\cdot|s)|s\big) \right], \tag{4.5}$$

其中， $\mathcal{L}_\pi$ 是底层离线强化学习算法的策略损失， λ 控制策略正则化的强度。

算法 1 局部化动力学感知域适应（LoDADA）

1：输入：源数据 D_{src}**，目标数据** D_{tar}**2：初始化：策略**π_ϕ**、价值函数** Q_θ**、批大小** B**、簇数** K**、直径阈值** δ**、评论器重要性比率** α**、正则化权重** λ**、过滤比率** ξ_1**、** ξ_2**、** ξ_3**3：基于** s'**，将** $D_{tar} \cup D_{src}$ **局部化为** $\{\mathcal{N}^i\}_{i=1}^K$ **4：对于 i=1、 2... 、 K，执行：5：基于公式 (4.1) 训练分类器** $D_n(z)$ **^ 6：根据公式 (4.3) 计算 KL 散度** D_{KL}^n **7：结束循环 8：根据 ^** D_{KL}^n **对** $\{\mathcal{N}^i\}_{i=1}^K$ **排序，并按簇选择性过滤，以获得** $\tilde{D}_{src}$**9：对每个梯度步执行：~10：从** D_{src}**中采样大小为** $N/2$ **的小批量** $b_{\text{src}} := \{(s, a, r, s')\}$ **11：从** D_{tar}**中采样大小为** $N/2$ **的小批量** $b_{\text{tar}} := \{(s, a, r, s')\}$ **12：根据公式 (4.4) 优化价值函数** Q_θ **13：根据公式 (4.5) 优化策略**π_ϕ **14：结束循环**

形式上，我们在图1中展示了 LoDADA 的框架，并在算法1中给出了算法。实现和技术细节见 C.6 节。我们使用 IQL（Kostrikov 等人，2021）作为骨干算法。

5 实验

本节通过在异动力学强化学习的 ODRL基准（Lyu 等人，2024b）上开展广泛实验，评估 LoDADA 的性能。我们首先在多种类型和偏移级别的动力学偏移条件下，进行 MuJoCo 运动任务（HalfCheetah、Ant、Walker2d 和 Hopper）实验。此外，我们还在 AntMaze 的导航任务和 Adroit 的操作任务上评估 LoDADA，以展示其在标准运动基准之外的有效性。最后，我们开展消融研究，以更好地理解各个组件的作用以及对关键超参数的敏感性。我们还进行了实验，将 LoDADA 的运行时复杂度与其他数据过滤策略进行比较，以展示本算法的计算效率。关于 Adroit 任务的结果和运行时复杂度比较，详见附录 D.3和D.4。

表1：在中等级别离线数据集上，MuJoCo 任务（HalfCheetah、Ant、Walker2D、Hopper）在重力和摩擦偏移级别为 0.1、0.5、2.0、5.0 时的性能对比。源域保持不变；目标域发生偏移。我们报告归一化目标域得分（五个随机种子的均值和标准差）。最佳和次佳得分分别以绿色和蓝色标出。

Env	偏移级别	BOSA	IQL	DARA	IGDF	OTDF	Ours
HalfCheetah 重力	0.1	9.31±1.94	12.90±1.01	9.62±4.27	7.65±6.63	12.17±3.58	**14.09±3.83**
	0.5	39.92 ±5.63	43.68±2.96	40.87±5.46	39.73±4.62	37.12±4.88	**45.00±2.95**
	2.0	28.47 ±1.23	**31.12±0.3**	30.49±0.54	30.16±1.5	25.38±3.12	28.02±1.67
	5.0	11.27±1.77	20.87±21.45	32.42±18.54	30.08±34.02	24.01±17.88	**71.11±1.63**
HalfCheetah摩擦	0.1	12.53±3.61	23.69±16.46	**26.39±11.35**	20.56±11.22	10.31±1.68	9.97±0.47
	0.5	63.65±9.14	68.26±1.42	68.41±3.19	66.30±1.79	66.32±0.46	**68.51±0.44**
	2.0	45.53±2.01	42.68±2.64	46.17±1.72	46.22±1.27	45.84±0.34	**48.19±0.75**
	5.0	38.26±10.82	48.97±4.68	45.31±9.83	**54.75±5.6**	38.36±9.82	54.65±1.56
Ant重力	0.1	**25.58±2.21**	11.03±1.24	12.53±1.11	13.55±2.13	18.82±5.51	22.38±1.28
	0.5	17.75±2.48	10.53±1.06	9.08±0.88	11.34±1.89	16.77±3.13	**18.61±2.79**
	2.0	37.17±0.96	38.26±3.09	38.54±2.19	41.59±4.63	37.11±0.74	**45.82±1.33**
	5.0	33.48±1.70	31.51±1.97	32.51±1.71	34.44±2.77	34.58±0.25	**48.29±3.68**
Ant摩擦	0.1	**58.95±0.71**	55.12±0.24	55.56±0.46	54.88±0.60	53.38±0.42	53.76±1.37
	0.5	**59.72±3.57**	58.92±0.80	59.28±0.80	55.64±5.03	58.20±0.84	55.54±3.80
	2.0	20.18±3.79	17.54±2.47	19.84±3.20	19.63±4.15	32.33±3.45	**62.27±2.23**
	5.0	9.88±1.17	7.79±0.31	7.78±0.26	13.97±7.68	13.17±8.28	**28.62±3.43**
Walker2d 重力	0.1	18.75±12.02	20.12±5.74	16.04±7.60	24.95±16.19	63.77±6.43	**71.41±1.99**
	0.5	40.09±2.37	29.72±16.02	42.05±10.52	37.36±5.49	32.47±7.04	**44.65±1.49**
	2.0	8.91±2.28	32.20±1.05	25.69±10.67	34.89±6.73	37.23±2.16	**40.27±1.77**
	5.0	5.25±0.50	5.44±0.08	5.42±0.29	5.66±0.32	4.92±0.08	**8.55±3.85**
Walker2d摩擦	0.1	7.88±1.88	5.65±0.06	5.72±0.23	7.21±1.80	**30.40±7.54**	20.21±7.90
	0.5	53.42±20.06	67.91±7.03	68.92±11.43	70.83±4.22	69.06±9.44	**81.41±6.47**
	2.0	39.06±17.36	**72.91±0.37**	65.40±7.13	62.55±10.36	59.93±14.53	62.01±4.36
	5.0	5.89±3.42	5.38±0.09	5.28±0.19	5.21±0.36	14.98±9.15	**15.93±3.09**
Hopper Gravity	0.1	27.82±13.41	23.4±11.62	13.10±0.98	13.89±4.70	33.41±1.92	**37.68±6.32**
	0.5	17.74±7.93	17.78±6.85	14.68±5.24	19.81±9.56	35.22±5.94	**43.72±9.72**
	2.0	13.18±2.10	14.71±1.67	14.94±1.08	14.71±0.87	**19.82±1.77**	16.99±3.74
	5.0	7.97±0.61	7.78±0.37	7.81±0.48	8.03±0.50	8.78±0.03	**9.28±0.26**
料斗摩擦	0.1	25.55±2.69	**26.13±4.24**	24.16±4.50	24.02±5.00	24.76±1.21	**28.06±7.97**
	0.5	25.86±2.17	24.75±3.58	24.84±4.63	29.12±3.01	29.89±6.17	**33.57±1.31**
	2.0	10.32±0.06	10.15±0.03	10.15±0.06	10.35±0.60	**10.52±0.29**	10.44±0.13
	5.0	6.92±1.97	7.92±0.06	7.84±0.04	7.90±0.17	**8.07±0.04**	8.04±0.05
总计		826.06	894.82	886.84	916.98	1007.10	**1207.05**

5.1 实验设置

任务与环境。我们在 ODRL基准（Lyu 等人，2024b）上，在具有不同动力学偏移的 MuJoCo 环境中评估 LoDADA。每个环境都包含三种动力学偏移：重力、摩擦和形态变化。源数据集采集自 D4RL（Fu 等人，2020）中的原始 MuJoCo 环境，目标数据集则由相应动力学偏移环境中的行为策略采集。所有数据集均使用 SAC（Haarnoja 等人，2018）采集，该策略在原始环境或动力学偏移环境中训练至中等性能水平。我们使用 ODRL 提供的数据集，其中源数据集包含 100 万条状态转移，目标数据集包含 5000 条状态转移。为构建动力学偏移的目标环境，我们遵循 ODRL 的方法，修改 MuJoCo 模拟器中与重力、摩擦和形态相关的参数。对于重力，我们重新缩放全局重力的大小，同时保持其方向不变。对于摩擦，我们重新缩放摩擦参数。在这两种情况下，我们均考虑四个偏移级别 $\{0.1, 0.5, 2.0, 5.0\}$，其中原始重力或摩擦会乘以相应的因子。形态变化通过改变特定肢体或躯干的尺寸来构造

机器人。

除运动任务外，我们还遵循ODRL，在具有不同地图布局的AntMaze问题上进行评估。我们聚焦于中等规模的蚁迷宫设置，并使用6种不同的地图布局作为目标环境，记为 $\{1, 2, 3, 4, 5, 6\}$。我们使用*play* 和多样化数据集作为源数据集，二者均取自D4RL（Fu等人，2020），而目标数据集由直接在目标环境中训练的智能体采集，并由ODRL提供。多样化数据集由从随机位置出发并到达随机目标的轨迹组成，而play数据集包含从人工选定位置出发并到达固定人工选定目标的轨迹。因此，这两个源数据集表现出不同的状态覆盖。

最后，我们在Adroit基准的操作任务上评估LoDADA。我们考虑*Pen*和门环境。与MuJoCo实验类似，我们通过引入运动学变化和形态偏移来构建目标环境。对于运动学变化，我们限制手部关节的旋转范围，称为关节损坏。对于形态偏移，我们缩小手指的尺寸，称为手指收缩。我们在实验中考虑中等和困难两种偏移级别。源数据集取自D4RL的人类水平，而目标数据集来自ODRL。关于所有这些环境的更多细节见第C节。我们将Adroit结果留至第D.3节。

基线方法。 我们将 LoDADA 与多种强大的基线方法进行比较，包括 DARA（Eysenbach 等人，2021）、BOSA（Liu 等人，2024a）、IQL（Kostrikov 等人，2021）、IGDF（Wen 等人，2024）和 OTDF（Lyu 等人，2025年）。DARA 将根据源域与目标域动力学之间的差异学习得到的奖励增广应用于源数据集，以模拟目标环境中的轨迹。BOSA 通过两个目标施加支持集约束，以缓解分布外状态—动作对和分布外动力学问题，同时过滤不可靠的状态转移。IQL 是一种离线强化学习算法，它通过学习价值函数和优势加权策略来提升性能；在混合数据集上训练、且不进行显式领域适应的离线强化学习任务中，该算法表现出色（Wen 等人，2024）。IGDF 学习对比表示，以估计基于互信息的领域差距，进而过滤源转移。OTDF 是最新的先进方法，它使用最优传输匹配源域与目标域转移，并根据最优传输成本进行非参数过滤。所有方法的实现细节，包括网络架构和超参数，均见C 节。

5.2 MuJoCo 运动任务的结果

我们首先总结 LoDADA 和基线方法在各种动力学偏移下的 MuJoCo 运动任务表现，包括重力、摩擦、形态偏移和局部扰动。我们使用归一化分数 $\text{归一化分数} = \frac{\text{分数}-\text{随机分数}}{\text{专家分数}-\text{随机分数}} \times 100$进行评估，其中 随机分数 由随机策略取得，而 专家分数由经过训练、达到目标域专家水平的 SAC（Haarnoja 等人，2018）取得。

重力和摩擦偏移的结果。 如 表1所示，LoDADA 的总分为 1207.05，比排名第二的 OTDF（Lyu 等人，2025年）高出 19.9%，比骨干算法 IQL（Kostrikov 等人，2021）高出 34.9%。在 32 项评估任务中（8 个环境 $\times$ 4 偏移级别），LoDADA 在 32 项任务中的 27 项取得最佳（21 项）或次佳（6 项）表现，并在其余 5 项任务中保持竞争力。 **这些结果表明，LoDADA 通过有效利用局部过滤提升了目标域性能。**

LoDADA 中的策略优化模块基于 IQL 构建。与原始 IQL 相比，我们引入了：(i) 评论家学习中的自适应加权；(ii) 额外的基于 CVAE 的行为克隆正则项。如表1所示，这些改进使其在 32 个任务中的 27 个任务上显著优于 IQL。**这些观察结果表明，将局部过滤与目标感知正则化相结合能够提升适应能力**

表2：在中等水平离线数据集上，MuJoCo 任务（Halfcheetah、Ant、Walker2d、Hopper）的性能比较。目标域保持不变；通过在不同区域添加局部扰动，对源数据集施加不同偏移级别。我们报告归一化目标域得分（均值 ´s 个随机种子的标准差）。最佳和次佳得分分别以绿色和蓝色突出显示。

Env	# 组	偏移级别	BOSA	DARA	IQL	IGDF	OTDF	Ours
HalfCheetah	3	0.1-0.5-2.0	27.36±2.22	19.86±4.42	17.57±3.41	14.59±3.13	29.26±3.10	**34.97±0.85**
		0.2-1.0-4.0	25.28±3.10	16.94±3.31	16.07±3.45	13.16±4.83	24.21±2.82	**28.36±2.76**
		0.3-1.5-6.0	20.75±2.66	6.59±0.45	10.73±2.88	8.73±2.31	22.84±0.54	**27.24±4.37**
	5	0.1-0.2-0.3-0.4-0.5	19.30±2.38	12.43±1.95	10.83±1.05	11.27±2.35	28.99±1.85	**34.28±2.91**
		0.2-0.4-0.6-0.8-1.0	23.59±5.81	13.18±2.09	14.77±2.41	11.43±3.96	21.41±3.95	**25.73±5.38**
		0.3-0.6-0.9-1.2-1.5	**23.95±2.23**	7.10±0.97	8.98±2.61	7.37±3.03	11.53±2.04	18.74±1.55
Walker2d	3	0.1-0.5-2.0	5.28±0.54	17.85±10.67	36.44±15.19	14.02±10.20	34.52±11.33	**37.45±18.71**
		0.2-1.0-4.0	6.42±1.10	19.79±3.58	15.29±9.44	14.62±8.16	20.66±7.06	**31.10±18.00**
		0.3-1.5-6.0	7.16±2.91	9.86±4.44	12.28±8.14	11.14±3.69	16.78±9.39	**18.69±10.51**
	5	0.1-0.2-0.3-0.4-0.5	7.17±1.48	8.10±3.46	6.96±2.01	6.99±2.77	33.87±16.17	**35.84±16.01**
		0.2-0.4-0.6-0.8-1.0	6.51±4.60	13.90±9.23	12.07±7.55	9.37±2.01	13.87±6.95	**24.92±5.96**
		0.3-0.6-0.9-1.2-1.5	6.43±2.53	8.65±3.18	7.91±3.80	8.99±5.44	**18.52±15.46**	12.98±3.88
Hopper	3	0.1-0.5-2.0	7.84±2.84	8.70±6.67	8.14±6.31	13.08±10.44	33.96±4.57	**40.77±5.04**
		0.2-1.0-4.0	11.91±2.60	9.61±5.56	11.97±5.99	8.35±3.56	32.58±6.89	**37.51±9.16**
		0.3-1.5-6.0	11.85±3.19	8.11±3.54	10.17±5.50	6.59±2.59	31.77±3.04	**33.41±2.72**
	5	0.1-0.2-0.3-0.4-0.5	10.42±3.55	6.37±1.67	6.92±1.85	15.81±9.78	22.25±8.74	**40.30±4.61**
		0.2-0.4-0.6-0.8-1.0	11.63±1.87	14.07±5.65	9.82±4.29	11.24±7.86	21.42±9.11	**36.06±4.59**
		0.3-0.6-0.9-1.2-1.5	13.97±5.15	11.14±4.01	12.01±2.06	15.63±10.74	29.95±5.01	**35.02±5.70**
Ant	3	0.1-0.5-2.0	55.51±10.37	69.67±8.59	69.65±10.93	63.79±9.55	70.30±5.24	**85.28±6.52**
		0.2-1.0-4.0	31.75±10.39	54.02±5.97	59.71±7.84	55.91±7.68	54.55±6.35	**61.51±6.89**
		0.3-1.5-6.0	14.41±4.68	67.13±5.07	**69.03±9.18**	50.24±7.41	34.47±4.88	53.54±7.07
	5	0.1-0.2-0.3-0.4-0.5	44.36±8.50	75.03±7.08	66.50±8.95	63.40±10.26	58.51±10.29	**78.96±7.34**
		0.2-0.4-0.6-0.8-1.0	30.26±11.22	56.84±5.10	**67.07±4.40**	55.65±13.74	47.99±8.48	61.92±10.42
		0.3-0.6-0.9-1.2-1.5	13.14±5.58	**51.81±6.79**	46.98±9.14	51.36±8.39	15.23±4.79	48.51±9.77
总计			436.25	586.77	607.87	542.73	729.44	**943.09**

到目标域。

形态偏移结果。对于存在形态变化的 MuJoCo 运动任务，结果及详细分析请参见D.1节。关键结论是，LoDADA 在 16 个任务中的 13 个任务上取得了最佳得分，相比第二佳方法提升了 32.7%。

5.3 局部扰动实验

在此实验中，我们评估了局部过滤方法在局部扰动动力学下的表现。与之前采用全局偏移的设置（例如，重力或摩擦的变化会均匀影响状态转移）不同，此处的动力学会随状态空间中的不同区域而变化。这种设置更具挑战性，因为分布不匹配具有异质性，无法通过依赖全局假设的方法（例如 DARA（Eysenbach 等人，2021）和 IGDF（Wen 等人，2024）来有效解决。

我们保持目标域不变，仅扰动源数据集以模拟局部变化。策略在扰动后的源数据集上训练，并在未改变的目标域中评估。对于目标数据集，我们从 D4RL（Fu 等人，2020）中随机采样 5,000 条状态转移，而不引入任何动力学偏移。对于源数据集，我们从原始 D4RL 数据集出发，使用 K-means 将下一状态空间划分为 15 个区域。随后我们将这些区域分为 M 组，并向状态变量注入高斯噪声，同一组内的所有样本共享相同的噪声方差。我们使用上文定义的归一化分数评估性能。我们考虑全部四个 MuJoCo 环境

（Ant、Hopper、HalfCheetah、Walker2d），并设置 $M \in \{3, 5\}$。当 $M = 3$ 时，噪声方差为$\{0.1, 0.5, 2.0\}$、$\{0.2, 1.0, 4.0\}$和 $\{0.3, 1.5, 6.0\}$。当 $M = 5$ 时，方差为 $\{0.1, 0.2, 0.3, 0.4, 0.5\}$、$\{0.2, 0.4, 0.6, 0.8, 1.0\}$和 $\{0.3, 0.6, 0.9, 1.2, 1.5\}$。

如表2所示，LoDADA 的总分为 943.09，比排名第二的 OTDF 高出 29.3%，比 IQL 高出 55.1%。在 24 项评估任务（4 个环境 × 6 偏移级别）中，LoDADA 在 24 项任务中的 22 项表现最佳（19 项）或排名第二（3 项）。相比之下，DARA 和 IGDF 在这种设置下往往表现不佳，因为它们依赖于全局失配假设。**这些结果表明，LoDADA 在局部扰动下尤为有效；与全局偏移相比，簇感知过滤在这种情况下更能发挥作用。**

表3：在地图中存在动力学偏移的多样化源数据集上，AntMaze 任务的性能比较。源域保持不变；目标域发生偏移。我们报告归一化目标域分数（五个随机种子的均值 ± 标准差），每项取 50 个回合的均值。最佳和次佳分数分别以绿色和蓝色突出显示。

Env	偏移级别	BOSA	DARA	IQL	IGDF	OTDF	Ours
Antmaze Medium	1	43.6 ± 11.2	53.6 ± 9.9	**76.5±8.4**	60.4 ± 9.1	51.6 ± 10.53	66.8 ± 4.6
	2	22.4 ± 5.5	33.6 ± 3.3	52.8 ± 7.9	45.2 ± 7.8	56.4 ± 4.34	**69.2±8.9**
	3	39.6 ± 11.5	50.8 ± 3.0	57.6 ± 7.7	65.2 ± 6.6	71.6±5.90	**73.2±7.3**
	4	40.0 ± 3.2	48.8 ± 8.3	58.4 ± 4.8	60.4 ± 6.2	52.4 ± 8.53	**65.2±5.9**
	5	21.2 ± 7.6	41.6 ± 12.7	66.4±8.3	56.8 ± 7.6	57.2 ± 7.16	**66.8±4.8**
	6	24.8 ± 14.0	59.2 ± 12.4	72.4 ± 10.7	73.6 ± 1.7	65.2 ± 9.86	**75.2±6.7**
总计		191.6	287.6	384.1	361.6	354.4	**416.4**

5.4 Antmaze 上的结果

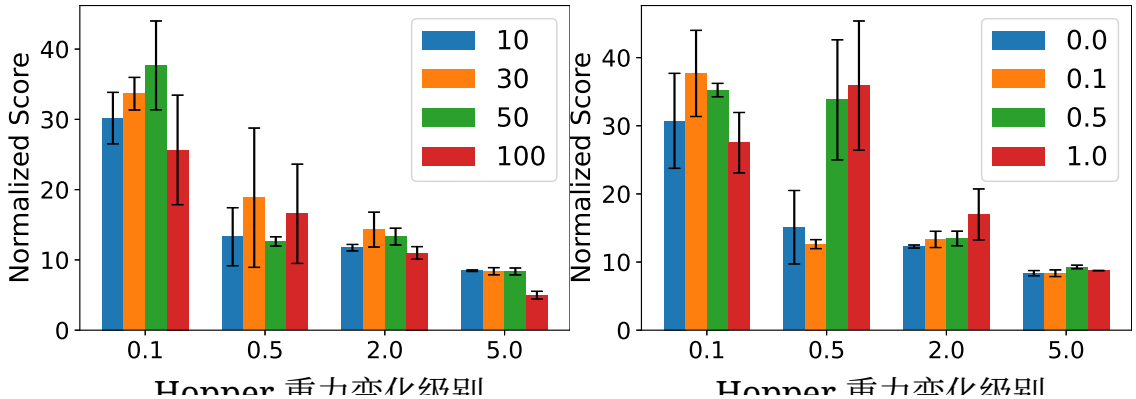
本节展示5.1节中所述导航任务的结果。我们使用多样化源数据集，在六个目标环境中基于中等规模蚁迷宫设置，将 LoDADA 与基线方法进行比较。结果汇总于表3。如表3所示，在采用多样化数据集的中等规模蚁迷宫基准上，LoDADA 的表现始终优于基线方法。总体而言，LoDADA 的总分最高，达到 416.4，比排名第二的方法高出 8.4%。在 6 张地图中，LoDADA 在 5 张地图上取得最佳表现，在剩余的一张地图上取得第二高分。**这些结果表明，LoDADA 的泛化能力不仅限于运动任务，还能在导航问题上取得出色表现。**

在结构化地图环境中，我们的局部过滤策略能够有效保留更接近目标域转移的源转移。具体而言，LoDADA 仅保留状态空间中地图位置与目标数据集中的转移相似的源转移。相比之下，DARA 和 IGDF 等以往方法往往无法保留这些有用的转移，从而限制了其有效性。采用 play 数据集的结果见D.2节 的表9。

5.5 消融研究

在本节中，我们进行若干消融研究，以深入了解 LoDADA 及关键超参数对运动任务的影响。作为示例，我们研究 Hopper 环境在重力偏移下的敏感性，偏移级别设为四档（0.1、0.5、2.0、5.0），重点考察簇数 K 和策略正则化权重 λ。

簇数 K。 K 决定了我们簇感知过滤机制所采用的状态空间划分的局部性程度。较小的 K 会形成较粗的簇，可能将动力学各异的转移合并在一起；而较大的 K 则可能导致划分过细，使每个簇中的样本不足。为研究这一权衡，我们对 $K \in \{10, 30, 50, 100\}$ 进行扫描，并在源数据集和目标数据集均为中等规模的 hopper-gravity 环境中运行 LoDADA。如 图 2a所示，当 K 过小（例如，10）或过大（例如，100）时，性能都会下降，而取中间值时结果始终更好。**这些结果表明，适度的局部性对于有效的簇感知过滤至关重要，能够兼顾动力学同质性和数据充足性**。 因此，在所有 $K = 30$ 中，我们将其设为50。



（a）针对 K 的消融研究。 （b）针对 λ 的消融研究。

图 2：参数敏感性研究。

策略正则化权重 λ。 λ 用于控制策略优化中正则化项的强度，从而促使学习得到的策略接近目标行为策略。较大的 λ 会施加更强的正则化，可能限制策略改进；而较小的 λ 则允许更激进的策略更新，但可能导致不稳定，或对不完善的过滤后数据过拟合。为研究这一权衡，我们对 $\lambda \in \{0, 0.1, 0.5, 1.0\}$ 进行扫描，并在具有中等水平源数据集和目标数据集的 hopper-gravity 环境中运行 LoDADA。 图 2b 所示结果表明，降低数据集正则化的作用（即， $\lambda = 0$）会导致性能不佳。LoDADA 对 λ 的选择较为敏感，而且不同任务的最佳 λ 可能有所不同（例如，hopper-gravity-0.1 对应 0.1，hopper-gravity-0.5 对应 1.0）。然而，如图 2b所示，**选择 $\lambda = 0.5$ 作为正则化权重，通常能在 MuJoCo 的不同设置中有效地用作默认策略正则化权重**。

6 结论

在本研究中，我们通过一种局部化动力学感知数据过滤方法，研究了动力学偏移离线强化学习。我们给出了理论性能界，该性能界促使我们从局部视角进行数据过滤，并引入数据集正则化项，以促使学习得到的策略与目标行为保持一致。基于这一洞见，我们提出了 LoDADA，并在全局和局部动力学偏移的多种类型和程度下开展了广泛实验。实验结果表明，在众多任务中，LoDADA 的性能均优于现有基线方法，尤其是在局部扰动动力学条件下取得了显著提升。未来的工作包括探索 LoDADA 在现实环境中的有效性，以及探索更具原则性的聚类机制，以更好地捕捉细粒度的状态-动作结构。

影响声明

本文旨在推动动力学偏移离线强化学习的研究。通过利用有限的目标域数据，在动力学失配条件下实现有效的策略学习，我们的方法有望减少在现实环境中进行大规模数据收集的需求。这对于机器人学、自主系统和医疗治疗等安全关键或成本高昂的场景尤其有益，因为在新动力学条件下收集数据可能不安全、成本高昂或难以实施。我们认为，本研究不会带来任何显著的负面社会影响。

A 主要理论结果的证明

A.1 定理 4.1

定理 *4.1*的证明。 该证明与 Lyu 等人 (2024a) 中定理 4.3 的证明类似。不过，我们通过计算源表示分布与目标表示分布之间的 KL 散度来推导这一结果，而不是反过来计算。这是因为，如 定理 4.2所示，我们提出的 LoDADA 方法需要刻画从筛选后的源数据中学习到的策略与目标最优策略之间的距离。由于 KL 散度具有非对称性，为使本文内容完整，我们在此重现该证明。

我们希望建立两个域中的表示偏差与这两个域之间的动力学差异之间的联系。为此，我们引入互信息的概念：

$$
\begin{aligned}
h\left(z; s'_{\text{src}}, s'_{\text{tar}}\right) &= I\left(z; s'_{\text{src}}\right) - I\left(z; s'_{\text{tar}}\right) \\
&= \int_{\mathcal{Z}}\int_{\mathcal{S}} P\left(z, s'_{\text{src}}\right) \log \frac{P\left(z, s'_{\text{src}}\right)}{P(z)P\left(s'_{\text{src}}\right)} dz ds'_{\text{src}} - \int_{\mathcal{Z}}\int_{\mathcal{S}} P\left(z, s'_{\text{tar}}\right) \log \frac{P\left(z, s'_{\text{tar}}\right)}{P(z)P\left(s'_{\text{tar}}\right)} dz ds'_{\text{tar}} \\
&= \int_{\mathcal{Z}}\int_{\mathcal{S}} P\left(z, s'_{\text{src}}\right) \log \frac{P\left(z|s'_{\text{src}}\right)}{P(z)} dz ds'_{\text{src}} - \int_{\mathcal{Z}}\int_{\mathcal{S}} P\left(z, s'_{\text{tar}}\right) \log \frac{P\left(z|s'_{\text{tar}}\right)}{P(z)} dz ds'_{\text{tar}} \\
&= \int_{\mathcal{Z}}\int_{\mathcal{S}}\int_{\mathcal{S}} P(z, s'_{\text{src}}, s'_{\text{tar}}) \log \frac{P(z|s'_{\text{src}})}{P(z)}\, dz\, ds'_{\text{src}}\, ds'_{\text{tar}} \\
&\quad - \int_{\mathcal{Z}}\int_{\mathcal{S}}\int_{\mathcal{S}} P(z, s'_{\text{tar}}, s'_{\text{src}}) \log \frac{P(z|s'_{\text{tar}})}{P(z)}\, dz\, ds'_{\text{tar}}\, ds'_{\text{src}} \\
&= \int_{\mathcal{Z}}\int_{\mathcal{S}}\int_{\mathcal{S}} P\left(z, s'_{\text{src}}, s'_{\text{tar}}\right) \log \frac{P\left(z|s'_{\text{src}}\right)}{P\left(z|s'_{\text{tar}}\right)} dz ds'_{\text{src}} ds'_{\text{tar}} \\
&= D_{\text{KL}}\left(P\left(z|s'_{\text{src}}\right) \| P\left(z|s'_{\text{tar}}\right)\right).
\end{aligned}
$$

请注意，KL 散度的定义已经涉及对 s'_{src} 和 s'_{tar} 的期望。虽然也可以写成 $\mathbb{E}_{s'_{\text{src}}, s'_{\text{tar}}}\left[D_{\text{KL}}\left(P(z|s'_{\text{src}}) \| P(z|s'_{\text{tar}})\right)\right]$，但这不应影响结果。同时，我们还有

$$
\begin{aligned}
h\left(z; s'_{\text{src}}, s'_{\text{tar}}\right) &= I\left(z; s'_{\text{src}}\right) - I\left(z; s'_{\text{tar}}\right) \\
&= \int_{\mathcal{Z}}\int_{\mathcal{S}} P\left(z, s'_{\text{src}}\right) \log \frac{P\left(z, s'_{\text{src}}\right)}{P(z)P\left(s'_{\text{src}}\right)}\, dz\, ds'_{\text{src}} - \int_{\mathcal{Z}}\int_{\mathcal{S}} P\left(z, s'_{\text{tar}}\right) \log \frac{P\left(z, s'_{\text{tar}}\right)}{P(z)P\left(s'_{\text{tar}}\right)}\, dz\, ds'_{\text{tar}} \\
&= \int_{\mathcal{Z}}\int_{\mathcal{S}} P\left(z, s'_{\text{src}}\right) \log \frac{P\left(s'_{\text{src}}|z\right)}{P\left(s'_{\text{src}}\right)}\, dz\, ds'_{\text{src}} - \int_{\mathcal{Z}}\int_{\mathcal{S}} P\left(z, s'_{\text{tar}}\right) \log \frac{P\left(s'_{\text{tar}}|z\right)}{P\left(s'_{\text{tar}}\right)}\, dz\, ds'_{\text{tar}}
\end{aligned}
$$

$$
\begin{aligned}
&= \int_{\mathcal{Z}}\int_{\mathcal{S}}\int_{\mathcal{S}} P\left(z, s'_{\text{src}}, s'_{\text{tar}}\right) \log \frac{P\left(s'_{\text{src}}|z\right)}{P\left(s'_{\text{src}}\right)}\, dz\, ds'_{\text{src}}\, ds'_{\text{tar}} \\
&\quad - \int_{\mathcal{Z}}\int_{\mathcal{S}}\int_{\mathcal{S}} P\left(z, s'_{\text{tar}}, s'_{\text{src}}\right) \log \frac{P\left(s'_{\text{tar}}|z\right)}{P\left(s'_{\text{tar}}\right)}\, dz\, ds'_{\text{tar}}\, ds'_{\text{src}} \\
&= \int_{\mathcal{Z}}\int_{\mathcal{S}}\int_{\mathcal{S}} P\left(z, s'_{\text{src}}, s'_{\text{tar}}\right) \log \frac{P\left(s'_{\text{src}}|z\right)}{P\left(s'_{\text{tar}}|z\right)}\, dz\, ds'_{\text{src}}\, ds'_{\text{tar}} \\
&\quad - \int_{\mathcal{S}} P\left(s'_{\text{src}}\right) \log P\left(s'_{\text{src}}\right)\, ds'_{\text{src}} + \int_{\mathcal{S}} P\left(s'_{\text{tar}}\right) \log P\left(s'_{\text{tar}}\right)\, ds'_{\text{tar}} \\
&= D_{\text{KL}}\left(P\left(s'_{\text{src}}|z\right) \parallel P\left(s'_{\text{tar}}|z\right)\right) + \mathbb{H}\left(s'_{\text{src}}\right) - \mathbb{H}\left(s'_{\text{tar}}\right).
\end{aligned}
$$

可以看出，所定义的函数也与动力学差异项$D_{\text{KL}}\left(P(s'_{\text{src}}|z) \parallel P(s'_{\text{tar}}|z)\right)$及两个熵项有关。不过，我们注意到源域和目标域都是指定且固定的，其状态分布也固定不变，这意味着熵项是常数。因此，最终我们得到

$$
\underbrace{D_{\text{KL}}\left(P\left(z|s'_{\text{src}}\right) \| P\left(z|s'_{\text{tar}}\right)\right)}_{\text{representation deviation}} = \underbrace{D_{\text{KL}}\left(P\left(s'_{\text{src}}|z\right) \| P\left(s'_{\text{tar}}|z\right)\right)}_{\text{dynamics deviation}} + \underbrace{\mathbb{H}\left(s'_{\text{src}}\right) - \mathbb{H}\left(s'_{\text{tar}}\right)}_{\text{constants}}.
$$

至此，证明完成。 □

A.2 定理4.2 的证明

证明定理 *4.2*。为简便起见，记为$\tilde{r} = \sum \gamma^t r\left(s_t,\ a_t\right)$。由于无法直接在不同域中与离线策略 π^* 和离线策略$\hat{}\pi^*_{\tilde{D}_{src}}$ 进行交互，我们引入 $\pi^*_{\text{src}} \in \Pi_{\text{no exploit}}$，它在筛选后的源域中最大化奖励。于是有

$$
\begin{aligned}
\mathbb{E}_{P_{\text{tar}},\pi^*}(\tilde{r}) - \mathbb{E}_{P_{\text{tar}},\hat{\pi}^*_{\tilde{D}_{src}}}(\tilde{r}) = &\underbrace{\mathbb{E}_{P_{\text{tar}},\pi^*}(\tilde{r}) - \mathbb{E}_{P_{\text{src}},\pi^*}(\tilde{r})}_{(a)} + \underbrace{\mathbb{E}_{P_{\text{src}},\pi^*}(\tilde{r}) - \mathbb{E}_{P_{\text{src}},\pi^*_{\text{src}}}(\tilde{r})}_{(b)} \\
&+ \underbrace{\mathbb{E}_{P_{\text{src}},\pi^*_{\text{src}}}(\tilde{r}) - \mathbb{E}_{P_{\text{tar}},\pi^*_{\tilde{D}_{tar}}}(\tilde{r})}_{(c)} + \underbrace{\mathbb{E}_{P_{\text{tar}},\pi^*_{\tilde{D}_{tar}}}(\tilde{r}) - \mathbb{E}_{P_{\text{tar}},\hat{\pi}^*_{\tilde{D}_{src}}}(\tilde{r})}_{(d)}.
\end{aligned}
$$

首先注意到，项 (b) 为 0，因为 π^*_{src} 是在筛选后的源域中的最优策略，该域应与目标域相同，这意味着 π^*_{src} 与 π^* 几乎相同，并且它们在相同域中应获得相同的奖励。根据$2R_{\max}\sqrt{1-e^{B-\epsilon}}$，项 (a) 的上界为假设 1。为了界定项 (c)，我们首先利用引理 1得到以下结果：

$$
\mathbb{E}_{P_{\text{src}},\pi^*_{\text{src}}}(\tilde{r}) - \mathbb{E}_{P_{\text{tar}},\pi^*}(\tilde{r}) \leq \mathbb{E}_{P_{\text{src}},\pi^*_{\text{src}}}(\tilde{r}) - \mathbb{E}_{P_{\text{tar}},\pi^*_{\text{src}}}(\tilde{r}) \leq 2R_{\max}\sqrt{1-e^{B-\epsilon}}.
$$

接下来我们可以使用这个 以及常规离线强化学习算法的性能界，以求得项（c）的界 ：

$$
\mathbb{E}_{P_{\text{src}},\pi^*_{\text{src}}}(\tilde{r}) \leq 2R_{\max}\sqrt{1-e^{B-\epsilon}} + \mathbb{E}_{P_{\text{tar}},\pi^*}(\tilde{r}) \leq 2R_{\max}\sqrt{1-e^{B-\epsilon}} + C + \mathbb{E}_{P_{\text{tar}},\pi^*_{\tilde{D}_{tar}}}(\tilde{r}).
$$

重新排列这些项能否帮助我们求出项 (c) 的上界？最后，我们将处理项 (d):

$$
(d) = \underbrace{\mathbb{E}_{P_{\text{tar}},\pi^*_{\tilde{D}_{tar}}}(\tilde{r}) - \mathbb{E}_{P_{\text{src}},\pi^*_{\tilde{D}_{tar}}}(\tilde{r})}_{I_1} + \underbrace{\mathbb{E}_{P_{\text{src}},\pi^*_{\tilde{D}_{tar}}}(\tilde{r}) - \mathbb{E}_{P_{\text{src}},\hat{\pi}^*_{\tilde{D}_{src}}}(\tilde{r})}_{I_2} + \underbrace{\mathbb{E}_{P_{\text{src}},\hat{\pi}^*_{\tilde{D}_{src}}}(\tilde{r}) - \mathbb{E}_{P_{\text{tar}},\hat{\pi}^*_{\tilde{D}_{src}}}(\tilde{r})}_{I_3}.
$$

我们通过 I_2 应用 引理 3:

$$
\begin{aligned}
I_2 &\leq \left|\mathbb{E}_{P_{\text{src}},\hat{\pi}^*_{\tilde{D}_{src}}}(\tilde{r}) - \mathbb{E}_{P_{\text{src}},\pi^*_{D_{tar}}}(\tilde{r})\right| \\
&= \left|\frac{1}{1-\gamma}\mathbb{E}_{\substack{\hat{\pi}^*_{\tilde{D}_{src}} \\ \rho_{\mathcal{M}_{\text{src}}}(s,a),s'\sim P_{\text{src}}}}\left[\mathbb{E}_{a'\sim\hat{\pi}^*_{\tilde{D}_{src}}}\left[Q^{\pi^*_{D_{tar}}}_{\mathcal{M}_{\text{src}}}(s',a')\right] - \mathbb{E}_{a'\sim\pi^*_{D_{tar}}}\left[Q^{\pi^*_{D_{tar}}}_{\mathcal{M}_{\text{src}}}(s',a')\right]\right]\right| \\
&= \frac{1}{1-\gamma}\mathbb{E}_{\substack{\hat{\pi}^*_{\tilde{D}_{src}} \\ \rho_{\mathcal{M}_{\text{src}}}(s,a),s'\sim P_{\text{src}}}}\left|\sum_{a'\in\mathcal{A}}\left(\hat{\pi}^*_{\tilde{D}_{src}}(a'|s') - \pi^*_{D_{tar}}(a'|s')\right)Q^{\pi^*_{D_{tar}}}_{\mathcal{M}_{\text{src}}}(s',a')\right| \\
&\leq \frac{R_{\max}}{1-\gamma}\mathbb{E}_{\substack{\hat{\pi}^*_{\tilde{D}_{src}} \\ \rho_{\mathcal{M}_{\text{src}}}(s,a),s'\sim P_{\text{src}}}}\left|\sum_{a'\in\mathcal{A}}\left(\hat{\pi}^*_{\tilde{D}_{src}}(a'|s') - \pi^*_{D_{tar}}(a'|s')\right)\right| \\
&= \frac{2R_{\max}}{1-\gamma}\mathbb{E}_{\substack{\hat{\pi}^*_{\tilde{D}_{src}} \\ \rho_{\mathcal{M}_{\text{src}}}(s,a),s'\sim P_{\text{src}}}}\left[D_{\text{TV}}\left(\hat{\pi}^*_{\tilde{D}_{src}}(\cdot|s') \,\|\, \pi^*_{D_{tar}}(\cdot|s')\right)\right].
\end{aligned}
$$

我们通过应用 I_1 来界定 引理 2:

$$
\begin{aligned}
I_1 &\leq \left|\mathbb{E}_{P_{\text{src}},\pi^*_{D_{tar}}}(\tilde{r}) - \mathbb{E}_{P_{\text{tar}},\pi^*_{D_{tar}}}(\tilde{r})\right| \\
&= \left|\frac{\gamma}{1-\gamma}\mathbb{E}_{\substack{\pi^*_{D_{tar}} \\ \rho_{\mathcal{M}_{\text{src}}}(s,a)}}\left[\mathbb{E}_{s'\sim P_{\text{src}}}\left[V^{\pi^*_{D_{tar}}}_{\mathcal{M}_{\text{tar}}}(s')\right] - \mathbb{E}_{s'\sim P_{\text{tar}}}\left[V^{\pi^*_{D_{tar}}}_{\mathcal{M}_{\text{tar}}}(s')\right]\right]\right| \\
&= \left|\frac{\gamma}{1-\gamma}\mathbb{E}_{\substack{\pi^*_{D_{tar}} \\ \rho_{\mathcal{M}_{\text{src}}}(s,a)}}\left[\int_{\mathcal{S}}\left(P_{\mathcal{M}_{\text{src}}}(s' \mid s,a) - P_{\mathcal{M}_{\text{tar}}}(s' \mid s,a)\right)V^{\pi^*_{D_{tar}}}_{\mathcal{M}_{\text{tar}}}(s')\,ds'\right]\right| \\
&\leq \frac{\gamma}{1-\gamma}\mathbb{E}_{\substack{\pi^*_{D_{tar}} \\ \rho_{\mathcal{M}_{\text{src}}}(s,a)}}\left[\left|\int_{\mathcal{S}}\left(P_{\mathcal{M}_{\text{src}}}(s' \mid s,a) - P_{\mathcal{M}_{\text{tar}}}(s' \mid s,a)\right)\right| \times \left|V^{\pi^*_{D_{tar}}}_{\mathcal{M}_{\text{tar}}}(s')\right| ds'\right] \\
&\leq \frac{2\gamma}{1-\gamma}R_{\max}\left[D_{\text{TV}}\left(P_{\mathcal{M}_{\text{src}}}(\cdot \mid s,a) \| P_{\mathcal{M}_{\text{tar}}}(\cdot \mid s,a)\right)\right] \\
&\leq \frac{2\gamma}{1-\gamma}R_{\max}\sqrt{1-e^{B-\epsilon}}.
\end{aligned}
$$

对于 I_3，同样可以得到相同的上界。将所有上界合并后，即可得到所需结果。 □

B 有用引理

引理 1. 设 R_{max} 为任意轨迹的最大（熵正则化）回报。则以下不等式成立:

$$
\left|\mathbb{E}_{\pi,P_{src}}\left[\sum\gamma^t r\left(s_t,a_t\right)\right] - \mathbb{E}_{\pi,P_{tar}}\left[\sum\gamma^t r\left(s_t,a_t\right)\right]\right| \leq 2R_{\max}\sqrt{1-e^{B-\epsilon}}.
$$

请注意，Eysenbach 等人 (2021) 的原始结果是在策略类 $\Pi_{\text{no exploit}}$中推导得出的，该策略类定义为动力学差异受到约束的策略集合。由于在我们的情形中，假设源域与目标域动力学之间的 KL 散度有界于 $\epsilon - B$，上述结果不再局限于策略类 $\Pi_{\text{no exploit}}$。

证明。该证明与 Eysenbach 等人 (2021) 中引理 B.2 的证明类似，唯一的区别是我们在最后一步使用 B-H 界来处理全变差距离，而不是使用 Pinsker 不等式。为方便读者参考，我们在此重述证明。我们知道 $z = f(s, a)$ 是单射，并应用 Hölder 不等式和 Bretagnolle-Huber 界（其效果不会差于 Pinsker 不等式）（Canonne，2022），从而得到所需结果：

$$\begin{aligned}\left|\mathbb{E}_{P_{\text{src}}}\left[\sum \gamma^t r\left(s_t, a_t\right)\right] - \mathbb{E}_{P_{\text{tar}}}\left[\sum \gamma^t r\left(s_t, a_t\right)\right]\right| &= \left|\sum_\tau \left(P_{\text{src}}(\tau) - P_{\text{tar}}(\tau)\right)\left(\sum \gamma^t r\left(s_t, a_t\right)\right)\right| \\ &\leq \left\|\sum \gamma^t r\left(s_t, a_t\right)\right\|_\infty \cdot \left\|P_{\text{src}}(\tau) - P_{\text{tar}}(\tau)\right\|_1 \\ &\leq \left(\max_\tau \sum \gamma^t r\left(s_t, a_t\right)\right) \cdot 2\sqrt{1 - e^{-D_{\text{KL}}(P_{\text{src}}(\tau) \| P_{\text{tar}}(\tau))}} \\ &\leq 2R_{\max}\sqrt{1 - e^{B-\epsilon}}.\end{aligned}$$

证明至此完成。 □

引理 2（望远镜引理，Luo 等人（2019）中的引理 4.3）。。记 $\mathcal{M}_1 = (\mathcal{S}, \mathcal{A}, P_1, r, \gamma)$ 和 $\mathcal{M}_2 = (\mathcal{S}, \mathcal{A}, P_2, r, \gamma)$ 表示两个仅转移动力学不同的 *MDP*。于是，对于任意策略 π，我们有

$$J_{\mathcal{M}_1}(\pi) - J_{\mathcal{M}_2}(\pi) = \frac{\gamma}{1-\gamma}\mathbb{E}_{\rho^\pi_{\mathcal{M}_1}(s,a)}\left[\mathbb{E}_{s'\sim P_1}\left[V^\pi_{\mathcal{M}_2}\left(s'\right)\right] - \mathbb{E}_{s'\sim P_2}\left[V^\pi_{\mathcal{M}_2}\left(s'\right)\right]\right].$$

引理 3. 记 $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$ 为底层 *MDP*。设有两个策略 π_1、π_2，则这两个策略在该 *MDP* 中的性能差异为：

$$J_{\mathcal{M}}\left(\pi_1\right) - J_{\mathcal{M}}\left(\pi_2\right) = \frac{1}{1-\gamma}\mathbb{E}_{\rho^{\pi_1}_{\mathcal{M}}(s,a), s'\sim P}\left[\mathbb{E}_{a'\sim\pi_1}\left[Q^{\pi_2}_{\mathcal{M}}\left(s', a'\right)\right] - \mathbb{E}_{a'\sim\pi_2}\left[Q^{\pi_2}_{\mathcal{M}}\left(s', a'\right)\right]\right].$$

证明。这是Lyu 等人（2024a）中的引理 B.3，请参阅其中的证明。 □

C 实验细节

本节将详细介绍实验设置和超参数，相关内容见第 5 节。

C.1 运动任务

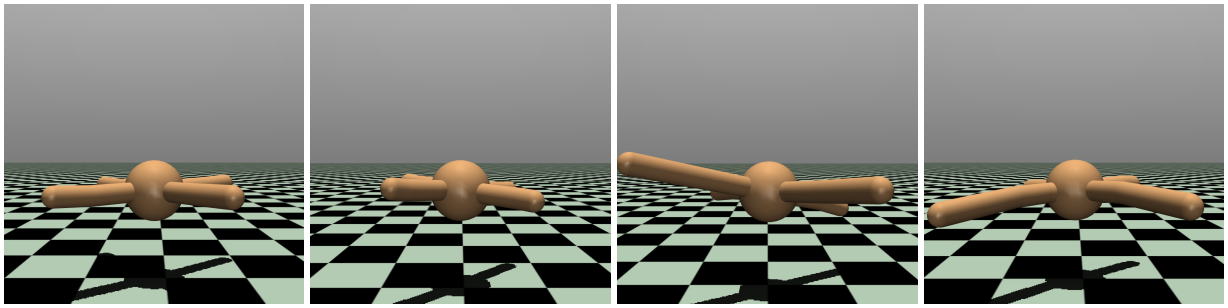
在本节中，我们将详细介绍如何修改 MuJoCo xml 文件来构建目标环境。在运动任务中，我们遵循 ODRL，设置了三种动力学偏移：重力、摩擦和形态偏移。在重力和摩擦偏移中，我们考虑偏移级别为$\{0.1, 0.5, 2.0, 5.0\}$。在形态偏移中，我们考虑中等和困难级别的偏移。

重力偏移。遵循 ODRL基准（Lyu 等人，2024b），我们通过编辑重力属性来修改环境的重力。例如，将目标域中 HalfCheetah 的重力修改为源域重力的 0.5 倍，代码如下。

```
# 重力<选项 重力="00 -4.905" 时间步长="0.01"/>
```

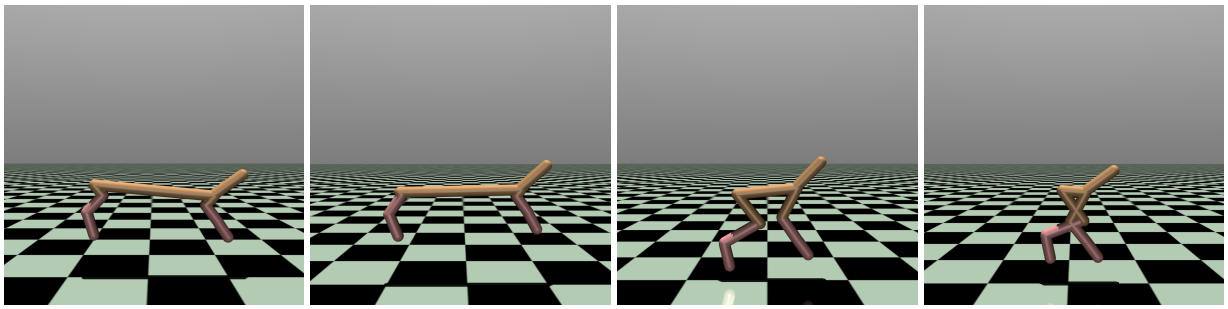
摩擦偏移。 摩擦偏移通过修改几何元素中的摩擦属性来实现。摩擦分量分别调整为源域中摩擦分量的 0.1、0.5、2.0、5.0 倍。

形态偏移。 形态偏移通过在 MuJoCo 中修改仿真机器人的特定肢体或躯干的尺寸来实现，且不改变状态空间和动作空间。可视化结果见 图3、图4、图5 和图6。



(a) 全腿（M） (b) 全腿（H） (c) 半腿（M） (d) 半腿 (H)

图3：Ant 形态变化环境的可视化。

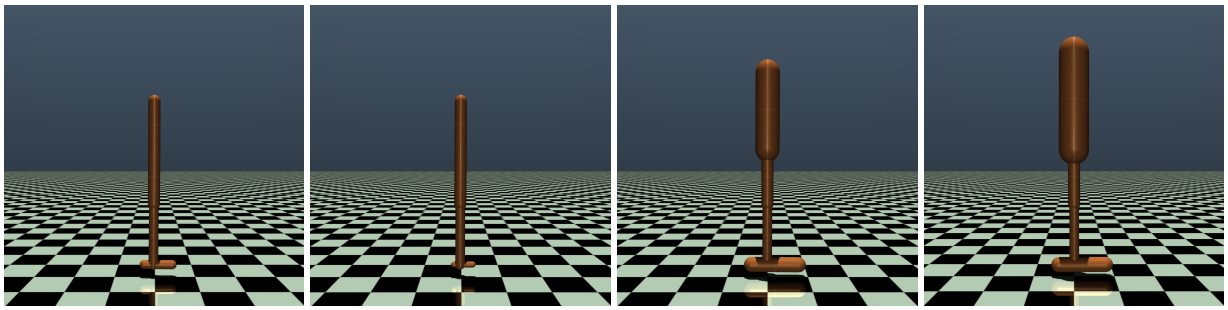


(a) 大腿 (M) (b) 大腿 (H) (c) 躯干 (M) (d) 躯干 (H)

图4：HalfCheetah形态变化环境的可视化。

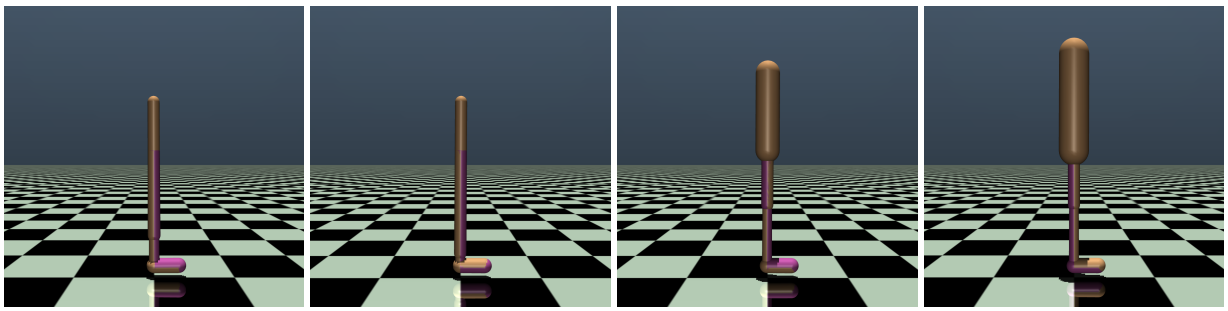
C.2 导航任务

在本节中，我们介绍对AntMaze地图的详细修改，以构建目标环境。在该设定中，我们遵循ODRL基准，并考虑中等尺寸以及六种不同的目标地图。



(a) 脚 (M) (b) 脚 (H) (c) 躯干 (M) (d) 躯干 (H)

图5：Hopper 形态变化环境的可视化。



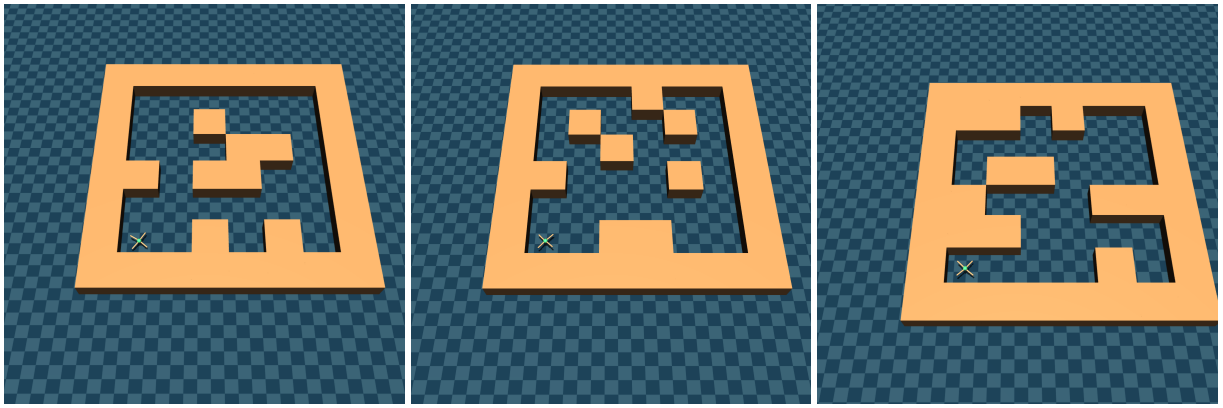
(a) 腿 (M) (b) 腿（H） (c) 躯干（M） (d) 躯干（H）

图6：Walker2d 形态变化环境的可视化。

地图可视化请参见图7。



(a) antmaze-medium-1 (b) antmaze-medium-2 (c) antmaze-medium-3



(d) antmaze-medium-4 (e) antmaze-medium-5 (f) antmaze-medium-6

图7：AntMaze地图结构的可视化。我们考虑具有6种不同地图布局的中等尺寸地图。

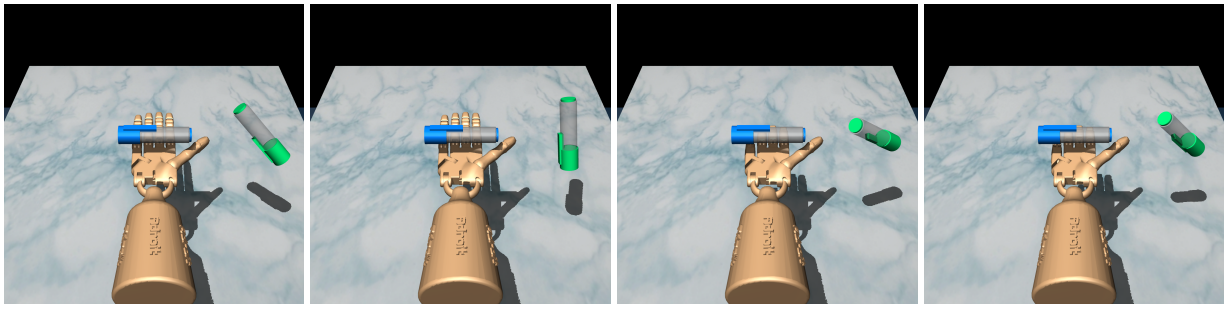
C.3 操作任务

本节将详细介绍对Adroit环境所做的修改。在此设置中，我们遵循ODRL基准，将*Pen* 和门作为任务。*Pen*任务是控制24自由度Shadow手机器人转动一支笔，而门任务则是打开一扇门。在此环境中，我们考虑两种类型的动力学偏移，分别称为运动学偏移和形态任务。我们只考虑中等和困难级别的偏移。可视化结果见图8和图9。

运动学偏移。Adroit机器人手的运动学偏移发生在食指和拇指的所有手指关节上。我们对躯干的活动范围施加限制。运动学偏移采用两个偏移级别，分别为中等和困难。

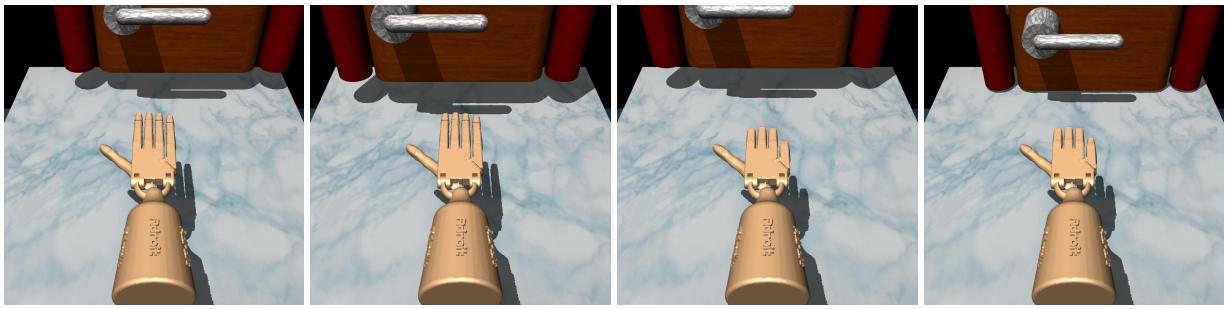
形态偏移。Adroit机器人手的形态偏移发生在手指上。我们将这些手指的尺寸缩小至中等和困难等级，分别乘以 0.25 和 0.125，得到手指的

尺寸。



(a) 关节损坏（M） (b) 关节损坏（H） (c) 手指收缩（M） (d) 手指收缩（H）

Fi图 8：Adroit 笔的形态与运动学偏移环境可视化。



(a) 关节损坏（M） (b) 关节损坏 (H) (c) 手指收缩 (M) (d) 手指收缩 (H)

Fi图 9：Adroit 门的形态与运动学偏移环境可视化 .

C.4 数据集

请注意，我们遵循 D4RL，采用归一化分数指标，以更好地刻画智能体在不同任务中的表现。目标域中的归一化分数定义如下：

$$\mathrm{NS} = \frac{J - J_r}{J_e - J_r} \times 100,$$

其中，J 表示智能体在目标域中获得的回报，J_r和 J_e 分别表示同一域中随机策略和专家策略获得的回报。

我们总结了 J_r 和 J_e 在不同动力学偏移情景下的参考分数，详见表 4、表 5和表 6。

C.5 基线算法的技术细节

本节介绍基线方法的实现细节，遵循 ODRL 基准（Lyu 等人，2024b）。此外，我们在表 7中列出了所有方法所用的超参数设置。

表 4：MuJoCo 数据集在不同动力学偏移情景下的参考最低分数和参考最高分数。这些参考分数用于计算目标域中的归一化性能。

任务名称	动力学偏移类型	参考最低分数 (J_r)	参考最高分数 (J_e)
HalfCheetah	重力 - 0.1	-280.18	2466.85
	重力 - 0.5	-280.18	9509.15
	重力 - 2.0	-280.18	9509.15
	重力 - 5.0	-280.18	3756.24
	摩擦 - 0.1	-280.18	41696.55
	摩擦 - 0.5	-280.18	7357.07
	摩擦 - 2.0	-280.18	11255.97
	摩擦 - 5.0	-280.18	10199.33
	局部	-280.18	12135.0
Hopper	重力 - 0.1	-26.34	3234.3
	重力 - 0.5	-26.34	3234.3
	重力 - 2.0	-26.34	3234.3
	重力 - 5.0	-26.34	3234.3
	摩擦 - 0.1	-26.34	3234.3
	摩擦 - 0.5	-26.34	3234.3
	摩擦 - 2.0	-26.34	3234.3
	摩擦 - 5.0	-26.34	3234.3
	局部	-20.272305	3234.3
Walker2d	重力 - 0.1	10.08	2074.90
	重力 - 0.5	10.08	5194.71
	重力 - 2.0	10.08	5056.45
	重力 - 5.0	10.08	3665.39
	摩擦 - 0.1	10.08	3360.18
	摩擦 - 0.5	10.08	4229.35
	摩擦 - 2.0	10.08	5180.04
	摩擦 - 5.0	10.08	4988.84
	局部	1.629008	4592.3
Ant	重力 - 0.1	-325.6	2782.09
	重力 - 0.5	-325.6	4317.07
	重力 - 2.0	-325.6	6705.12
	重力 - 5.0	-325.6	6226.89
	摩擦 - 0.1	-325.6	7938.96
	摩擦 - 0.5	-325.6	8301.34
	摩擦 - 2.0	-325.6	5167.38
	摩擦 - 5.0	-325.6	4545.02
	局部	-325.6	3879.7

表 5：各种动力学偏移情形下 AntMaze 中等数据集的参考最低分数和参考最高分数。这些参考分数用于计算目标域中的归一化性能。

任务名称	动力学偏移类型	参考最低分数（J_r）	参考最高分数（J_e）
AntMaze-medium	1	0.0	1.0
	2	0.0	1.0
	3	0.0	1.0
	4	0.0	1.0
	5	0.0	1.0
	6	0.0	1.0

表 6：Adroit 数据集在各种动力学偏移情景下的参考最低分数和参考最高分数。这些参考分数用于计算目标域中的归一化性能。

任务名称	动力学偏移类型	参考最低分数（J_r）	参考最高分数（J_e）
Pen	broken-joint-medium	-12.17	6408.38
	关节损坏-hard	-12.17	6408.38
	手指收缩-medium	-12.17	6408.38
	手指收缩-hard	-12.17	6408.38
Door	broken-joint-medium	-52.34	2880.57
	关节损坏-hard	-52.34	2880.57
	手指收缩-medium	-52.34	2880.57
	手指收缩-硬	-52.34	2880.57

BOSA（**Liu 等人，2024a**）。BOSA 通过支持策略优化处理分布外状态–动作对，并通过数据筛选实现支持价值优化，以解决分布外动力学问题。具体而言，策略通过最大化以下目标函数进行更新：

$$\mathcal{L}_{\text{actor}} = \mathbb{E}_{s\sim D_{\text{src}}\cup D_{\text{trg}},\, a\sim\pi_\phi(s)}\big[Q(s,a)\big], \quad \text{s.t.} \quad \mathbb{E}_{s\sim D_{\text{src}}\cup D_{\text{trg}}}\big[\hat{\pi}_{\theta_{\text{offline}}}(\pi_\phi(s)\mid s)\big] > \epsilon.$$

其中，ϵ 是阈值，$\hat{\pi}_{\theta_{\text{offline}}}$ 是针对合并的离线数据集学习得到的策略。价值评论家通过以下内容更新

$$\begin{aligned}\mathcal{L}_{\text{critic}} = {} & \mathbb{E}_{(s,a)\sim D_{\text{src}}}\big[Q(s,a)\big] \\ & + \mathbb{E}_{(s,a,r,s')\sim D_{\text{src}}\cup D_{\text{trg}},\, a'\sim\pi_\phi(\cdot|s')}\Big[I\left(\hat{p}_{\text{trg}}(s'\mid s,a) > \epsilon'\right)\left(Q_{\theta_i}(s,a)-y\right)^2\Big],\end{aligned}$$

其中，$I(\cdot)$ 是指示函数，$\hat{p}_{\text{trg}}(s'\mid s,a) = \text{argmax}\, \mathbb{E}_{(s,a,s')\sim D_{\text{trg}}}\big[\log \hat{p}_{\text{trg}}(s'\mid s,a)\big]$ 是估计的目标域动力学，ϵ' 是阈值。上述目标被转换为松弛的拉格朗日形式以进行优化。行为策略和动力学模型均由条件变分自编码器建模。我们按照原始论文中的说明实现 BOSA。在实际应用中，BOSA 通过从源域数据集和目标域数据集中抽取样本，训练 500K 个梯度步数。

IQL（Kostrikov 等人，2021）。IQL 通过最小化以下期望分位数回归损失函数，同时学习状态价值函数和状态-动作价值函数

$$\mathcal{L}_V = \mathbb{E}_{(s,a)\sim D_{\text{src}}\cup D_{\text{trg}}}\left[L_2^\tau\big(Q_\theta(s,a) - V_\psi(s)\big)\right],$$

其中，$L_2^\tau(u) = |\tau - I(u<0)|\|u\|^2$，$I(\cdot)$ 是指示函数，θ 是目标网络参数。随后，状态–动作价值函数按以下方式更新：

$$\mathcal{L}_Q = \mathbb{E}_{(s,a,r,s')\sim D_{\text{src}}\cup D_{\text{trg}}}\left[\big(r(s,a) + \gamma V_\psi(s') - Q_\theta(s,a)\big)^2\right].$$

优势函数为 $A(s,\ a) = Q(s,\ a) - V(s)$。策略通过优势加权行为克隆进行优化

$$\mathcal{L}_{\text{actor}} = \mathbb{E}_{(s,a)\sim D_{\text{src}}\cup D_{\text{trg}}}\left[\exp\big(\beta\cdot A(s,a)\big)\log\pi_\phi(a\mid s)\right],$$

其中，β 是逆温度系数。我们遵循 IQL 的官方代码库实现 IQL，并在来自两个域的离线数据集上训练 IQL 智能体，共进行 500K 个梯度步数。

DARA（Eysenbach 等人，2021）。DARA 是 DARC 的离线版本（Eysenbach 等人，2021）。它训练两个域分类器 $q_{\theta_{\text{SAS}}}$（目标域 | s_t，a_t，s_{t+1}），$q_{\theta_{\text{SA}}}$（目标域 | s_t，a_t），其目标函数如下

$$\begin{aligned}\mathcal{L}(\theta_{\text{SAS}}) &= \mathbb{E}_{D_{\text{tar}}}\left[\log q_{\theta_{\text{SAS}}}(\text{target}\mid s_t,a_t,s_{t+1})\right] + \mathbb{E}_{D_{\text{src}}}\left[\log\big(1 - q_{\theta_{\text{SAS}}}(\text{target}\mid s_t,a_t,s_{t+1})\big)\right],\\ \mathcal{L}(\theta_{\text{SA}}) &= \mathbb{E}_{D_{\text{tar}}}\left[\log q_{\theta_{\text{SA}}}(\text{target}\mid s_t,a_t)\right] + \mathbb{E}_{D_{\text{src}}}\left[\log\big(1 - q_{\theta_{\text{SA}}}(\text{target}\mid s_t,a_t)\big)\right],\end{aligned}$$

用于估计动力学差异

$$\log\frac{P_{\mathcal{M}_{\text{tar}}}(s_{t+1}\mid s_t,a_t)}{P_{\mathcal{M}_{\text{src}}}(s_{t+1}\mid s_t,a_t)}$$

源域与目标域之间的动力学差异。DARA 利用训练后的分类器近似这一项，并提出按以下方式修改源领域奖励

$$\begin{aligned}\hat{r}_{\text{DARA}} &= r - \lambda\times\delta r,\\ \delta r(s_t,a_t) &= -\log\frac{q_{\theta_{\text{SAS}}}(\text{target}\mid s_t,a_t,s_{t+1})\,q_{\theta_{\text{SA}}}(\text{source}\mid s_t,a_t)}{q_{\theta_{\text{SAS}}}(\text{source}\mid s_t,a_t,s_{t+1})\,q_{\theta_{\text{SA}}}(\text{target}\mid s_t,a_t)},\end{aligned}$$

其中，λ 是一个重要的超参数，用于控制奖励惩罚的强度，默认设为 0.1。随后，DARA 使用来自两个域的数据训练 500K 梯度步数。我们遵循原始论文实现 DARA，并将奖励惩罚项裁剪到 $[-10,10]$。我们使用 IQL 作为 DARA 的基础算法。

IGDF（Wen 等人，2024）。IGDF 通过对比学习捕捉源域与目标域之间的动力学差异。它使用来自目标域的 $h(\cdot)$ 训练评分函数 $(s,\ a,\ s'_{\text{tar}}) \sim D_{\text{tar}}$ 作为正样本，并将混合转移 $(s,\ a,\ s'_{\text{src}})$ 作为负样本，其中 $(s,\ a) \sim D_{\text{tar}}$，$s'_{\text{src}} \sim D_{\text{src}}$。评分函数通过以下对比学习目标进行优化：

$$\mathcal{L}_{\text{contrastive}} = -\mathbb{E}_{(s,a,s'_{\text{tar}})}\mathbb{E}_{S'-}\left[\log\frac{h(s,a,s'_{\text{tar}})}{\sum_{s'\sim S'-\cup s'_{\text{tar}}}h(s,a,s')}\right],$$

其中， S'^{-} 是负样本中下一状态的集合。具体而言，IGDF 采用两个神经网络 $\phi(s,\ a)$ $\psi(s')$ 来学习状态-动作对和状态的表征，并通过对它们进行线性参数化来近似评分函数：

$$h(s,a,s') = \exp\left(\phi(s,a)^\top \psi(s')\right).$$

基于测得的评分函数，IGDF 提议在训练价值函数时筛除源域数据：

$$\mathcal{L}_{\text{critic}} = \frac{1}{2}\,\mathbb{E}_{D_{\text{tar}}}\left[(Q_\theta - \mathcal{T}Q_\theta)^2\right] + \frac{1}{2}\,\alpha \cdot h(s,a,s')\,\mathbb{E}_{(s,a,s')\sim D_{\text{src}}}\left[\mathbf{1}\left(h(s,a,s') > h_{\xi\%}\right)(Q_\theta - \mathcal{T}Q_\theta)^2\right],$$

其中， α 是用于加权源域数据时序差分误差的重要性系数， ξ 是与 OTDF 中类似的数据选择比例。我们使用其官方代码库运行 IGDF，并采用 IQL 作为其骨干算法。

OTDF（Lyu 等人，2025年）。OTDF 通过使用最优传输对齐源域与目标域的转移动力学，实现异动力学离线策略适应。它首先使用

$$W(u,u') = \min_{\mu\in\mathcal{M}} \sum_{t=1}^{|D_{\text{src}}|}\sum_{t'=1}^{|D_{\text{tar}}|} C(u_t,u'_{t'})\,\mu_{t,t'},$$

给定的代价函数 C 求解。假设得到的最优耦合为 μ^*，则源域中的一个数据点与目标域数据集之间的偏差定义为：

$$d(u_t) = -\sum_{t'=1}^{|D_{\text{tar}}|} C(u_t,u'_{t'})\,\mu^*_{t,t'}, u_t = (s_t^{\text{src}}, a_t^{\text{src}}, (s'_{\text{src}})_t) \sim D_{\text{src}}.$$

计算并归一化 d 后，通过以下方式训练价值函数：

$$\mathcal{L}_Q = \mathbb{E}_{(s,a,s')\sim D_{\text{tar}}}\left[(Q_\theta - \mathcal{T}Q_\theta)^2\right] + \mathbb{E}_{(s,a,s')\sim D_{\text{src}}}\left[exp(\hat{d})\cdot\mathbf{1}(d > d_{\xi\%})\,(Q_\theta - \mathcal{T}Q_\theta)^2\right],$$

ˆ其中， d 是归一化后的 d ， $\mathbf{1}(\cdot)$ 是指示函数。随后，使用损失函数训练条件变分自编码器（CVAE）：

$$\mathcal{L}_{\text{CVAE}} = \mathbb{E}_{(s,a)\sim D_{\text{tar}},\, z\sim E_\nu(s,a)}\left[(a - D_\varsigma(s,z))^2\right] + D_{\text{KL}}\big(E_\nu(s,a)\,\|\,\mathcal{N}(0,I)\big),$$

其中， $E_\nu(s,\ a)$ 是编码器， $D_\varsigma(s,\ z)$ 是解码器，用于对目标动力学进行建模并生成多个潜在转移。最后，他们通过以下方式训练策略

$$\hat{\mathcal{L}}_\pi = \mathcal{L}_\pi - \beta\,\mathbb{E}_{s\sim D_{\text{src}}\cup D_{\text{tar}}}\left[\log\left(\sum_{i=1}^{M}\exp\left(\log\hat{\pi}^i_{\text{tar}}(\pi(\cdot|s)\mid s)\right)\right)\right],$$

其中，ˆ π^i_{tar} 表示在目标域中构建的第 i 个高斯行为策略。数据选择机制会筛选出其生成的状态转移与目标状态转移高度匹配的源样本。我们使用官方代码库运行 OTDF。

表 7：LoDADA 和基线方法的超参数。

方法	超参数	值
所有方法共用	演员网络	(256, 256)
	评论家网络	(256, 256)
	学习率	3×10^{-4}
	优化器	Adam
	折扣因子	0.99
	经验回放缓冲区大小	10^6
	非线性函数	ReLU
	目标更新率	5×10^{-3}
	源域批次大小	128
	目标域批次大小	128
DARA	温度系数	0.2
	最大对数标准差	2
	最小对数标准差	-20
	分类器网络	(256, 256)
	奖励惩罚系数 λ	0.1
BOSA	温度系数	0.2
	最大对数标准差	2
	最小对数标准差	-20
	策略正则化系数 λ_{policy}	0.1
	转移系数 $\lambda_{\text{transition}}$	0.1
	阈值参数 ϵ，ϵ'	log(0.01)
	价值权重 ω	0.1
	CVAE 集成规模	1（行为），5（动力学）
IGDF	表征维度	{16, 64}
	对比编码器网络	(256, 256)
	编码器预训练步数	7000
	重要性系数	1.0
	数据选择比例 ξ	75%
OTDF	CVAE训练步数	10000
	CVAE学习率	0.001
	采样潜变量数量 M	10
	高斯标准差	$\sqrt{0.1}$
	代价函数	余弦
	策略系数 β	0.5
	数据选择比例 ξ	80%
IQL	温度系数	0.2
	最大对数标准差	2
	最小对数标准差	-20
	逆温度 β	3.0
	期望分位数 τ	0.7
LoDADA	簇数 K	{30, 50}
	直径阈值	{1.5, 5.0}
	过滤比率（ξ_1，ξ_2，ξ_3）	(90%, 80%, 70%)
	输入噪声标准差 σ	1.0
	重要性权重 α	1.0

C.6 LoDADA 的技术细节

LoDADA 算法（局部化动力学感知域适应）总结见 算法 2，它利用聚类和基于散度的筛选来改进非动力学策略学习。首先，根据下一状态变量对目标域数据进行聚类，形成若干簇；然后，根据源样本与各簇质心的距离是否在直径阈值范围内，将源样本分配至相应簇。对于每个簇，分类器会估计源分布与目标分布之间的 KL 散度，并据此对簇进行排名。随后，根据各簇的散度排名自适应地筛选源样本，并按照 KL 散度将簇分为低、中、高三个类别。低散度类别中的簇与目标数据集最相似，因此允许保留比例最高的 ξ_1 源样本；中散度类别中的簇允许保留适中比例的 ξ_2源样本；而高散度类别中的簇表明差异更大，因此只允许保留少量 ξ_3源样本。

过滤后，每个保留的源样本都会根据其归一化 KL 估计值分配一个重要性权重，使得来自更相似簇的样本在训练过程中具有更高的影响，而来自较不相似簇的样本则被降低权重。在训练过程中，使用来自两个域的小批量来更新价值函数、计算目标值，并通过散度加权损失优化状态-动作价值函数。最后，更新目标网络，并使用条件变分自编码器（CVAE）在经过过滤的源样本上优化策略，从而实现跨域的有效适应。对于策略正则化权重 λ，我们发现，在所有摩擦实验中使用 $\lambda = 0.1$ 、在重力实验中使用 $\lambda = 0.5$ 、在形态实验中使用 $\lambda = 1.0$ ，以及在所有局部扰动实验中使用 $\lambda \in \{0.5, 1.0\}$ ，均可获得良好性能。

在蚁迷宫设置中，我们将簇数遍历设置为 $\{10, 20, 30, 50\}$，并观察到较小的簇数始终能带来更好的性能。我们推测，与 MuJoCo 环境中的运动任务不同，源 AntMaze 环境和目标 AntMaze 环境中的对齐转移会受到地图布局的影响，集中在少量簇中（即存在重叠的地图位置）。因此，使用较少的簇就已足够，并且能够获得更好的性能。对于策略正则化权重 λ，我们遍历 $\lambda = \{0.1, 0.5, 1.0\}$ 以获得最佳性能。在 Adroit 设置中，我们沿用 MuJoCo 任务中的类似超参数。我们将簇数遍历设置为 $\{30, 50\}$ ，并将策略正则化权重遍历设置为 $\lambda = \{0.1, 1.0\}$。

C.7 计算基础设施

我们在单个图形处理器 NVIDIA RTX A5000-24564MiB 和 8 个中央处理器 AMD Ryzen Threadripper 3960X 24-Core 上运行实验。每项实验需要 12GB 随机存取存储器，并需要 20GB 可用磁盘空间来存储数据。

D 更多实验结果

本节提供正文中省略的更多实验结果。我们展示了 LoDADA 与其他基线方法在存在形态变化的任务上的完整归一化分数对比。

D.1 形态变化下的结果

如表 8所示，我们的方法将总分提升至 597.88，较排名第二的 OTDF 提高了 32.7%，较骨干算法 IQL 提高了 63%。在这 16

算法 2 LoDADA：局部化动力学感知域适应

1: **输入：**源域数据 D_{src}，目标域数据 D_{tar}2: **初始化：**策略 π_ϕ，价值网络 V_ψ， Q_θ，目标Q函数 $Q_{\theta'}$，批大小 B，簇 m数量 K，直径阈值 δ，评论器重要性比率 α，正则化权重 λ，过滤率 io ξ_1， ξ_2， ξ_3，目标更新率 η，带有编码器 $E_\nu(s,\ a)$ 和解码器 $D_\varsigma(s,\ z)$ 的条件变分自编码器，采样e的潜变量数量 M**3：使用**$\mathcal{L}_{\text{CVAE}} = \mathbb{E}_{(s,a)\sim\mathcal{D}_{\text{tar}},\ z\sim E_\nu}\left(s,\ a\right)\left[\left\|a - D_\varsigma\left(s,\ z\right)\right\|_2^2 + D_{\text{KL}}\left(E_\nu\left(s,\ a\right)\|\mathcal{N}\left(\mathbf{0},\ I\right)\right.\right.$**训练条件变分自编码器策略，以对目标域数据集中的行为策略建模。**$\left.\left.\right)\right]$.**4：根据** s'**将** D_{tar} **聚类为** $\{\mathcal{N}^i\}_{i=1}^K$ **个簇，质心为** c^i**，直径为** d^i**。5：对于每个源样本**（$s,\ a,\ s'$）$\in D_{src}$**执行**

6 c^i $\{\mathcal{N}^i\}_{i=1}^K$**：在**$\leq \delta \cdot \frac{1}{K}\sum_{i=1}^K d^i$**中找到最近的质**

7 s' c^i **心：如果 dist(，)**$\leq \delta \cdot \frac{}{K}\sum_{i=1}^K d^i$**，则将**（$s,\ a,\ s'$）**分配到邻域** $\mathcal{N}^i$**。9：结束 if。10：结束 for。11：对于 i=**1**、2...、K，执行。12：根据公式 (4.1) 训练分类器** D_n（z）ˆ**。13：根据公式 (4.3) 计算 KL 散度的聚类估计** D_{KL}^n**。14：结束 for**

15：按ˆ D_{KL}^n 从小到大对 $\{\mathcal{N}^i\}_{i=1}^K$ 排序，得到 $\{\mathcal{N}^{(j)}\}_{j=1}^K$

16：**对**每个邻域 $\mathcal{N}^{(j)}, j = 1, \ldots, K$**执行**

17：秩点估计 $\{d_i^j\}^{|\mathcal{N}^{(j)}}$ $_{j-1}^{0}$ |的采样源数据

18：纳入排名前 $\xi_{g(j)}$%的样本，其中 $g(j) = \begin{cases} , & j \leq \lceil K/3\rceil,2 \quad \lceil K/3\rceil < j \leq \lceil 2K/3\rceil,, 3, \end{cases}$

$j > \lceil 2K/3\rceil$.

19：**结束循环**

20: 通过ˆ $d_i =$ d_i-最大值$_i$ d_i最大值$_i$ d_i-最小值$_i$ d_i， $i \in \{1, ..., |\tilde{D}_{src}|\}$ 归一化 KL 散度的点估计

21: **对于**每个梯度步**执行**.22: 从 D_{src} 中采样一个大小为 $B/2$ 的小批量 b_{src}，并从 D_{tar} 中采样一个大小为 $B/2$ 的小批量 b_{tar} 23: 通过以下方式更新状态价值函数 V_ψ：

$\mathcal{L}_V = \mathbb{E}_{(s,a)\sim\tilde{D}_{src}\cup D_{tar}}\left[L_2^\tau\left(Q_{\theta'}(s,\ a) - V_\psi\right.\right.$ $\left.\left.(s)\right)\right]$

24: 通过以下方式计算目标价值： $y = r + \gamma V_\psi(s')$ 25: 通过以下方式在 $b_{\text{src}} \cup b_{\text{tar}}$ 上优化状态-动作价值函数 Q_θ：

$$\mathcal{L}_Q = \mathbb{E}_{D_{tar}}[(Q_\theta - y)^2] + \mathbb{E}_{(s,a,s',d)\sim\tilde{D}_{src}}[\exp(-\alpha \cdot \hat{d}_i)(Q_\theta - y)^2].$$

26: 通过以下方式更新目标网络： $\theta' \leftarrow \eta\theta + (1-\eta)\theta'$

27: D从条件变分自编码器中解码出 M 个动作，并构建高斯分布 $\{\hat{}\pi_i^{\text{tar}}(\cdot \mid s)\}_{i=1}^M$。

28：计算优势 A，并使用 $b_{\text{src}} \cup b_{\text{tar}}$ 在 π_ϕ 上优化策略

$$\mathcal{L}_\pi = \mathbb{E}_{(s,a)\sim\tilde{D}_{\text{src}}\cup\mathcal{D}_{\text{tar}}}\left[\exp\left(A\right)\log\pi_\phi(a \mid s)\right] - \lambda\mathbb{E}_{s\sim\tilde{D}_{\text{src}}\cup\mathcal{D}_{\text{tar}}}\left[\log\left(\sum_{i=1}^M \hat{\pi}_{\text{tar}}^i\left(\pi_\phi(\cdot \mid s) \mid s\right)\right)\right].$$

29：**结束循环**

表 8：在 5 个随机种子下，HalfCheetah、Ant、Walker2d 和 Hopper 环境中不同**形态偏移**下的性能比较。我们用**M** 和**H** 分别表示动力学偏移的中等和困难级别。

Env	Type	级别	BOSA	IQL	DARA	IGDF	OTDF	Ours
HalfCheetah	大腿形态	M	25.12±4.33	22.22±1.13	20.68±2.49	19.31±4.35	21.94±6.30	**26.42±3.44**
		H	20.74±3.22	15.90±1.47	16.35±0.97	15.57±2.06	22.75±2.76	**25.81±5.29**
	躯干形态	M	0.66±0.90	1.95±1.69	1.40±1.66	1.84±0.37	3.38±1.30	**9.96±4.07**
		H	16.44±12.82	**30.87±10.00**	29.89±5.88	28.85±3.89	26.06±6.00	23.44±4.82
Ant	形态学-半腿	M	64.52±7.46	67.15±4.33	65.82±5.31	69.07±5.76	73.86±0.79	**77.48±0.77**
		H	37.21±6.15	54.27±5.60	61.31±5.71	60.08±2.09	59.76±9.27	**64.47±4.62**
	形态学-全腿	M	31.09±7.09	30.05±7.41	26.49±4.52	37.87±17.24	32.98±4.91	**77.55±3.44**
		H	13.89±1.46	5.65±3.30	3.21±3.26	7.24±1.40	12.84±2.06	**24.25±0.02**
Walker2d	躯干形态	M	8.63±3.49	11.44±0.91	14.80±1.69	14.16±2.19	11.47±4.93	**29.67±11.80**
		H	1.93±0.53	4.78±1.13	6.86±2.38	5.38±1.34	7.05±0.57	**7.76±2.42**
	morph-腿	M	28.10±8.64	40.61±4.78	43.87±6.22	41.71±10.14	45.07±10.09	**52.86±8.82**
		H	10.17±1.60	20.54±5.17	20.33±3.49	18.78±3.84	15.13±3.25	**59.01±1.76**
Hopper	形态足	M	12.87±0.09	**25.62±18.89**	20.05±7.32	14.91±4.52	22.56±8.65	13.11±0.05
		H	9.84±0.17	10.72±2.84	25.74±10.70	18.35±8.08	66.78±9.79	**67.53±14.03**
	躯干形态	M	13.43±0.37	13.66±0.98	14.70±4.48	13.06±4.33	16.84±1.44	**27.81±9.08**
		H	11.28±0.24	11.30±0.44	10.82±0.19	10.66±1.19	**12.23±0.15**	10.75±0.89
总计			305.92	366.73	382.32	376.84	450.70	**597.88**

在评估任务（8 个环境 × 2 个偏移级别）中，LoDADA 在 16 项任务中的 13 项取得了最佳表现。这些结果表明，我们提出的方法在形态变化下始终能够增强鲁棒性。与基线方法 IQL 相比，我们的方法在 16 项任务中的 13 项上持续取得显著的性能提升。

D.2 使用 play 源数据集在 AntMaze 上的结果

本节介绍使用 play 源数据集在 AntMaze 基准测试上进行的补充实验结果。我们采用与多样化数据集设置相同的评估流程，并报告 6 个 AntMaze 中等环境上的归一化目标环境性能。结果汇总于表9。如表9所示，在 play数据集设置下，我们提出的 LoDADA 实现了最佳总体性能，总分为 398.8，优于所有基线方法。与排名第二的方法相比，LoDADA 的总分提高了 5.3%，表明其在源数据集状态覆盖有限的情况下仍具有鲁棒性。我们的方法在 6 个目标地图中的 3 个上取得最佳性能，并在另外 2 个地图上排名第二。具体而言，LoDADA 在地图 $\{2, 3, 4\}$ 上表现最佳，并在地图 $\{5, 6\}$上取得第二佳表现。

D.3 Adroit任务上的结果

我们报告了所提方法及基线方法在 Adroit 操控基准测试上的表现，见 表10。如 表10所示，我们提出的 LoDADA 在 Pen 和门任务上始终优于所有基线方法。总体而言，LoDADA 的总分最高，达到 402.24，显著超过排名第二的方法（OTDF），其总分为 330.41。这表明，我们的方法在现实应用中面对严重的形态和运动学不匹配时具有很强的鲁棒性。这些结果表明，LoDADA 不仅适用于导航任务，也能在具有挑战性的操控基准测试中取得出色表现。通过利用局部动力学-

表9：在地图存在动力学偏移的 play离线数据集上，Antmaze任务的性能对比。源域保持不变；目标域发生偏移。我们报告归一化的目标域得分（均值 ′和五个随机种子的标准差）。最佳和次佳得分分别以绿色和蓝色突出显示。

Env	偏移级别	BOSA	DARA	IQL	IGDF	OTDF	Ours
Antmaze Medium	1	33.6 ± 8.0	52.4 ± 7.8	**76.4±10.1**		56.8 ± 6.87	52.8 ± 9.4
	2	25.2 ± 4.6	29.6 ± 6.2	46.8 ± 4.1	42.4 ± 8.2	63.2 ± 7.95	**67.6±6.7**
	3	30.0 ± 8.1	56.4 ± 11.1	66.4 ± 6.7	64.8 ± 7.0	69.6 ± 5.90	**71.6±3.8**
	4	34.0 ± 8.9	52.8 ± 6.3	48.8 ± 6.9	58.0 ± 1.4	44.8 ± 3.90	**71.6±6.8**
	5	22.8 ± 13.9	34.4 ± 9.3	**73.6±8.2**	54.4 ± 6.2	55.2 ± 3.03	60.8 ± 7.2
	6	12.8 ± 4.8	62.8 ± 3.0	66.8 ± 8.1	**74.8±4.1**	70.8 ± 5.93	74.4 ± 3.0
总计		158.4	288.4	378.8	352.4	360.4	**398.8**

表10：在 5 个随机种子下，Door 和 Pen 环境中发生**形态变化**和**运动学偏移**时的性能比较。我们用**M** 和**H** 分别表示动力学偏移的中等和困难级别。

Env	Type	级别	BOSA	IQL	DARA	IGDF	OTDF	Ours
Pen	运动学断裂关节	M	42.56±11.31	45.46±8.54	46.17±6.26	50.74±11.20	52.14±6.51	**64.31±11.33**
		H	19.36±5.42	24.02±5.16	22.41±4.73	18.62±10.47	20.95±7.23	**28.16±18.04**
	形态学收缩手指	M	10.33±2.79	15.64±2.97	15.33±2.39	13.19±4.41	18.02±5.64	**19.88±8.02**
		H	19.47±5.01	19.40±5.05	18.24±2.59	19.05±3.87	20.34±2.85	**33.06±11.48**
Door	运动学断裂关节	M	20.65±10.72	40.17±7.10	36.82±4.63	37.76±7.02	39.82±4.75	**55.12±12.50**
		H	19.80±15.61	52.37±5.52	49.31±5.53	52.61±4.83	58.24±6.37	**59.81±7.23**
	形态学收缩手指	M	41.05±15.36	54.85±7.79	54.49±5.88	54.73±5.28	59.15±4.89	**71.48±8.80**
		H	38.92±18.85	61.84±4.16	58.91±3.05	61.07±6.89	61.74±3.24	**70.42±4.26**
总计			212.15	313.75	301.67	307.78	330.41	**402.24**

感知过滤，我们的方法能够保留与目标域动力学高度匹配的源转移，而依赖全局假设的既有方法则难以应对形态和运动学偏移。

D.4 数据筛选方法的时间比较

为了理解 LoDADA 相较于其他数据筛选方法的优势，我们研究了本方法与 IGDF（Wen 等人，2024年）和 OTDF（Lyu 等人，2025年）之间数据筛选过程的运行时间。我们仅比较筛选过程的运行时间，而非整个训练过程，因为这些方法都可以共享相同的策略优化骨干，因此运行时间的主要差异出现在训练开始之前。具体而言，我们选择 ant-gravity-0.5 环境，并修改回放缓冲区，使其进行无放回采样，即每个轮次都会遍历整个数据集。为公平起见，我们将批大小固定为 32，轮次数固定为 7000。如表11所示，在目标数据集固定的情况下，对于 OTDF，由于其在源数据集和目标数据集之间逐点计算最优传输，其运行时间会随数据规模显著增长，这使其数据效率低下，尤其是在真实场景中。对于 IGDF，尽管其运行时间在三种情况下均为最少，但基于表1 和表2中的结果，我们认为其性能在三种数据筛选方法中最差，这表明效率提升是以策略质量大幅下降为代价的，且其过滤机制不足以在动力学偏移下捕获可靠转移。这进一步表明，本方法通过平衡模型性能与数据效率，相较于其他数据筛选方法具有优越性。

表11：Ant-Gravity-0.5 中不同源样本数量下的运行时间比较。

# 源	IGDF	OTDF	LoDADA
1000000	∼41.49秒	∼4597.82秒	∼134.62秒
100000	∼38.71秒	∼628.16秒	∼80.18秒
10000	∼55.64秒	∼126.39秒	∼72.92秒

参考文献

ABEL, D., JINNAI, Y., GUO, S. Y., KONIDARIS, G. 和 LITTMAN, M. (2018). 终身强化学习中的策略与值迁移。载于《国际机器学习会议》。PMLR。3 BARDE, P., ROY, J., JEON, W., PINEAU, J., PAL, C. 和 NOWROUZEZAHRAI, D. (2020). 对抗式软优势拟合：无需策略优化的模仿学习。载于《神经信息处理系统进展》33 12334–12344。3 CANONNE, C. L. (2022). 关于 KL 与 TV 之间一个不等式的简短注记。arXiv 预印本 arXiv:2202.07198。16 CORONATO, A., NAEEM, M., DE PIETRO, G. 和 PARAGLIOLA, G. (2020). 智能医疗应用中的强化学习：综述。《医学中的人工智能》109 101964。1 DVURECHENSKY, P., GASNIKOV, A. 和 KROSHNIN, A. (2018). 计算最优传输：通过加速梯度下降实现的复杂度优于 Sinkhorn 算法。载于《国际机器学习会议》。PMLR。2 EYSENBACH, B., CHAUDHARI, S., ASAWA, S., LEVINE, S. 和 SALAKHUTDINOV, R. (2021). 动力学分布外强化学习：使用域分类器进行迁移训练。载于《国际学习表征会议》。网址：https://openreview.net/forum?id=eqBwg3AcIAK 1, 3, 5, 9, 10, 15, 16, 23 EYSENBACH, B., GUPTA, A., IBARZ, J. 和 LEVINE, S. (2019). 你所需的一切就是多样性：无奖励函数的技能学习。载于《国际学习表征会议》。网址：https://openreview.net/forum?id=SJx63jRqFm 3 FU, J., KUMAR, A., NACHUM, O., TUCKER, G. 和 LEVINE, S. (2020). D4RL：深度数据驱动强化学习数据集。arXiv 预印本 arXiv:2004.07219。8, 9, 10 GU, J., HE, Y., LIU, Z. 和 XU, P. (2025). 具有线性函数逼近的策略正则化分布鲁棒马尔可夫决策过程。arXiv 预印本 arXiv:2510.14246。3 GUO, Y., WANG, Y., SHI, Y., XU, P. 和 LIU, A. (2024). 通过领域适应和奖励增强模仿实现动力学分布外强化学习。载于《神经信息处理系统进展》第37卷。2, 3 GUO, Y., YANG, Y., XU, P. 和 LIU, A. (2026). MOBODY：基于模型的动力学分布外离线强化学习。载于《第十四届国际学习表征会议》。网址：https://openreview.net/forum?id=7c0YS3cuno 2, 3 HAARNOJA, T., ZHOU, A., ABBEEL, P. 和 LEVINE, S. (2018). 软演员-评论家：带随机演员的离策略最大熵深度强化学习。载于《国际机器学习会议》。PMLR。8, 9 HE, Y., LIU, Z., WANG, W. 和 XU, P. (2025). 在线交互下分布鲁棒动力学分布外强化学习的样本复杂度。载于《第四十二届国际机器学习大会》。网址：https://openreview.net/forum?id=pJdMOKqdSV 1, 3

KAELBLING, L. P., LITTMAN, M. L. 和 MOORE, A. W. (1996). 强化学习：综述。《人工智能研究杂志》4 237–285。1 KIRAN, B. R., SOBH, I., TALPAERT, V., MANNION, P., AL SALLAB, A. A., YOGAMANI, S. 和 PÉREZ, P. (2021). 自动驾驶中的深度强化学习：综述。《IEEE智能交通系统汇刊》23 4909–4926。1 KOSTRIKOV, I., NAIR, A. 和 LEVINE, S. (2021). 基于隐式 Q 学习的离线强化学习。arXiv 预印本 arXiv:2110.06169。3, 7, 9, 23 LEE, J., JEON, W., LEE, B., PINEAU, J. 和 KIM, K.-E. (2021). OptiDICE：通过稳态分布校正估计进行离线策略优化。载于《国际机器学习会议》。PMLR。3 LEVINE, S., KUMAR, A., TUCKER, G. 和 FU, J. (2020). 离线强化学习：教程、综述及开放问题展望。arXiv 预印本 arXiv:2005.01643。4 LIU, J., ZHANG, H. 和 WANG, D. (2022). DARA：离线强化学习中的动态感知奖励增强。arXiv 预印本 arXiv:2203.06662。1, 2, 3 LIU, J., ZHANG, Z., WEI, Z., ZHUANG, Z., KANG, Y., GAI, S. 和 WANG, D. (2024a). 超越分布外状态动作：支持集约束的跨域离线强化学习。载于《AAAI人工智能会议论文集》第38卷。9, 22 LIU, Y., SWAMINATHAN, A., AGARWAL, A. 和 BRUNSKILL, E. (2020). 无需大量探索即可实现有理论保证的优良批量离策略强化学习。《神经信息处理系统进展》33 1264–1274。3 LIU, Z., WANG, W. 和 XU, P. (2024b). 分布鲁棒动力学分布外强化学习的上下界。arXiv 预印本 arXiv:2409.20521。3 LIU, Z. 和 XU, P. (2024). 分布鲁棒动力学分布外强化学习：基于线性函数逼近的可证明高效性。载于《人工智能与统计学会议》。PMLR。3 LUO, Y., XU, H., LI, Y., TIAN, Y., DARRELL, T. 和 MA, T. (2019). 具有理论保证的基于模型深度强化学习算法框架。载于《国际学习表征会议》。网址：https://openreview.net/forum?id=BJe1E2R5KX 16 LYU, J., BAI, C., YANG, J.-W., LU, Z. 和 LI, X. (2024a). 通过捕捉表征失配实现跨领域策略适应。载于《国际机器学习会议》。PMLR。2, 3, 4, 13, 16 LYU, J., XU, K., XU, J., YANG, J.-W., ZHANG, Z., BAI, C., LU, Z., LI, X. 等 (2024b). ODRL：动力学分布外强化学习基准。《神经信息处理系统进展》37 59859–59911。7, 8, 17, 20 LYU, J., YAN, M., QIAO, Z., LIU, R., MA, X., YE, D., YANG, J.-W., LU, Z. 和 LI, X. (2025年). 基于最优传输和数据集约束的跨领域离线策略适应。载于《第十三届国际学习表征会议》。2, 3, 5, 7, 9, 24, 29

MACHADO, M. C., BELLEMARE, M. G. 和 BOWLING, M. (2017). 用于强化学习选项发现的拉普拉斯框架。载于《国际机器学习会议》。PMLR。3 NACHUM, O., CHOW, Y., DAI, B. 和 LI, L. (2019). Dualdice：与行为无关的折扣平稳分布校正估计。《神经信息处理系统进展》32。3 PENG, X. B., KUMAR, A., ZHANG, G. 和 LEVINE, S. (2019). 优势加权回归：简单且可扩展的离线策略强化学习。arXiv 预印本 arXiv:1910.00177。6 SUTTON, R. S., BARTO, A. G. 等 (1998). 强化学习：导论，第 1 卷。剑桥麻省理工学院出版社。1 WANG, R., YANG, Y., LIU, Z., ZHOU, D. 和 XU, P. (2024). 用于动力学分布外强化学习的回报增强型决策变换器。arXiv 预印本 arXiv:2410.23450。1, 2 WEN, X., BAI, C., XU, K., YU, X., ZHANG, Y., LI, X. 和 WANG, Z. (2024). 跨领域离线强化学习中用于数据筛选的对比表征。arXiv 预印本 arXiv:2405.06192。2, 3, 5, 9, 10, 23, 29 XU, H., ZHAN, X., LI, J. 和 YIN, H. (2021). 采用软行为正则化的离线强化学习。arXiv 预印本 arXiv:2110.07395。3 XU, K., BAI, C., MA, X., WANG, D., ZHAO, B., WANG, Z., LI, X. 和 LI, W. (2023). 通过价值引导的数据筛选实现跨领域策略适应。《神经信息处理系统进展》36 73395–73421。2, 3 YU, T., THOMAS, G., YU, L., ERMON, S., ZOU, J. Y., LEVINE, S., FINN, C. 和 MA, T. (2020). Mopo：基于模型的离线策略优化。《神经信息处理系统进展》33 14129–14142。3