复现：ODRL: A Benchmark for Off-Dynamics Reinforcement Learning

摘要

动态不匹配领域的强化学习（Off-Dynamics Reinforcement Learning, ODRL）聚焦于源域与目标域动态差异对策略迁移能力的影响，是强化学习领域的一项重要研究方向。本文复现了 ODRL: A Benchmark for Off-Dynamics Reinforcement Learning [4] 提出的 ODRL 基准框架以及其中的核心实验，验证了不同算法的适应能力和局限性，实验结果表明，现有算法在动态不匹配场景中的通用性不足，不同任务和动态变化类型对算法的适应能力提出了不同的挑战，源域性能与目标域性能间并无明确的正相关性。未来研究应重点关注通用动态感知算法的开发，提升目标域有限数据的利用效率，并引入更复杂和多样化的任务场景，以进一步推动动态失配强化学习在实际应用中的发展和突破。

关键词：深度强化学习；策略迁移；动态不匹配；基准框架

1 引言

在深度强化学习 (Deep Reinforcement Learning) 领域中，策略跨域迁移 (policy transfer) 是一个存在大量研究但仍未完全解决的重要问题。特别是在动态不匹配 (dynamics mismatch) 的域中, 不同域之间的动力学差异可能导致在源域训练好的策略在目领域的表现大幅下降。现阶段研究为了解决这一问题，提出了一些针对的强化学习算法。但由于缺乏统一的基准平台对这些算法进行全面评估，阻碍了该领域的发展和进步。

本次复现的论文 ODRL: A Benchmark for Off-Dynamics Reinforcement Learning [4] 是首个专门针对动态不匹配的强化学习算法的基准测试平台。该测试平台提供在线/离线源域与目标域的四种实验设置，以匹配不同算法的需求；覆盖多样化的任务和动态变化（如摩擦、重力、运动学、形态学等）；实现了现有算法的统一代码框架，后续增加算法评测十分便捷。

该测试平台的开发不仅填补了动态不匹配领域强化学习算法缺乏标准化基准的空白，也为科研学者设计动态感知算法提供了可靠的测试平台。同时这篇论文还通过对现有算法进行广泛的评估，旨在推动此类算法在真实世界中的应用和发展。综上所述，ODRL 为动态不匹配问题强化学习算法的研究提供了重要的基准测试，能够推动该领域的算法设计向更加通用和高效方向发展。

2 相关工作

跨域策略迁移（policy transfer across domains）是深度强化学习中的一个重要研究方向，如何有效地在存在动态不匹配的领域间进行策略的泛化仍是强化学习领域中一个尚未解决的问题，在 [2] 中，这类问题被称作动态失配强化学习，而本次复现的论文将这一概念推广，并作出了以下定义：智能体能够在源域 M_{src} 中获得充分的交互数据，且能在目标域 M_{tar} 中获得有限度的交互数据，源域 M_{src} 和目标域 M_{tar} 之间存在动态不匹配，智能体旨在充分利用好两个域的数据进行训练，并在目标域 M_{tar} 中得到更好的性能表现。 [4] 下面是现有研究在相关算法和相关基准框架上的工作：

2.1 现有相关算法

动态随机化（Domain Randomization）：动态随机化通过引入随机化参数来模拟出更多样化的环境，从而增强策略的鲁棒性，使其适应未见过的动态变化 [7]。例如在环境重力参数中加入一定随机化，可以训练出更能适应不同重力环境的智能体，这类方法的效果十分依赖于随机化范围的设计，如果参数范围设置不合理，可能导致过拟合或适应性不足

系统辨识（System Identification）：系统辨识旨在通过学习目标域的动态特性来减小源域与目标域的动态差距。这类方法通常需要对目标域进行动态参数估计或建模，从而实现更加精准的策略迁移 [12]。这类方法会利用目标域的数据，动态地调整从源域中获得的环境参数。这类方法需要大量的目标域数据，在复杂变化的环境或缺乏目标域数据的情况下，可能难以成功学习到目标域和源域的差异。

元学习（Meta-Learning）：元学习方法通过在多任务训练中学习一个通用的元策略（meta-policy），在多个不同动态的任务中训练一个元策略，元策略学习到如何快速调整自己以适应新的任务，使得智能体能够快速适应新任务中的动态变化 [6]。该方法特别适用于动态变化多样的场景，这类方法更多任务设置，高维稀疏奖励任务中，元学习的表现可能不稳定。

动态感知算法（Dynamics-Aware Algorithms）：动态感知算法直接从动态变化的建模入手，设计能够自适应动态变化的强化学习算法。这类方法更侧重于数据的选择与特征的提取，以提升策略的泛化能力，具体的实现有通过对比学习或价值目标对源域数据进行筛选，从而过滤掉不符合目标域动态特征的数据的数据过滤方法（VGDF）[10]、使用领域分类器（domain classifiers）来测量源域与目标域的动态差异，并利用这一差异调整训练过程的 DARC 算法 [2]、使用重要性采样（Importance Weighting）技术调整目标域数据的分布的 H2O 算法 [5]，都可以归类在动态感知算法下。

2.2 现有基准框架

现有的许多基准框架已经广泛应用于强化学习领域，但它们大多专注于特定任务或问题（如多任务学习、离线强化学习），而非复现论文所提到的动态不匹配领域的强化学习场景，以下是目前已有的相关基准测试框架：

D4RL（Datasets for Deep Data-Driven Reinforcement Learning）：D4RL 是一个评测离线强化学习的基准框架，提供了大量的单域离线数据集（如机器人控制和导航任务），以评估算法在不同数据质量（random、medium、expert）上的表现 [3]。主要关注离线数据集的适应性，未涉及跨域或动态不匹配的场景，无法很好地评估动态不匹配相关算法的性能。

DeepMind Control Suite (DMC)：DMC 是一个广泛应用于连续控制任务的强化学习基准测试框架，包含多个基于 MuJoCo 的经典任务（如 Ant、Walker2D 和 Hopper），适用于在线强化学习研究 [8]。DMC 提供了大量关于连续控制的问题，但任务设计集中在单一动态环境中，未涉及跨域或动态不匹配的场景，也不涉及离线数据集。

Meta-World：Meta-World 是一个多任务强化学习和元学习的基准框架，由 50 个基于模拟机器人操作的任务组成，适用于多任务训练和跨任务迁移 [11]。Meta-World 具有评估多任务场景下的迁移学习的能力，但重点在于多任务和元学习，而非动态变化，因此对动态不匹配问题缺乏足够支持。

CARL 和 Continual World 是两个针对上下文变化和持续学习的基准框架。CARL 专注于上下文变化场景，环境的动态属性或奖励函数随上下文切换而改变 [1]；Continual World 聚焦于多任务的持续学习能力，通过评估任务切换时的适应性表现 [9]。这两个框架的任务主要针对上下文切换的场景，缺乏明确的源域和目标域概念，不完全适用于动态不匹配强化学习的评估。

与上述框架相比，ODRL 更专注于动态失配问题，设计了涵盖摩擦、重力、运动学和形态学变化的动态任务，以评估算法在不同动态不匹配场景下的适应能力。相比现有框架，ODRL 的提出填补了动态失配领域的空白，为动态感知算法提供了一个标准化的评估平台。

3 本文方法

3.1 本文方法概述

复现的论文首先正式定义了动态失配强化学习问题：在动态失配强化学习中，源域（source domain）和目标域（target domain）分别表示为 $M_{src} = \langle S, A, P_{src}, r, \rho_0, \gamma \rangle$ 和 $M_{tar} = \langle S, A, P_{tar}, r, \rho_0, \gamma \rangle$，其中两域的状态空间 S、动作空间 A 和奖励函数 r 相同，但状态转移概率 P_{src} 和 P_{tar} 不同，最终目标是通过源域的充分数据和目标域的有限数据，在目标域中获得更优的策略，这一定义扩展了之前仅支持在线源域和目标域的设定，涵盖了更广泛的动态不匹配场景。

ODRL 包含三类任务（运动控制、导航、灵巧操作）和四种实验设置（在线-在线、离线-在线、在线-离线、离线-离线），不仅适配了不同的研究需求，还提供了统一的算法实现与多质量级别的离线数据集。

3.2 实验设置

ODRL 提供了如图 1 [4] 所示四种实验设置，模拟实际应用中的多种场景，包括在线-在线：源域和目标域均为智能体在线交互学习；离线-在线：源域使用离线数据集训练，目标域为在线交互学习；在线-离线：源域为在线交互学习，目标域为离线数据集训练；离线-离线：源域和目标域均为离线数据集。

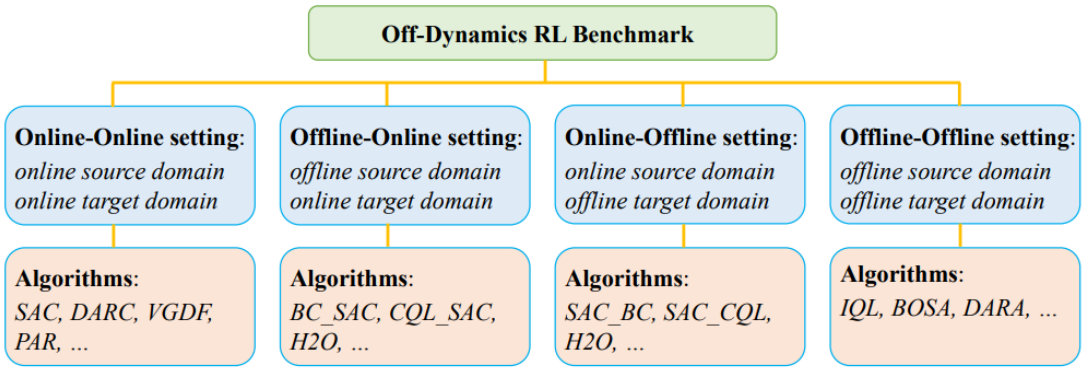

图 1. 实验设置

3.3 任务设计

如图 2 [4] 所示 ODRL 提供了三类任务：运动控制（Locomotion）任务：包括 Ant、Hopper、Walker2D 和 HalfCheetah 四种不同形态的机器人，涉及摩擦变化、重力变化、运动学变化和形态学变化四种动态变化类型，每种变化分为三种不同强度；导航（Navigation）任务：使用 AntMaze 任务，通过改变迷宫地图的结构，评估策略在不同地形中的迁移能力，提供了三种不同大小的地图，每种大小地图有六种变化。灵巧操作（Dexterous Manipulation）任务：包括 Pen、Door、Relocate 和 Hammer 四种 Adroit 任务，通过改变环境的摩擦变化、重力变化和机械手的形态学特征来实现动态变化，这些任务涵盖了从低维连续控制到高维稀疏奖励的多样化场景，能够全面评估算法的适应能力。

为了模拟现实世界中的复杂动态变化，ODRL 设计了以下四类动态变化类型：摩擦变化（Friction Shift）：通过改变机器人与地面之间的摩擦系数；重力变化（Gravity Shift）：改变环境中的重力大小，但保持方向不变；运动学变化（Kinematic Shift）：限制机器人关节的转动范围；形态学变化（Morphology Shift）：修改机器人的结构尺寸。

以上三类任务和四类动态变化，再加上每种变化的不同强度，可以适配了不同的研究需求，为动态不匹配领域的强化学习算法提供了一个标准化的评估平台。

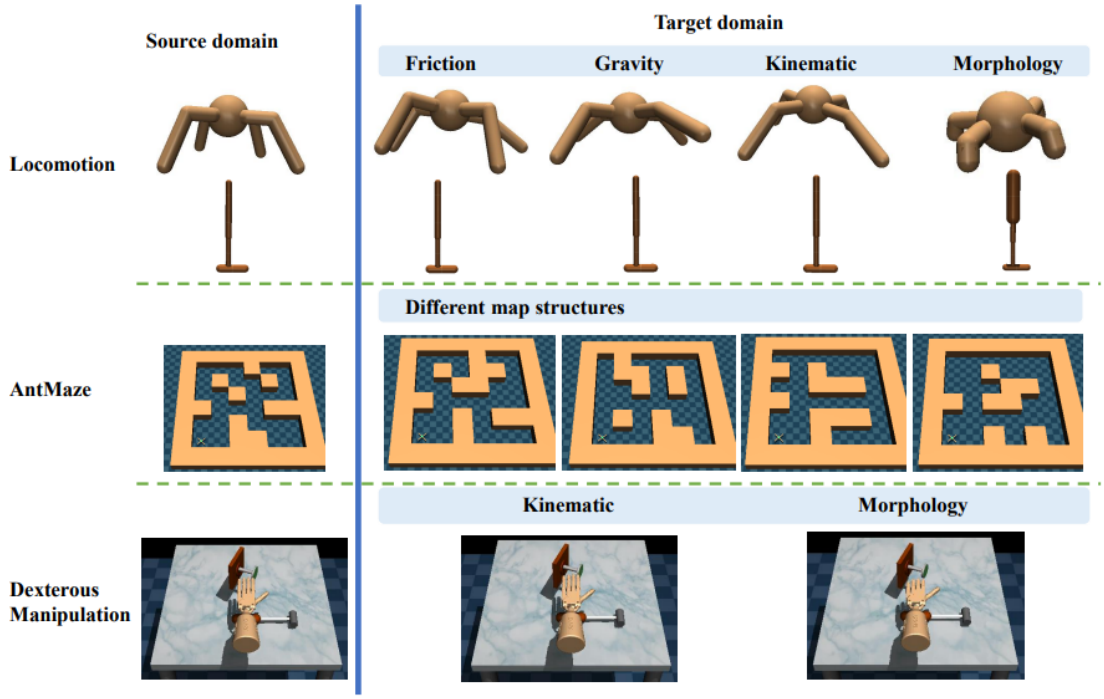

图 2. 任务设计

3.4 评价指标

为了评估算法在目标域中的适应能力，ODRL 提供了以下两个标准化评价指标：直接评估智能体在目标域中交互所获得的累积奖励；使用标准化分数进行评价，通过归一化策略在目标域中的表现，以便与随机策略和专家策略进行比较，计算公式如下，其中，J_π 为策略的累计奖励，J_r 为随机策略的奖励，J_e 为专家策略的奖励：

$$NS = \frac{J_\pi - J_r}{J_e - J_r} \times 100$$

4 复现细节

4.1 与已有开源代码对比

论文提供公开源代码于 https://github.com/OffDynamicsRL/off-dynamics-rl. 原论文已提供了一个较为完善的开源代码实现，包含了 ODRL 框架的主要任务与相关算法的实现。但原始代码存在一定不足：原始代码在硬件需求和超参数设置上较为苛刻，默认配置下难以在本地环境运行；原代码未提供针对训练过程的详细分析工具，特别是在奖励曲线和标准化分数变化的可视化方面，训练完成后无法进行只直观分析；原始实现中超参数的调整功能较为有限，部分关键参数被硬编码，修改起来较为不便。

基于上述问题，我在复现过程中进行了以下改进：编写了一套绘图代码，用于绘制训练过程中奖励和标准化分数随训练轮数的变化曲线，帮助更直观地分析训练过程；改写了更加灵活的超参数配置接口，并调整了部分超参数（如批量大小、学习率等）以降低显存占用，使代码能够在本地计算资源上顺利运行，这些改进不仅提高了代码的适用性，也为后续实验结果分析提供了支持。

4.2 实验环境搭建

复现工作在本地计算机上完成，硬件环境为：GPU：Intel Core i7-10875H；GPU：NVIDIA RTX 2060 (Notebooks)；内存：32GB，软件环境为：Anaconda 虚拟环境，具体环境配置根据原论文提供的 README 文件和 Requirements 文件进行配置。同时为了适应上述硬件环境，我将源代码训练过程中的计算资源使用进行了优化，包括对批大小和经验缓存区大小等参数进行了适当调整。

4.3 创新点

尽管大部分复现工作均为基于原论文的开源代码实现，但在以下几个方面的改进和补充体现了本次复现工作的价值和创新性：通过对超参数的优化和代码结构的调整，成功在中端硬件设备上复现了原始论文中的主要实验结果；在复现过程中，我编写了一套绘图代码，用于绘制训练过程中的奖励和标准化分数随训练轮数的变化曲线，直观展示模型的适应能力，便于不同算法的性能比较。这些可视化工具为实验结果的直观展示和后续分析提供了重要支撑。

本次复现工作不仅成功运行了原论文的核心实验，还通过对代码的调整和功能的补充，使其能够适应资源受限的实验环境，并提供了更加直观的结果分析工具，还通过对算法和任务的深入分析，揭示了不同动态不匹配的强化学习算法的表现特点和局限性。

5 实验结果分析

在本次复现实验中，我们对如图 3 [4] 所示的原论文的核心实验进行了重现与分析，使用 ODRL 基准框架对多个动态不匹配领域的强化学习算法进行了评测和结果分析。

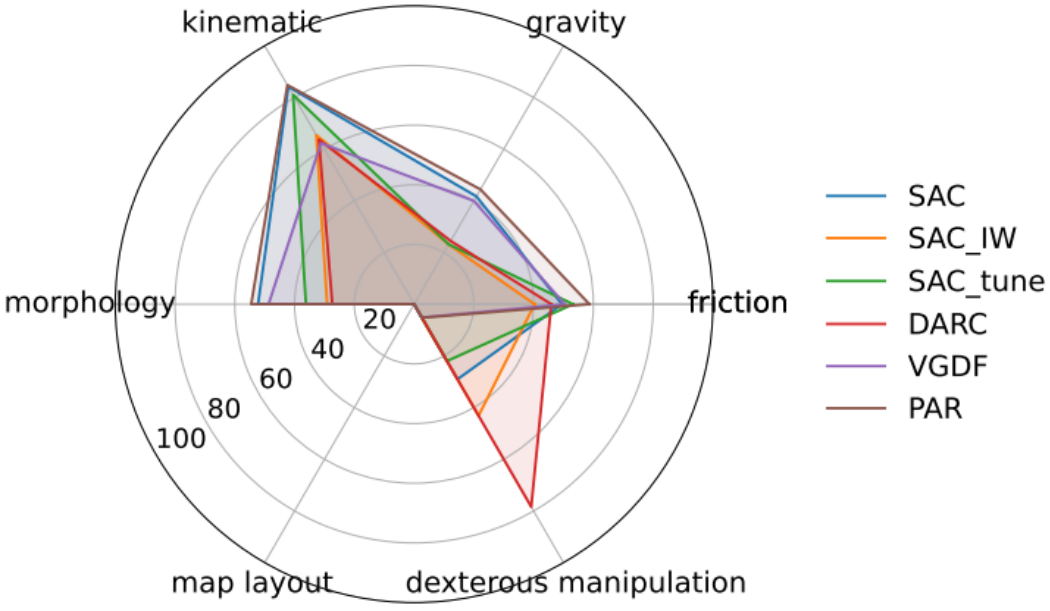

图 3. 原论文核心实验

5.1 算法性能比较评测

如图 4所示，这是我对原论文核心实验的复现结果，将多个动态不匹配领域的强化学习算法在六个不同领域中进行评测，以分析这些算法的综合能力，以下是一些观察结果：

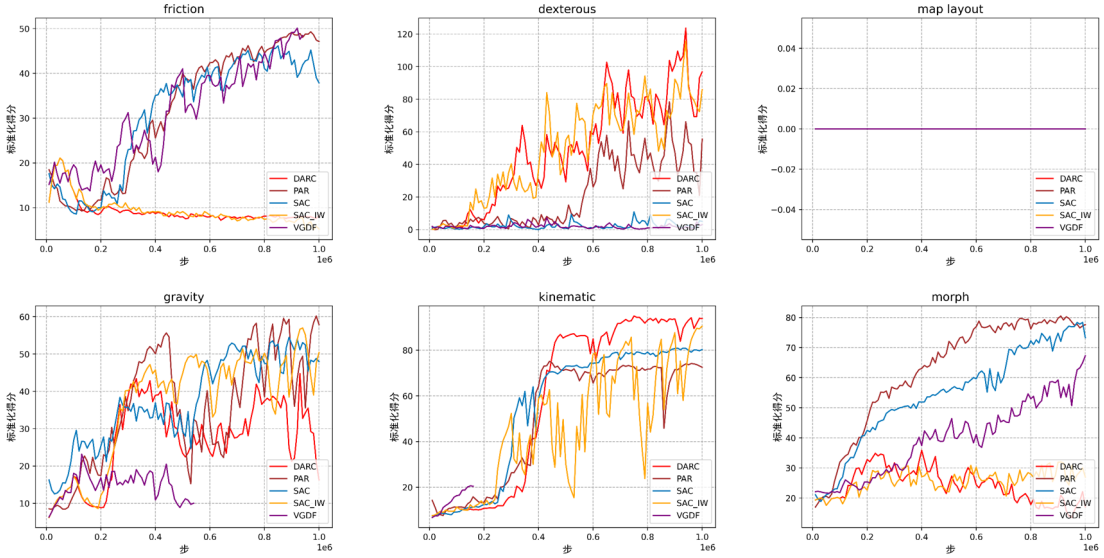

图 4. 复现结果

目前测评的动态不匹配领域的强化学习算法中，没有任何一个算法能够在所有任务中都表现出优越的性能，不同算法在特定任务和动态变化类型下展现出了各自的优劣，但无法形成统一的动态感知适应能力，本领域算法的通用性发展仍是尚未解决的难题；

PAR 算法在运动任务中表现优异，但在导航任务和灵巧操控任务中都表现不佳，这种差异可能与源域奖励的惩罚项设置有关，尤其是在稀疏奖励任务中会产生过估现象；

AntMaze 导航任务具有极高的挑战性，没有算法能够在该任务中获得有意义的回报。这表明基于状态的强化学习方法在动态变化的迷宫中难以适应，也可能是因为相比其它动态变化，迷宫构造的改变会对整体最优策略以及状态价值产生更加剧烈的扰动，这一结果表明在复杂路径规划任务中的强化学习算法设计仍有待发展；

动态不匹配强化学习算法在灵巧手操作任务（如开门、抓取等）中的表现普遍较差，只有两种基于域分类器的算法在某些设置下能够获得相对较好的性能，这可能表明现有算法在处理高维稀疏奖励问题时仍有明显不足；

基线算法 SAC 通过直接使用源域和目标域的混合数据训练，能够在许多任务中获得较好的表现，虽然这种直接混合的方法在形态变化和重力变化场景中表现欠佳，但仍然说明其具有一定的适应能力，作为基线算法去衡量其它算法的效果比较合适。

5.2 源域与目标域性能的关联性

同时，我还根据原文实验，对比源域与目标域的策略性能，进一步分析了两者之间的关联性，如图 5所示，智能体在源域与目标域的性能之间没有明确的相关性，例如在 ant-friction-0.5 任务中，VGDF 算法目标域中的表现相对优秀，但在源域中的性能较差，相反，DARC 和 SACIW 在源域中表现较强，但在目标域中的适应性较弱，这一观察结果表明算法在源域中的

性能并不是预测目标域表现的可靠指标，算法在跨域适应过程中需要更加注重目标域特性的直接学习，而不是仅依赖源域的良好表现。

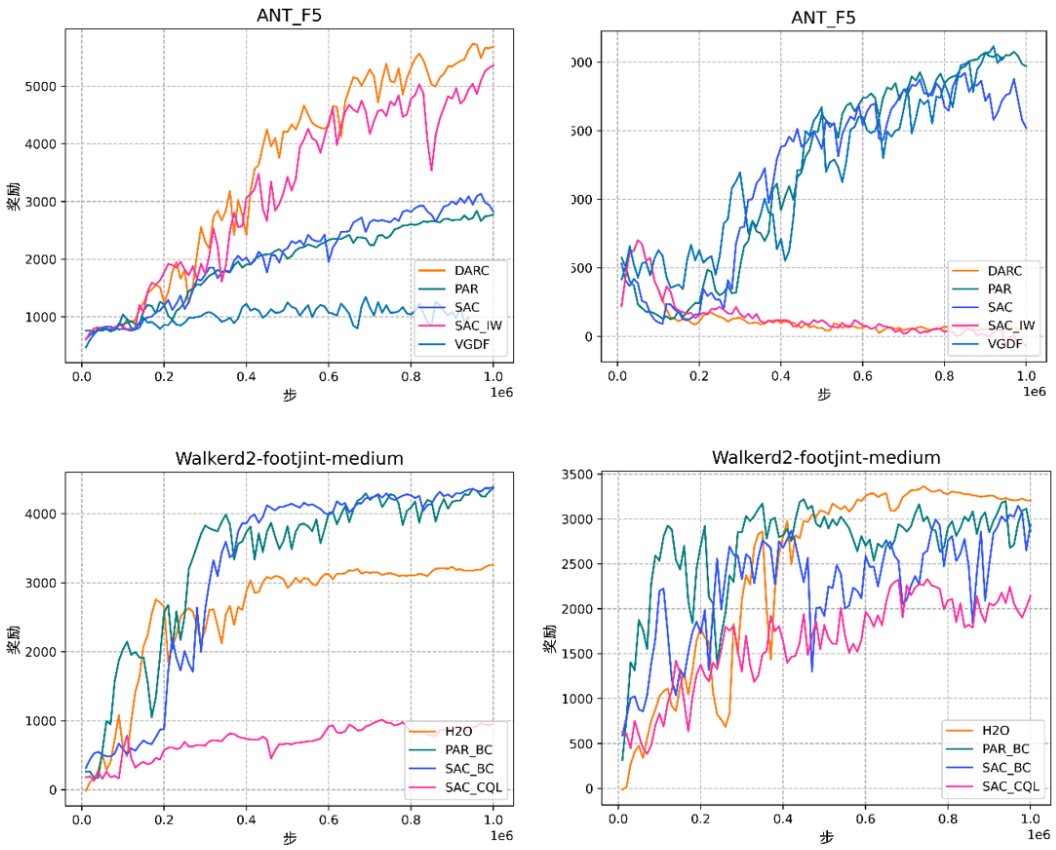

图 5. 源域与目标域性能的关联性

6 总结与展望

本次复现实验重点复现了 ODRL: A Benchmark for Off-Dynamics Reinforcement Learning [4] 中的测试框架对多个动态不匹配领域的强化学习算法的评测与分析的实验。验证了不同动态变化类型和任务场景下现有算法的性能与局限性。结果表明，目前的动态失配强化学习算法无法在所有任务和动态变化中保持优异表现，尤其在复杂导航和灵巧操控任务中适应性显著不足。此外，实验揭示了源域性能与目标域性能之间的非线性关系，表明源域的良好表现并不能直接保证目标域的高效适应。复现过程中，通过超参数调整、实验环境优化及可视化分析工具的开发，也为我未来的科研学习打下坚实的基础。

未来的研究重点可以放在通用动态感知算法的设计，以应对多种任务场景的适应性问题，如何高效利用目标域的稀缺数据，结合源域数据提升目标域性能，也是一个值得深入探索的方向。在任务设计方面，本基准框架的任务复杂性和动态变化的多样性可以进一步增加。通过对现有算法的进一步优化和新方法的开发，动态不匹配强化学习有望在更复杂的实际应用场景中展现更高的通用性和鲁棒性。

参考文献

[1] Carolin Benjamins, Theresa Eimer, Frederik Schubert, Aditya Mohan, Sebastian Döhler, André Biedenkapp, Bodo Rosenhahn, Frank Hutter, and Marius Lindauer. Contextualize me–the case for context in reinforcement learning. *arXiv preprint arXiv:2202.04500*, 2022.

[2] Benjamin Eysenbach, Swapnil Asawa, Shreyas Chaudhari, Sergey Levine, and Ruslan Salakhutdinov. Off-dynamics reinforcement learning: Training for transfer with domain classifiers. *arXiv preprint arXiv:2006.13916*, 2020.

[3] Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4rl: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*, 2020.

[4] Jiafei Lyu, Kang Xu, Jiacheng Xu, Mengbei Yan, Jingwen Yang, Zongzhang Zhang, Chenjia Bai, Zongqing Lu, and Xiu Li. Odrl: A benchmark for off-dynamics reinforcement learning. *arXiv preprint arXiv:2410.20750*, 2024.

[5] Haoyi Niu, Yiwen Qiu, Ming Li, Guyue Zhou, Jianming Hu, Xianyuan Zhan, et al. When to trust your simulator: Dynamics-aware hybrid offline-and-online reinforcement learning. *Advances in Neural Information Processing Systems*, 35:36599–36612, 2022.

[6] Roberta Raileanu, Max Goldstein, Arthur Szlam, and Rob Fergus. Fast adaptation to new environments via policy-dynamics value functions. In *Proceedings of the 37th International Conference on Machine Learning*, pages 7920–7931, 2020.

[7] Reda Bahi Slaoui, William R Clements, Jakob N Foerster, and Sébastien Toth. Robust domain randomization for reinforcement learning. 2019.

[8] Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, et al. Deepmind control suite. *arXiv preprint arXiv:1801.00690*, 2018.

[9] Maciej Wołczyk, Michał Zając, Razvan Pascanu, Łukasz Kuciński, and Piotr Miłoś. Continual world: A robotic benchmark for continual reinforcement learning. *Advances in Neural Information Processing Systems*, 34:28496–28510, 2021.

[10] Kang Xu, Chenjia Bai, Xiaoteng Ma, Dong Wang, Bin Zhao, Zhen Wang, Xuelong Li, and Wei Li. Cross-domain policy adaptation via value-guided data filtering. *Advances in Neural Information Processing Systems*, 36:73395–73421, 2023.

[11] Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on robot learning*, pages 1094–1100. PMLR, 2020.

[12] Wenxuan Zhou, Lerrel Pinto, and Abhinav Gupta. Environment probing interaction policies. *arXiv preprint arXiv:1907.11740*, 2019.