基于 Transformer 强化学习设计 A/B 测试中的时间序列实验

Xiangkun Wu[1,*] 浙江大学 数学科学学院，杭州，中国 12235031@zju.edu.cn

Qianglin Wen[2,*] 环境建模与数据分析实验室 云南大学 昆明，中国 qianglin@mail.ynu.edu.cn

Yingying Zhang[3,*] 华东师范大学统计学院 上海，中国 yyzhang@fem.ecnu.edu.cn

Hongtu Zhu[4] University of North Carolina 教堂山 教堂山，北卡罗来纳州，美国 htzhu@email.unc.edu

Ting Li[5,†] 上海财经大学统计与数据科学学院 上海，中国 tingli@mail.shufe.edu.cn

Chengchun Shi[6,†] 伦敦政治经济学院统计系 伦敦，英国 C.Shi7@lse.ac.uk

* 同等贡献。† 通讯作者。

arXiv:2602.01853v1 [cs.LG] 2 Feb 2026

摘要

A/B 测试已成为现代科技公司进行策略评估的黄金标准。然而，将其应用于时间序列实验——即随时间推移依次分配策略的实验——仍然面临挑战。现有设计存在两项局限：（i）未能充分利用完整历史信息进行处理分配；（ii）依赖强假设来近似目标函数（例如，估计处理效应的均方误差），以优化设计。我们首先提出一项不可能性定理，说明由于时间序列实验中的动态依赖，未能以完整历史信息为条件会导致设计次优。为同时解决这两项局限，我们进而提出一种 Transformer 强化学习（RL）方法，利用 Transformer 根据完整历史信息来进行分配，并通过强化学习直接优化均方误差，而不依赖严格的假设。在合成数据、公开可用的派单模拟器以及真实世界网约车数据集上进行的实证评估表明，我们的方法始终优于现有设计。

1 引言

本文研究 A/B 测试，该方法已被众多科技公司采用，包括 Airbnb、DoorDash 和美团等多边平台，以及亚马逊和阿里巴巴等大型电子商务企业，

以及谷歌、Meta 和领英等搜索与社交网络领域的领先企业。A/B 测试的主要目标是策略评估：将新开发的网站设计、策略或产品（称为处理组）与现有版本（称为对照组）进行比较。从本质上说，A/B 测试运用因果推断方法（参见 Imbens & Rubin，2015；Ding，2024；Wager，2024）来设计在线实验，将实验单位随机分配至处理组和对照组，并根据收集的数据推断平均处理效应。

我们关注这样一类应用：实验单位会随着时间推移依次接受一系列处理，从而产生构成时间序列的数据（例如，参见 Bojinov & Shephard，2019年；Farias 等，2022；Shi 等，2023b；Li 等，2024a；Xiong 等，2024b）。一个具有代表性的应用是优步、Lyft 和滴滴出行等大型网约车平台上的 A/B 测试。这些公司作为双边市场运营，将乘客与司机匹配，为用户提供高效便捷的交通服务。它们经常开发并更新订单派发策略（Xu 等人，2018；Wan 等，2021；Zhou 等，2021）、车辆重新调度策略（Pouls 等，2020；Jiao 等，2021）、定价与补贴策略（Fang 等，2017；Chen 等，2019），并通过 A/B 测试进行验证（Chamandy，2016；Shi 等，2023a）。这些策略会随时间依次实施。此外，在网约车领域，请求乘车的乘客人数（需求）和司机的在线时长（供给）会随时间动态变化（例如，参见 Luo 等人，2024，图 1）。与策略序列一起，这些变量构成了三个相互作用的时间序列：策略影响需求和供给，而需求和供给也彼此相互作用。

此类时间序列实验中的 A/B 测试面临三项关键挑战：

1. **残留效应** 在此类实验中普遍存在（Shi 等，2023b；Xiong 等，2024b；Ni 和 Bojinov，2025；Wen等人，2025）。也就是说，策略通常会产生延迟 效应：每个时间点实施的策略既会影响公司的当前结果，也会影响其未来结果，从而违反经典的稳定单元处理值假设（例如，参见 Imbens 和 Rubin，2015）。在网约车领域，订单派发、车辆重新调度、定价和补贴等策略都会产生延迟效应。它们会改变城市中司机的空间分布，而这类变化是逐渐发生的，因此对未来结果的影响也会有所延迟。忽略这些残留效应，并将时间序列数据视为独立同分布数据来采用经典 A/B 测试方法，往往会导致平均处理效应（ATE）估计量不显著（Shi 等，2023b）。

2. **较小的处理效应**使得人们极难根据新旧策略对公司结果的影响加以区分（Athey 等，2023）。在网约车领域，新开发的订单派发策略所带来的改进非常有限，通常仅为 0.5% 至 2%（Tang 等，2019）。

3. **实验持续时间有限**进一步阻碍了对平均处理效应的准确估计。具体而言，时间序列实验通常仅持续数周（Bojinov 等，2023），导致 A/B 测试的样本量相对较小。

为应对后两个挑战，本文研究如何在考虑残留效应、从而解决第一个挑战的同时，谨慎设计时间序列实验，以提高平均处理效应估计的准确性。实验设计是统计学中的经典问题（Fisher，1935），其研究动机源于农业领域的应用，近年来也在机器学习领域日益受到关注（见第 2 节）。其目标是以最优方式生成实验数据，从而最大化所获取的信息量，实现准确估计。在 A/B 测试的背景下，这意味着在每个时间点决定实施哪种策略（处理组或对照组），使得根据所得实验数据估计的平均处理效应具有尽可能小的均方误差（MSE）。

我们的主要贡献在于开发了这样一种 A/B 测试设计，利用现代神经网络架构和先进的机器学习算法（我们的方案概要见图 1）：

• 从理论上讲，我们给出一个否定性结果，表明现有设计若将处理分配限制为仅取决于初始动作、当前观测或短期历史（图 2 展示了示意图），则可能并非最优。具体而言，我们建立了一项**不可能性定理**，证明在将范围限制为用于平均处理效应估计的双重稳健估计量（Tsiatis，2007）时，若分配策略未能充分利用完整历史，那么它们通常不可能实现最优，至少在渐近意义上如此。

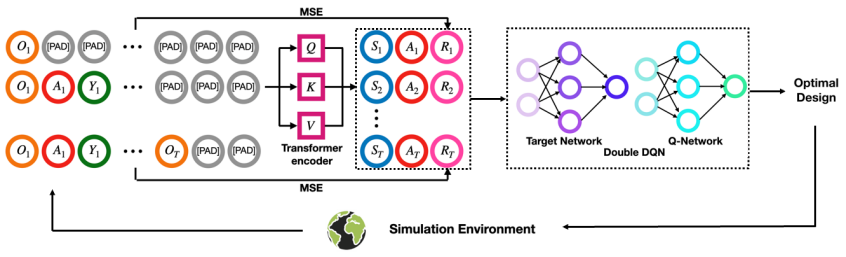


图 1：所提出的 Transformer 强化学习算法示意图。我们的算法使用 Transformer 编码器概括完整的历史上下文并生成状态$\{S_t\}_t$。随后，将这些状态输入双重深度 Q 网络智能体，由其输出使平均处理效应估计量（编码为（负）奖励$\{R_t\}_t$）的均方误差最小化的最优策略。

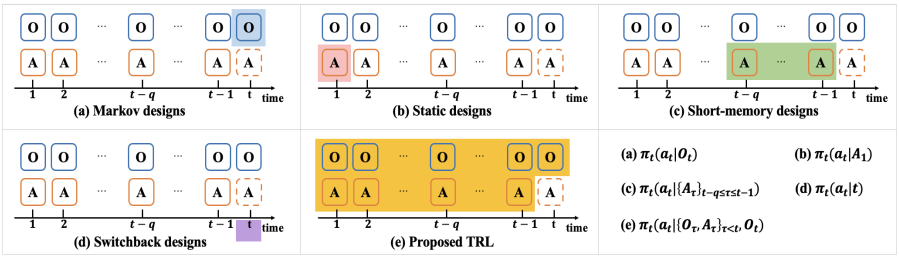


图 2：现有设计 (a)–(d) 和我们提出的设计 (e) 的处理分配示意图。具体而言，现有设计仅根据当前观测 (a)、初始动作 (b)、有限历史 (c) 或时间索引 (d) 来确定处理分配。相比之下，所提出的设计则依据完整的历史信息 (e) 进行处理分配。

• 基于这一洞见，我们采用**Transformer**（Vaswani 等人，2017），使每个时刻的策略决策都能依赖完整的数据历史，从而捕捉时间序列实验中固有的动态依赖。

• 我们接下来注意到，现有研究通常依赖较强的假设（详情见第 2.2 节），以将均方误差近似为设计优化的目标函数。为解决这一局限，我们采用**强化学习**（RL，Sutton 和 Barto，2018）直接最小化均方误差，而无需这些假设。

• 我们通过以下数据实证展示了所提出设计的实用性：(i) 合成数据；(ii) 一个公开可用的派单模拟器，该模拟器能够逼真地模拟司机和乘客的行为；以及 (iii) 从一家网约车公司收集的真实数据集。

总之，我们提出的方法能够处理残留效应和时间依赖关系，对数据生成过程的假设要求极少，并能取得显著的实证收益；这些优势均得益于现代计算能力的运用。

2 相关研究

Our 本文提出的方法与三个研究分支相关：A/B 测试、实验设计和强化学习优化。 r opti我们将在下文详细回顾各个分支的相关研究。

2.1 A/B 测试

关于 A/B 测试的文献非常丰富；近期的两篇综述可参阅 Larsen 等（2024）和 Quin 等（2024）。传统的 A/B 测试方法通过计算处理组与对照组在商品交易总额等各类业务指标上的均值差估计量来估计处理效应，以判断新策略是否优于现有策略

然而，正如引言中所述，这种方法忽略了延迟效应，在存在残留效应时会产生有偏估计。

处理残留效应主要有四种方法：

1. 第一种方法采用重要性抽样（参见例如 Precup et al., 2000；Zhang et al., 2013；Thomas et al., 2015；Guo et al., 2017；Bojinov & Shephard, 2019年；Hu & Wager, 2023；Zhou et al., 2025），该方法利用目标策略与生成实验数据的行为策略之比，对每个时间步的结果重新加权。然而，该估计量往往具有高方差——这一现象被称为时域诅咒（Liu et al., 2018；Xie et al., 2019年）。

2. 第二种方法以标准均值差估计量为基础，但通过在处理组与对照组切换后丢弃一部分数据来校正其偏差（Hu & Wager, 2022）。然而，该方法会降低有效样本量，在实验时长有限的情况下，这一问题可能尤为突出。

3. 第三种方法采用强化学习框架，将实验数据建模为马尔可夫决策过程（MDP，Puterman，2014），从而能够将现有的离策略评估算法（例如 Le 等，2019；Hao 等，2021；Shi 等，2021b；2022；Chen 和 Qi，2022；Uehara 等，2022；Bian 等，2025）应用于 A/B 测试；相关示例参见 Glynn 等人（2020）；Farias 等（2022）；Tang 等（2022）；Farias 等（2023）；Shi 等（2023b；2024）；Wen 等人（2025）。然而，这种方法要求时间序列满足马尔可夫假设，而在实践中这一假设可能不成立（Sun 等，2024）。

4. 最后一种方法通过使用部分可观测马尔可夫决策过程（POMDP）、粒子滤波器或更一般的时间序列模型来开展 A/B 测试，从而放宽马尔可夫假设（例如参见 Menchetti 等，2021；Sun 等，2024；Liang 和 Recht，2025；Ni 和 Bojinov，2025）。尽管如此，这些方法仍依赖特定的模型假设（例如线性、参数结构或短期依赖），这限制了它们捕捉实践中常见的复杂、非线性和长程时间依赖关系的能力。

上述大多数工作侧重于在给定实验数据的情况下估计平均处理效应，而我们关注的是一个不同的问题：如何设计并生成实验数据本身。尽管第三种方法与我们的方法都利用强化学习，但二者目标不同：前者将强化学习用作在残留效应下进行平均处理效应估计的建模框架，而我们将强化学习用作均方误差最小化的计算工具。

为应对较小的处理效应和有限的实验持续时间等其他挑战，已有许多算法被提出用于样本高效的 A/B 测试。除实验设计之外，其他方法还包括开发更准确的 ATE 估计量（Chen 与 Simchi-Levi, 2023；Chen 等, 2024；Sakhi 等, 2025）、功效更强的假设检验（Wang 等, 2025；Zhang 等, 2025b），以及从历史或同期实验数据中借用信息（Li 等, 2023b；Jung 与 Bellot, 2024；Li 等, 2024b；Wu 等, 2025）。

最后，我们指出还有另外两条研究方向：一条聚焦于序贯监测，研究如何在多个中期阶段开展 A/B 测试，同时控制总体第一类错误或错误发现（Johari et al., 2017; Yang et al., 2017; Shi et al., 2021a; Maharaj et al., 2023; Wan et al., 2023; Waudby-Smith et al., 2024; Lindon & Kallus, 2025）；另一条则关注如何处理残留效应之外的干扰效应，例如多个实验单位之间的溢出效应（Munro et al., 2021; Li et al., 2022; Li & Wager, 2022; Leung & Loupos, 2022; Shi et al., 2023a; Dai et al., 2024; Lu et al., 2024; Zhan et al., 2024）。

2.2 实验设计

实验设计在多个学科中得到了广泛研究，其应用范围涵盖生物学、医学、心理学和工程学等领域。下文将相关文献分为三类：(i) 统计学领域的早期研究，主要受临床试验应用的推动；(ii) 机器学习领域的最新进展，尤其是与强化学习相关的研究；以及 (iii) 管理科学、运筹学和其他相关领域的研究，重点关注科技行业中的现代 A/B 测试：

(i) 统计学早期研究提出了多种最优性准则（例如，D-最优性和 A-最优性），以及 Neyman 分配等最优设计及其推广形式，以优化协变量分布（例如，

Neyman，1934；de Aguiar 等，1995；Jones & Goos，2009；Yin & Zhou，2017；Loux，2013；Jones 等，2021；Li & Del Castillo，2024；Atkinson，2025；以及其中的参考文献）。我们的方案与临床试验中优化处理组分配策略的研究相关。此类研究已提出多种设计，包括协变量自适应设计（Lin 等人，2015；Zhou 等，2018；Bertsimas 等，2019；Ma 等，2024）、响应自适应设计（Hu & Rosenberger，2006；van der Laan，2008；Robertson 等，2023；Wei 等，2023；2025），以及协变量调整的响应自适应设计（Zhang 等，2007；Hu 等，2015；Lin 等，2015；Zhao 等，2022；Li 等，2024c）。然而，这些设计大多未考虑残留效应。交叉设计是一个例外（Hedayat & Stufken，2003；Li 等，2015；Louis 等，2019；Koner，2024）。但这类设计通常需要较长的洗脱期，从而降低有效样本量，因此在处理效应较小且时间有限的现代 A/B 测试中实用性受限。

（ii）近年来，实验设计在机器学习中日益受到关注（参见例如 Kato et al., 2024；Liu et al., 2024a；Weltz et al., 2024；Zhang & Wang, 2024；Pillai et al., 2025；Zhu et al., 2025）。具体而言，一系列研究利用深度学习和强化学习来数值计算最优设计（参见例如 Foster et al., 2021年；Blau et al., 2022；Lim et al., 2022；Annadani et al., 2024；Ai et al., 2025；Barlas & Salako, 2025）。我们的方法与这些工作具有相同的基本原理，但在两个方面有所不同。首先，这些工作考虑的是一般性设计问题，而我们专门聚焦于最小化 A/B 测试中平均处理效应估计量的均方误差。其次，我们提出利用 Transformer 神经网络架构，使处理分配策略能够以全部历史为条件，从而在时间依赖关系下实现更有效的设计（关于使用 Transformer 的理由，见第 3 节）。在强化学习文献中，该设计问题也被称为行为策略搜索（Hanna et al., 2017；Mukherjee et al., 2022；Liu & Zhang, 2024；Liu et al., 2025）。此外，经典的 D-最优设计已被用于马尔可夫决策过程中更高效的策略学习（Agarwal et al., 2019年，第 3.3 节）。

(iii) 关于技术行业 A/B 测试实验设计的研究日益增多，其中许多发表于 Management Science 和 Operations Research（例如，Ugander 等，2013；Viviano，2020；Bajari 等，2021；Wan 等，2022；Bajari 等，2023；Johari 等，2022；Leung，2022；Jia 等，2023；Ni 等，2023；Viviano 等，2023；Liu 等，2024b；Xiong 等，2024a；Zhao，2024a；b；Missault 等，2025；Xiong，2025；Yu 等，2025；Zhang 等，2025a）。我们的方案尤其与聚焦于时间序列实验的研究相关（Glynn 等，2020；Hu 和 Wager，2022；Basse 等，2023；Bojinov 等，2023；Chen 和 Simchi-Levi，2023；Li 等，2023a；Sun 等，2024；Xiong 等，2024b；Ni 和 Bojinov，2025；Wen等人，2025）。然而，这些研究存在两项局限。首先，正如我们的不可能性定理（定理 1）所示，处理分配时未根据完整历史进行条件化，可能导致次优结果。其次，这些研究需要强假设来简化均方误差的优化，例如：(a) 假设存在马尔可夫决策过程模型（Glynn 等，2020；Hu 和 Wager，2022；Li 等，2023a；Wen等人，2025）；(b) 对时间序列模型施加特定假设（Xiong 等，2024b；Sun 等，2024；Ni 和 Bojinov，2025）；(c) 假设延迟效应最多持续几个周期（Basse 等，2023；Bojinov 等，2023；Chen 和 Simchi-Levi，2023）。相比之下，我们的方案不需要这些假设。

2.3 用于优化的强化学习

强化学习（RL）是一种现代机器学习范式，用于学习最优策略，以便在序贯决策中最大化预期累积奖励。由于在策略优化方面行之有效，强化学习已被用作计算工具，以应对复杂的组合优化问题，包括经典的旅行商问题和背包问题（Bello 等，2016）、车辆路径规划、定向越野以及带奖赏旅行商问题（Kool 等，2018）、设备布局问题（Mirhoseini 等，2017；2018）和有向无环图约束（Zhu 等，2019）；有关全面综述，参见 Mazyavkina 等（2021）的综述。上述研究将解参数化为最优策略，并将奖励设为目标函数，从而使强化学习能够应用于优化。基于这一原理，我们提出将强化学习应用于最小化平均处理效应（ATE）估计量均方误差（MSE）这一具体问题，同时结合 Transformer 神经网络架构，有效捕捉时间序列实验中固有的时间依赖关系。

3 用于实验设计的 Transformer 强化学习

问题设定。考虑一家科技公司开展的一项在线实验，该实验持续T 个互不重叠的时间区间。在第 t个时间区间开始时，决策者观测到一组特征，记为 O_t。随后，他们决定在第 t个时间区间实施对照策略（-1）还是新策略（1），并以动作 $A_t \in \{-1, 1\}$表示。区间结束时，观测到一个数值结果 $Y_t \in \mathbb{R}$，数值越大表示结果越好。例如，在网约车平台上， O_t 可能包括用于描述第 t个时间区间开始前需求和供给的市场特征； A_t 可能表示订单派发、重新部署或补贴策略；而 Y_t 可以是已完成或已接单订单的百分比（完成率或响应率），也可以是司机收入。请注意，在时间序列实验中，数据具有时间依赖性：每个时点的结果 Y_t 和未来观测 O_{t+1} 可能不仅取决于当前的观测—动作对，还取决于过去的数据历史。我们将给定数据历史条件下它们的条件分布记为 $\mathcal{P}_t$，即，

$$\mathcal{P}_t(\mathcal{Y}, \mathcal{O} \mid H_{t-1}, O_t, A_t) = \mathbb{P}(Y_t \in \mathcal{Y}, O_{t+1} \in \mathcal{O} \mid H_{t-1}, O_t, A_t),$$

其中， $\mathcal{Y}$ 和 $\mathcal{O}$ 分别表示结果空间和观测空间的子集，而 $H_{t-1} = \{O_1, A_1, Y_1, \ldots, O_{t-1}, A_{t-1}, Y_{t-1}\}$ 表示截至时间 $t-1$的完整历史。

实验结束后，决策者收集观测-动作-结果三元组，以估计平均处理效应，其定义如下：

$$\mathrm{ATE} = \frac{1}{T}\sum_{t=1}^{T}\Big[\mathbb{E}_1(Y_t) - \mathbb{E}_{-1}(Y_t)\Big], \tag{1}$$

其中， $\mathbb{E}_1$ 和 $\mathbb{E}_{-1}$ 分别表示在假设情景下的期望：所有动作均设为新策略或对照策略。随后使用该估计量来确定部署哪种策略（新策略或对照策略）。正如第 2.1 节所述，基于实验数据进行平均处理效应估计已有大量相关研究。

相比之下，本文研究一个互补的设计问题：给定平均处理效应估计程序，在实验过程中应如何随时间分配动作，以最小化所得平均处理效应估计量的均方误差？从数学上讲，设 π_t 表示用于根据过去的数据历史生成 A_t 的处理分配策略，我们旨在确定使.....最小化的最优 $\pi = \{\pi_t\}_t$

$$\mathrm{MSE}(\pi) = \mathbb{E}_\pi(\widehat{\mathrm{ATE}} - \mathrm{ATE})^2, \tag{2}$$

whe $\widehat{\quad}$表示平均处理效应估计量， $\mathbb{E}_\pi$ 表示期望是对 dat a gen在 π 下生成的数据取的。

现有设计的局限性。有了这些表述，我们现在可以提出不可能性定理。为了在解析上计算均方误差，我们需要指定平均处理效应估计量。这里，我们将注意力限制在双重稳健估计量这一类上，因为它们广泛用于老虎机场景和序贯决策中的策略评估，无论是否满足马尔可夫假设（Jiang & Li, 2016；Chernozhukov 等，2018；Kallus & Uehara, 2022；Liao 等，2022）。在温和的正则性假设下（Chernozhukov 等，2018；Kallus & Uehara, 2022），这些估计量渐近无偏，因此其均方误差等于其渐近方差（记为 Var(π)），即

$$T\mathrm{MSE}(\pi) = T\mathrm{Var}(\pi) + o(1),$$

该估计量达到了半参数效率界——在大类正则且渐近线性估计量中可实现的最小均方误差（例如，Tsiatis，2007）。

$\widehat{\quad}$**定理 1**（不可能性定理）。 假设我们将 ATE 设为双稳健估计量。那么，存在一些数据生成过程 $\{\mathcal{P}_t\}_t$，使得对于所有 $1 \le t \le T$，最小化 Var（π）的最优策略 π 都取决于完整的过去历史，并且该最优策略是唯一的。

换句话说，对于一种处理分配策略，如果其中的 π_t 在任何时间步都不依赖于任何观测、动作或结果，那么该策略不可能是最优的。

如前文所述，现有设计存在两个局限：（i）其处理分配策略 π 不依赖完整历史，根据定理 1，这可能并非最优。（ii）为简化均方误差的计算，它们通常会对数据生成过程施加限制性假设。我们提供几个例子加以说明：

• 马尔可夫设计：Glynn 等人（2020）采用有限马尔可夫决策过程对实验数据进行建模。他们考虑的马尔可夫设计中，每个 π_t 仅通过当前观测 O_t 依赖于过去的历史，并通过求解一个简洁的凸优化问题，提出了一种学习最优马尔可夫设计的方法。

• 静态设计：Li 等人（2023a）将经典的内曼分配策略推广到马尔可夫决策过程。在他们的最优设计中，分配是静态的：每天的整个时间段内都分配相同的动作。因此，π_t 仅取决于初始动作。

• 短记忆设计：Sun 等人（2024）采用经典的 ARMA(p, q) 模型——一种特殊的部分可观测马尔可夫决策过程（POMDP）——来处理非马尔可夫环境。他们证明，在小信号假设下，最优的 π_t 仅取决于过去的 q 动作。与此同时，Ni 和 Bojinov（2025）将 π_t 限定为仅依赖于最近一次动作。

• 交替设计：Bojinov 等人（2023）和 Wen 等人（2025）研究了交替设计，即以固定时间间隔在对照组和新策略之间交替。Bojinov 等人（2023）表明，最优切换时长在很大程度上取决于残留效应持续的周期数。Wen 等人（2025）采用马尔可夫决策过程模型，允许残留效应贯穿整个实验，并表明最优切换时长取决于奖励的自相关性和残留效应的大小。在这两种情况下，π_t 仅取决于时间索引 t（对于 Wen 等人，2025，还取决于初始动作）。

总之，这些策略依赖于有限的历史信息——例如第一个动作、当前观测、时间索引或较短的历史——并施加马尔可夫决策过程或 ARMA 模型假设，或对残留效应施加约束。

我们的方案。为同时解决这两项局限，我们提出一种变换器强化学习（TRL）算法，用于最优设计在线实验。我们的方法整合了两个关键组成部分：（i）一个直接优化均方误差、而不依赖限制性建模假设的强化学习框架，以及（ii）一种基于变换器的神经网络架构，使处理分配能够以全部历史为条件。

（i）强化学习框架。我们假设可以访问一个仿真环境，该环境能够生成 $(O_t, A_t, Y_t)_t$，以近似实验数据生成过程（稍后我们将讨论如何构建这样的环境）。随后，我们应用强化学习与该环境交互，以优化设计。为应用强化学习，我们将状态-动作-奖励三元组定义如下：

• 状态：我们将状态 S_t 定义为截至时间点 t的完整历史 $\{O_1, A_1, Y_1, \ldots, O_{t-1}, A_{t-1}, Y_{t-1}, O_t\}$，以捕捉所有可能影响该时刻最优处理分配的潜在时间依赖关系；

• 动作：动作与 A_t相同，用于决定在时间 t实施哪种策略（标准策略或新策略）；

• 奖励：每个时间 t的奖励设定为

$$R_t = -\alpha^{T-t}\left[\widehat{\mathrm{ATE}}(t) - \mathrm{ATE}_{mc}\right]^2, \tag{3}$$

其中 ATE(t) 表示基于截至时间 t的全部数据三元组计算得到的平均处理效应估计量，ATE$_{mc}$表示从模拟器学习得到的蒙特卡罗估计量，$\alpha \in [0, 1)$ 表示折扣因子。

我们对奖励作几点说明。请注意，R_t 与 Y_t不同；它作为潜在均方误差的代理指标，而非公司的实际结果。具体而言，在运行强化学习之前，我们使用蒙特卡罗方法生成大量遵循标准策略或新策略的轨迹，并计算 ATE$_{mc}$，随后将其视为真实值，用以近似理论ATE。在强化学习过程中，在每个时间 t，我们使用截至时间 t 收集的观测-动作-结果三元组计算 ATE(t)，并度量其相对于 ATE$_{mc}$ 的平方偏差，作为均方误差的近似。随后将该平方误差按 α^{T-t} 进行折扣，以降低较早步骤的权重，因为这些步骤的 ATE 估计量因样本量较小而准确性较低。在 $\alpha = 0$的极端情形下，只有最后的奖励 R_T 起作用。最后，负号确保较小的均方误差产生较大的奖励。

在我们与网约车公司的合作中，有基于历史数据构建的物理模拟器，可用于模拟各种策略的效果。这些模拟器可以直接用作

强化学习的仿真环境。事实上，在开展在线实验之前，大多数策略都会使用此类模拟器进行离线评估，以展示其效果。如果没有可用的模拟器，则可以按顺序开展实验：每天结束后，我们利用收集到的实验数据估计数据生成过程并构建模拟器，然后应用所提出的流程来设计次日的实验。随着更多数据不断产生，我们会持续更新模拟器和实验设计，并重复这一过程。此外，我们可以结合实验数据与模拟器生成的额外合成数据来估计平均处理效应（参见 Lin 等人，2024；Xiong，2025）。我们提出的流程具有较强的通用性，因为平均处理效应估计量 ATE(t) 可以通过任何算法计算，包括此类数据整合方法。

(ii) 基于Transformer的DDQN。我们开发了一种双重深度 Q 网络的变体（Van Hasselt 等人，2016），利用 Transformer 架构来学习最优策略。具体而言，我们将 Q 函数定义为

$$Q_t(S_t, A_t) = \mathbb{E}_{\pi^{opt}}\left[\sum_{k=t}^{T} \gamma^{k-t} R_{k-t} \,\middle|\, S_t, A_t\right],$$

其中，π^{opt} 表示最优策略，即我们的目标处理分配策略，该策略相对于 Q 函数采取贪心选择。我们提出使用带掩码自注意力的 Transformer 来参数化 Q 函数。这种参数化方式有两个优点：(i) Q 函数的输入维度 S_t 会随时间变化，而 Transformer 能够轻松处理这种长度可变的输入，以编码完整历史；(ii) 与其他循环神经网络相比，Transformer 能够捕捉长期依赖关系（Zeng 等人，2023；Shi 与 Shide，2025），确保有效利用所有可能有助于制定最优策略的相关历史变量。

我们反复从仿真环境中采样轨迹，以学习 Q 函数。损失函数定义为 Q 函数与学习目标之间的平方损失。我们使用 AdamW 进行优化，并采用余弦学习率调度、梯度裁剪和混合精度训练，以提高效率。图 1 展示了我们流程的概要。

4 实验

本节通过数值实验评估我们提出的方法在有限样本下的性能。我们比较以下几种设计：

- **TRL**：提出的 Transformer 强化学习设计。
- **TMDP/NMDP**：Li 等人（2023a）针对马尔可夫决策过程提出的设计。这些设计每天切换处理组，并在每天内分配相同的处理组。
- **Switchback designs** 由 Hu 和 Wager（**HW**）、Bojinov、Simchi-Levi 和 Zhao（**BSZ**）、Xiong、Chin 和 Taylor（**XCT**），以及 Wen、Shi、Yang、Tang 和 Zhu（**WSY**）提出。这些设计每隔若干时间区间切换一次处理组。WSY 使用固定的切换间隔，而 HW、BSZ 和 XCT 则考虑切换间隔随机的常规切换回退设计。HW、BSZ 和 XCT 之间的区别在于所采用的平均处理效应估计量。这些设计都涉及一些超参数，我们考虑其最优配置。

我们在三个模拟环境中评估上述设计：(1) 合成模拟环境；(2) 基于真实数据的模拟环境；(3) 公开可用的派单模拟器。这些环境的详细信息见附录 A.2。对于每个环境，我们使用根据各项设计生成的数据，在 400 次模拟重复中计算 ATE 估计量的均方误差。

合成模拟器。我们考虑一种仿真环境，其中数据在不同天之间独立同分布。每天被划分为 M 个互不重叠的时间区间，在此期间，二维特征 O_t按照线性马尔可夫决策过程演变，结果 Y_t 则由线性奖励函数生成。我们改变天数 $n \in \{30, 35, 40, 45\}$，并设定区间数为 $M = 4$。我们考察设置 (i)–(iv)，改变方差和转移结构。转移系数随时间保持平稳，具体设定见附录 A.2。图 3 报告了设置 (i)-(ii) 下不同设计对应的各类平均处理效应估计量的经验均方误差及其置信区间（CI）。我们有以下三点观察：

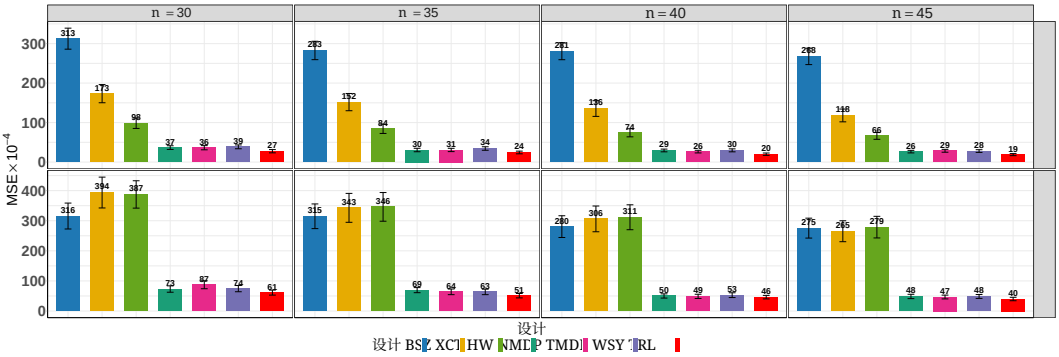


图 3：在合成环境中，不同设计在不同方差（设置（i））和转移结构（设置（ii））下的经验均方误差及其置信区间柱状图。

(a) 所提出的 TRL 总体上优于所有其他基线方法，在大多数设置下实现了最小的均方误差和最短的置信区间。(b) XCT、HW 和 BSZ 等交替设计在该环境中表现不佳，其均方误差显著高于其他设计。

（c）NMDP 和 TMDP 的表现与 TRL 相当，但均方误差略大。这是意料之中的，因为这些设计针对的是跨天具有独立同分布轨迹的设置，而每一天确实都遵循一个马尔可夫决策过程。尽管 TRL 具有理论最优性，TRL 的表现仍然更好，这可能得益于其 Transformer架构，该架构能够有效利用历史信息，并在有限样本下取得更优性能。

设置（iii）—（iv）的其他实验结果见附录中的图 6，呈现出类似的模式。

基于真实数据的模拟器。我们使用一家全球领先网约车公司的 A/A数据集来构建模拟器。该 A/A实验持续 40 天，在此期间始终采用单一订单派单策略。每天的轨迹数据被划分为 $M = 12$ 个互不重叠的时间区间，或 24 个互不重叠的时间区间。观测向量为二维，由以下两部分组成：(i) 订单请求数量；(ii) 上一个时间区间内司机的在线总时长，用于刻画双边市场的需求和供给。奖励定义为每个时间区间内司机获得的总收入。然而，原始数据集无法直接用于评估不同设计。为此，我们首先从 A/A数据集中估计数据生成过程，然后使用野生自助法（Wu等，1986）构建仿真环境。我们考虑一系列设置，其中天数 $n \in \{21, 28, 35, 42\}$、时间区间数量 $M \in \{12, 24\}$ 以及策略改进幅度（以平均处理效应与对照策略下平均结果之比衡量）从 $\{2.5\%、5\%\}$ 中选取，这反映了新开发调度策略通常产生的影响（Tang 等，2019）。

图4报告了策略改进幅度为 5% 时各种设计的经验均方误差及其置信区间。与之前一样，所提出的 TRL 设计始终优于所有其他设计，尤其是在 n 较小时。有趣的是，在这种环境下，TMDP 和 NMDP 得出的均方误差显著更大。这是因为在 A/A 数据集中观察到了结果之间的正相关性：在这种情况下，像 TMDP 和 NMDP 这样每天切换处理组的设计不再是最优的（Xiong 等，2024b；Wen等人，2025）。附录中的图 7 报告了策略改进幅度为 2。5% 时的结果，呈现出类似的模式。

公共调度模拟器。遵循 Xu 等人（2018）和 Li 等人（2023a），我们使用一个公开可用的实体调度模拟器[1]，它能够在共享出行市场中生成真实的乘客和司机行为模式，用于评估不同的设计。在该模拟器中，司机和订单在一个 9×9 空间网格上运行，每天包含 20 个时间步；若订单长时间未得到服务，则会被取消。我们将 Xu 等人（2018）的基于MDP的订单调度策略与一种基于距离的方法进行比较，后者旨在最小化

[1]https://github.com/callmespring/MDPOD。

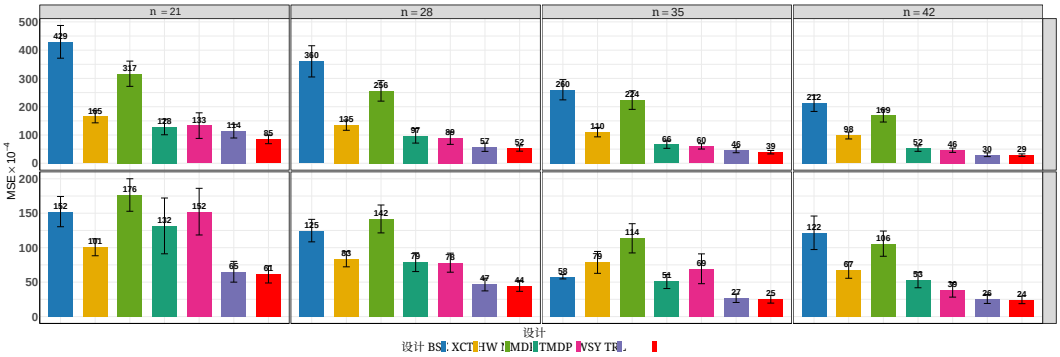


图4：基于真实数据的模拟中，不同设计的经验均方误差柱状图；新策略带来了 5% 的性能提升。

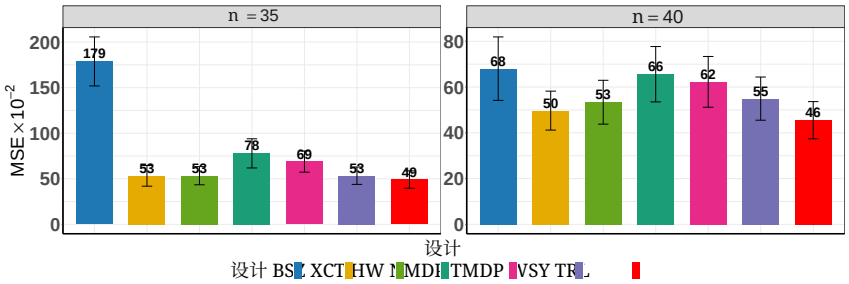


图5：调度环境中，不同天数下各设计的经验均方误差及其置信区间柱状图。

调度过程中司机与乘客之间的总距离。我们关注的结果是每个时间步的总收入，观测变量同样包括供给和需求。模拟持续 $n \in \{35, 40\}$ 天，每种设计均使用 100 个订单和一支由 25 名司机组成的固定车队进行评估。

图5报告了经验均方误差，表明所提出的 TRL 始终优于所有基线方法。随着 n 增加，TRL 的平均均方误差更低。交替设计通过交替处理组来减轻时空残留效应，从而降低偏差，因此其均方误差小于 NMDP 和 TMDP；而所提出的 TRL 则实现了最低的均方误差。

5 结论

本文研究了 A/B 测试中时间序列实验的设计。我们提出了一种新颖的基于 Transformer 的强化学习方法，使处理分配能够依赖于完整的数据历史，并放宽现有设计所需的严格建模假设。未来的工作可将我们的框架拓展至时间序列实验以外、更一般的实验设置，其中涉及多个实验单元，且单元之间可能出现溢出效应。

可复现性声明

我们采取了大量措施，以确保结果的可复现性。对于基于公开可用数据集的实验，我们将在补充材料中发布所有源代码，包括数据预处理、模型训练和评估脚本。

由于数据使用协议的限制，某些真实世界数据集无法公开共享。尽管如此，我们在附录 A.2 中详细描述了这些数据集的特征，并提供了相关模拟器的完整说明，以及附录 B.8 中真实世界数据集的可视化实证分析。

为减少随机性并实现客观评估，我们进行了 400 次独立运行，并报告了均方误差（MSE）及其置信区间（CI）。

实验在 NVIDIA A100（40GB）和 RTX 2080 图形处理器上进行，最快的合成数据实验在约三个图形处理器小时内完成训练。此外，我们还提供脚本，用于复现论文主体中基于公开数据或合成数据进行的实验所生成的所有图表。

伦理声明

这项工作不存在任何已知的伦理问题。所使用的数据集均为公开数据集或合成数据，且不包含个人身份信息。我们的方法旨在推动实验设计和因果推断研究，我们预计其不会被直接滥用。然而，与任何机器学习研究一样，这些方法仍存在被意外用于其他用途的潜在风险。我们鼓励未来的代码和方法使用者负责任地使用它们，并遵循伦理准则。

致谢

我们感谢匿名审稿人和元审稿人提出的建设性意见，这些意见显著改进了本文。
Xiangkun Wu 的研究得到了国家重点研发计划（项目编号：2024YFC2511003）的资助。
Qianglin Wen 的研究得到了国家重点研发计划（项目编号：2022YFA1003701）的资助。
Yingying Zhang 的研究得到了国家自然科学基金委员会（项目编号：12471280）的资助。
Ting Li 的研究得到了国家自然科学基金委员会（项目编号：12571304）、上海市浦江人才计划（项目编号：24PJIC030）、CCF–滴滴 GAIA 联合研究基金以及上海财经大学创新团队计划的资助。

参考文献

Alekh Agarwal、Nan Jiang、Sham M Kakade 和 Wen Sun。强化学习：理论与算法。*CS 系*，华盛顿大学西雅图分校，美国华盛顿州西雅图，技术报告，32：96，2019年。

Mingyao Ai、Holger Dette、Zhengfu Liu 和 Jun Yu。一种受多臂老虎机启发、针对模型不确定性的最优设计新方法。美国统计学会会刊，即将发表，2025。

YashasAnnadani、PanagiotisTigas、Stefan Bauer 和 Adam Foster。基于深度强化学习的摊销式主动因果归纳。神经信息处理系统进展，37：44216–44239，2024。

Susan Athey、Peter J Bickel、Aiyou Chen、Guido W Imbens 和 Michael Pollmann。随机实验中处理效应的半参数估计。英国皇家统计学会杂志 *B*辑：统计方法，85(5)：1615–1638，2023。

Anthony C Atkinson。最优实验设计。载于国际统计科学百科全书，第1835–1838页。施普林格，2025。

Patrick Bajari、Brian Burdick、Guido W Imbens、Lorenzo Masoero、James McQueen、Thomas Richardson 和 Ido MRosen。多重随机化设计。*arXiv* 预印本 *arXiv:2112.13495*，2021。

Patrick Bajari、Brian Burdick、Guido W Imbens、Lorenzo Masoero、James McQueen、Thomas S Richardson 和 Ido M Rosen。市场中的实验设计。*StatisticalScience*，38(3)：458–476，2023。

Yasir Zubayr Barlas 和 Kizito Salako。强化学习算法在序贯实验设计中的性能比较。*arXiv*预印本 *arXiv:2503.05905*，2025。

Guillaume W Basse、Yi Ding 和 Panos Toulis。考虑治疗习惯化的时间实验中因果效应的极小极大设计。*Biometrika*，110(1):155–168，2023。Irwan Bello、Hieu Pham、Quoc V Le、Mohammad Norouzi 和 Samy Bengio。结合强化学习的神经组合优化。*arXiv*预印本 *arXiv:1611.09940*，2016。Dimitris Bertsimas、Nikita Korolko 和 Alexander M Weinstein。在线临床试验中的协变量自适应优化。*OperationsResearch*，67(4):1150–1161，2019。Zeyu Bian、Chengchun Shi、Zhengling Qi 和 Lan Wang。双重非齐次环境中的离策略评估。美国统计学会会刊，即将发表，2025。Tom Blau、Edwin V Bonilla、Iadine Chades 和 Amir Dezfouli。利用深度强化学习优化序贯实验设计。载于国际机器学习会议，第2107–2128页。PMLR，2022。Iavor Bojinov 和 Neil Shephard。时间序列实验与因果估计目标：精确随机化检验与交易。美国统计学会会刊，114(528):1665–1682，2019。Iavor Bojinov、David Simchi-Levi 和 Jinglong Zhao。切换实验的设计与分析。*Management Science*，69(7):3759–3777，2023。Nicholas Chamandy。共享出行市场中的实验。网址：*https://eng. lyft. com/experimentation-in-a-ridesharing-marketplace-b39db027a66e*，2016。Haipeng Chen、Yan Jiao、Zhiwei Qin、Xiaocheng Tang、Hao Li、Bo An、Hongtu Zhu 和 Jieping Ye。Inbede：将上下文老虎机与时序差分学习相结合，用于网约车平台的联合定价与派单。载于*2019 IEEE*国际数据挖掘会议（*ICDM*），第61–70页。IEEE，2019。Hongyu Chen 和 David Simchi-Levi。利用替代变量开展高效切换实验：估计与实验设计。即将在*Management Science*发表，2023。Shuze Chen、David Simchi-Levi 和 Chonghuan Wang。对具有局部处理的马尔可夫决策过程开展实验。*SSRN 4906457*可查，2024。Xiaohong Chen 和 Zhengling Qi。关于离策略评估中非参数 q 函数估计的适定性与极小极大最优速率。载于国际机器学习会议，第 3558–3582 页。PMLR，2022。Victor Chernozhukov、Denis Chetverikov、Mert Demirer、Esther Duflo、Christian Hansen、Whitney Newey 和 James Robins。针对处理效应和结构参数的双重／去偏机器学习。*The Econometrics Journal*，21(1):C1–C68，2018。ISSN 1368-4221。Runpeng Dai、Jianing Wang、Fan Zhou、Shikai Luo、Zhiwei Qin、Chengchun Shi 和 Hongtu Zhu。针对空间或时空干扰下离策略评估的因果 Deep Sets。*arXiv*预印本 *arXiv:2407.17910*，2024。P Fernandes de Aguiar、B Bourguignon、MS Khots、DL Massart 和 R Phan-Than-Luu。D-最优设计。*Chemometrics and Intelligent Laboratory Systems*，30(2):199–210，1995。Peng Ding。因果推断入门。Chapman and Hall/CRC，2024。Zhixuan Fang、Longbo Huang 和 Adam Wierman。共享经济中的定价与补贴。载于第*26*届万维网国际会议论文集，第53–62页，2017。Vivek Farias、Andrew Li、Tianyi Peng 和 Andrew Zheng。实验中的马尔可夫干扰。神经信息处理系统进展，35:535–549，2022。Vivek Farias、Hao Li、Tianyi Peng、Xinyuyang Ren、Huawei Zhang 和 Andrew Zheng。实验中干扰的校正：抖音案例研究。载于第*17*届*ACM*推荐系统会议论文集，第455–466页，2023。

RA Fisher。《实验设计》。1935。

Adam Foster、Desi R Ivanova、Ilyas Malik 和 Tom Rainforth。深度自适应设计：摊销序贯贝叶斯实验设计。载于国际机器学习会议，第 3384–3395 页。PMLR，2021。Peter W Glynn、Ramesh Johari 和 Mohammad Rasouli。具有时间干扰的自适应实验设计：一种最大似然方法。神经信息处理系统进展，33:15054–15064，2020。Zhaohan Guo、Philip S Thomas 和 Emma Brunskill。利用选项和协方差检验进行长时域离策略策略评估。神经信息处理系统进展，30，2017。Josiah P Hanna、Philip S Thomas、Peter Stone 和 Scott Niekum。通过行为策略搜索实现数据高效的策略评估。载于国际机器学习会议，第 1394–1403 页。PMLR，2017。Botao Hao、Xiang Ji、Yaqi Duan、Hao Lu、Csaba Szepesvari 和 Mengdi Wang。通过自助法对拟合Q评估进行初始化，以实现离策略推断。载于国际机器学习会议，第 4074–4084 页。PMLR，2021。AS Hedayat 和 John Stufken。关于残留效应的不同假设下的最优且高效的交叉设计。生物制药统计学杂志，13(3):519–528，2003。Feifang Hu 和 William F Rosenberger。临床试验中反应自适应随机化理论。约翰·威立父子公司，2006。Jianhua Hu、Hongjian Zhu 和 Feifang Hu。一类基于效率和伦理的协变量调整反应自适应设计。美国统计学会会刊，110(509):357–367，2015。Yuchen Hu 和 Stefan Wager。几何混合下的切换实验。*arXiv*预印本 *arXiv:2209.00197*，2022。Yuchen Hu 和 Stefan Wager。序贯可忽略性假设下部分可观测马尔可夫决策过程的离策略评估。《统计年鉴》，51(4):1561–1585，2023。Guido W Imbens 和 Donald B Rubin。统计、社会与生物医学科学中的因果推断。剑桥大学出版社，2015。Su Jia、Nathan Kallus 和 Christina Lee Yu。聚类切换实验：时空干扰下的近最优速率。*arXiv*预印本 *arXiv:2312.15574*，2023。Nan Jiang 和 Lihong Li。强化学习的双重稳健离策略价值评估。载于国际机器学习会议，第 652–661 页。PMLR，2016。Yan Jiao、Xiaocheng Tang、Zhiwei Tony Qin、Shuaiji Li、Fan Zhang、Hongtu Zhu 和 Jieping Ye。利用深度强化学习实现真实场景中的网约车车辆重新调度。交通研究*C*辑：新兴技术，130:103289，2021。Ramesh Johari、Pete Koomen、Leonid Pekelis 和 David Walsh。窥探 A/B 测试：为何重要，以及该怎么做。载于第*23*届 *ACM SIGKDD*国际知识发现与数据挖掘大会论文集，第 1517–1525 页，2017。Ramesh Johari、Hannah Li、Inessa Liskovich 和 Gabriel Y Weintraub。双边平台中的实验设计：偏差分析。*Management Science*，68(10):7065–7791，2022。Bradley Jones 和 Peter Goos。裂区-裂裂区实验的D最优设计。*Biometrika*，96(1):67–82，2009。Bradley Jones、Katherine Allen-Moyer 和 Peter Goos。筛选实验的A最优设计与D最优设计。质量技术杂志，53(4):369–382，2021。Yonghan Jung 和 Alexis Bellot。跨多个不同实验数据集的高效策略评估。神经信息处理系统进展，37:136361–136392，2024。

Nathan Kallus 和 Masatoshi Uehara。用于马尔可夫决策过程离策略评估的高效双重强化学习。《机器学习研究杂志》，21(167):1–63，2020。Nathan Kallus 和 Masatoshi Uehara。通过双重强化学习高效破解离策略评估中的时间范围诅咒。*Operations Research*，70(6):3282–3302，2022。Masahiro Kato、Akihiro Oga、Wataru Komatsubara 和 Ryo Inokuchi。通过协变量选择进行处理效应估计的主动自适应实验设计。*arXiv* 预印本 *arXiv:2403.03589*，2024。Salil Koner。交叉设计统计方法进展的科学综述。*arXiv* 预印本 *arXiv:2410.08441*，2024。Wouter Kool、Herke Van Hoof 和 Max Welling。注意力机制，学会解决路由问题！*arXiv* 预印本 *arXiv:1803.08475*，2018。Nicholas Larsen、Jonathan Stallrich、Srijan Sengupta、Alex Deng、Ron Kohavi 和 Nathaniel T Stevens。在线对照实验中的统计挑战：A/B 测试方法综述。《美国统计学家》，78(2):135–149，2024。Hoang Le、Cameron Voloshin 和 Yisong Yue。约束条件下的批量策略学习。载于国际机器学习大会，第 3703–3712 页。PMLR，2019。Michael P Leung。针对空间干扰的速率最优整群随机设计。《统计年鉴》，50(5):3064–3087，2022。Michael P Leung 和 Pantelis Loupos。网络混杂条件下用于因果推断的图神经网络。*arXiv* 预印本 *arXiv:2211.07823*，2022。Hang Li 和 Enrique Del Castillo。黎曼流形上的最优实验设计。美国统计学会会刊，119(546):875–886，2024。Hannah Li、Geng Zhao、Ramesh Johari 和 Gabriel Y Weintraub。双边市场实验中的干扰、偏差和方差：平台指南。载于*ACM Web Conference 2022* 论文集，第 182–192 页，2022。Shuangning Li 和 Stefan Wager。微随机试验中的网络干扰。*arXiv* 预印本 *arXiv:2202.05356*，2022。Tianjing Li、Tsung Yu、Barbara S Hawkins 和 Kay Dickersin。交叉试验的设计、分析与报告：纳入荟萃分析的相关要求。*PloS one*，10(8):e0133023，2015。Ting Li、Chengchun Shi、Jianing Wang、Fan Zhou 等。用于序贯决策中高效策略评估的最优处理分配。神经信息处理系统进展，36:48890–48905，2023a。Ting Li、Chengchun Shi、Zhaohua Lu、Yi Li 和 Hongtu Zhu。动态条件分位数处理效应评估及其在网约车中的应用。美国统计学会会刊，119(547):1736–1750，2024a。Ting Li、Chengchun Shi、Qianglin Wen、Yang Sui、Yongli Qin、Chunbo Lai 和 Hongtu Zhu。结合实验数据与历史数据进行策略评估。载于第 *41* 届国际机器学习大会论文集，第 235 卷《机器学习研究论文集》，第 28630–28656 页。PMLR，2024 年 7 月 21–27 日。Xinyu Li、Wang Miao、Fang Lu 和 Xiao-Hua Zhou。利用外部对照数据提高临床试验推断效率。《生物统计学》，79(1):394–403，2023b。Yanping Li、Jingshen Wang 和 Waverly Wei。协变量调整的反应自适应随机化实验中的动态亚组识别。神经信息处理系统进展，37:85310–85339，2024c。Tengyuan Liang 和 Benja min Recht。当 n 等于 1 时的随机化推断。*Biometrika*，2025:asaf013，2025。doi: 10.1093/biomet/asaf013。

Peng Liao、Zhengling Qi、Runzhe Wan、Predrag Klasnja 和 Susan A Murphy。平均奖励马尔可夫决策过程中的批量策略学习。《统计年鉴》，50(6):3364，2022。

VincentLim、Ellen Novoseller、Jeffrey Ichnowski、HuangHuang 和KenGoldberg。基于策略的不可微隐式模型贝叶斯实验设计。*arXiv* 预印本 *arXiv:2203.04272*，2022。

Xi Lin、Jens Magelund Tarp 和 Robin J Evans。用于提高平均处理效应估计效率的数据融合：一篇结合模拟的实用综述。*arXiv* 预印本 *arXiv:2407.01186*，2024。

Yunzhi Lin、Ming Zhu 和 Zheng Su。追求平衡：临床试验中协变量自适应随机化技术概述。《当代临床试验》，45：21–25，2015。

Michael Lindon 和 Nathan Kallus。计数过程的任意时刻有效 A/B 测试。载于第 *28* 届国际人工智能与统计学会议，2025。

Pangpang Liu、Chengchun Shi 和 Will Wei Sun。用于基于人类反馈的强化学习的双重主动学习。*arXiv* 预印本 *arXiv:2410.02504*，2024a。

Qiang Liu、Lihong Li、Ziyang Tang 和 Dengyong Zhou。打破时域诅咒：无限时间范围的离策略估计。载于神经信息处理系统进展（*NeurIPS*），第 31 卷，2018。

Shuze Liu 和 Shangtong Zhang。利用离线数据指导行为策略设计，实现高效的策略评估。载于第 *41* 届国际机器学习会议论文集，第 32345–32368 页，2024年。

Shuze Liu、Claire Chen 和 Shangtong Zhang。用于强化学习的双重最优策略评估。载于 第十三届国际学习表征会议，2025。

Yang Liu、Yifan Zhou、Ping Li 和 Feifang Hu。簇自适应网络 A/B 测试：从随机化到估计。《机器学习研究杂志》，25(170)：1–48，2024b。

Thomas A Louis、W Lavori Philip、Marcia Polansky 等。临床研究中的交叉设计和自身对照设计。载于《统计学的医学应用》，第 83–103 页。CRC 出版社，2019年。

Travis M Loux. 一种简单、灵活且有效的协变量自适应处理分配方法。《医学统计学》，32(22):3775–3787，2013。

Xin Lu、Yuhao Wang 和 Zhiheng Zhang. 近似邻域干扰下辅助变量的调整。*arXiv* 预印本 *arXiv:2411.19789*，2024。

Shikai Luo、Ying Yang、Chengchun Shi、Fang Yao、Jieping Ye 和 Hongtu Zhu. 时间和/或空间依赖实验的策略评估。《英国皇家统计学会会刊*B*辑：统计方法学》，第 1–27 页，2024。

Wei Ma、Ping Li、Li-Xin Zhang 和 Feifang Hu。协变量自适应随机化程序及其性质：一个全新且统一的体系。《美国统计学会会刊》，119(545):151–162，2024。Akash Maharaj、Ritwik Sinha、David Arbour、Ian Waudby-Smith、Simon Z Liu、Moumita Sinha、Raghavendra Addanki、Aaditya Ramdas、Manas Garg 和 Viswanathan Swaminathan。企业级 A/B测试平台中的任意时刻有效置信序列。载于 *ACM Web Conference 2023* 配套论文集，第396–400页，2023。Nina Mazyavkina、Sergey Sviridov、Sergei Ivanov 和 Evgeny Burnaev。组合优化中的强化学习：综述。《计算机与运筹学》，134:105400，2021。Fiammetta Menchetti、Fabrizio Cipollini 和 Fabrizia Mealli。时间序列情境下干预因果效应的估计：C-ARIMA方法。*arXiv* 预印本 *arXiv:2103.06740*，2021。

Azalia Mirhoseini、Hieu Pham、Quoc V Le、Benoit Steiner、Rasmus Larsen、Yuefeng Zhou、Naveen Kumar、Mohammad Norouzi、Samy Bengio 和 Jeff Dean。运用强化学习优化设备放置。载于国际机器学习会议，第2430–2439页。PMLR，2017。

Azalia Mirhoseini、Anna Goldie、Hieu Pham、Benoit Steiner、Quoc V Le 和 Jeff Dean。设备放置的分层模型。载于国际学习表征会议，2018 年。

Paul Missault、Lorenzo Masoero、Christian Delb´e、Thomas Richardson 和 Guido Imbens。稳健高效的多单元切换实验。*arXiv* 预印本 *arXiv:2506.12654*，2025 年。

Subhojyoti Mukherjee、Josiah P Hanna 和 Robert D Nowak。Revar：通过降方差抽样强化策略评估。载于人工智能不确定性会议，第 1413–1422 页。PMLR，2022 年。

Evan Munro、Stefan Wager 和 Kuang Xu。市场均衡中的处理效应。*arXiv* 预印本 *arXiv:2109.11647*，2021 年。

Jerzy Neyman. 论代表性方法的两个不同方面：分层抽样方法与目的性选择方法。英国皇家统计学会会刊，97(4)：558–625，1934。

Tu Ni 和 Iavor Bojinov。提升切换实验的效率与鲁棒性：一种实用的模型辅助框架。可见于 *SSRN 5229804*，2025。

Tu Ni、Iavor Bojinov 和 Jinglong Zhao。考虑空间干扰和时间干扰的面板实验设计。可见于 *SSRN4466598*，2023。

Aditya Pillai、Gabriel Ponte、Marcia Fampa、Jon Lee、Mohit Singh 和 Weijun Xie。计算实验约束的D最优设计。载于*2025*年应用与计算离散算法会议（*ACDA*）论文集，第182–195页。SIAM，2025。

Martin Pouls、Anne Meyer 和 Nitin Ahuja。动态拼车中的空闲车辆重新部署。载于国际计算物流会议，第507–521页。施普林格，2020。

Doina Precup、Richard S Sutton 和 Satinder Singh。用于离策略策略评估的资格迹。2000年。Martin L Puterman。马尔可夫决策过程：离散随机动态规划。约翰·威立父子公司，2014年。Federico Quin、Danny Weyns、Matthias Galster 和 Camila Costa Silva。A/B 测试：系统性文献综述。《系统与软件杂志》，第 112011 页，2024年。David S Robertson、Kim May Lee、Boryana C L´opez-Kolkovska 和 Sof´ıa S Villar。临床试验中的响应自适应随机化：从误解到实践考量。《统计科学：数理统计学会评论期刊》，38(2):185，2023年。Otmane Sakhi、Alexandre Gilotte 和 David Rohde。通过离策略估计改进 A/B 测试的实践方法。*arXiv* 预印本 *arXiv:2506.10677*，2025年。Chengchun Shi、Shikai Luo、Hongtu Zhu 和 Rui Song。定性处理效应的在线序贯检验。《机器学习研究杂志》，22(286):1–51，2021a。Chengchun Shi、Runzhe Wan、Victor Chernozhukov 和 Rui Song。深度去偏离策略区间估计。载于国际机器学习会议，第 9580–9591 页。PMLR，2021b。Chengchun Shi、Sheng Zhang、Wenbin Lu 和 Rui Song。无限时域情形下强化学习价值函数的统计推断。《英国皇家统计学会会刊*B*辑：统计方法学》，84(3):765–793，2022年。

Chengchun Shi、Runzhe Wan、Ge Song、Shikai Luo、Hongtu Zhu 和 Rui Song。用于双边市场离策略评估的多智能体强化学习框架。应用统计年鉴，17(4):2701–2722，2023a。Chengchun Shi、Xiaoyu Wang、Shikai Luo、Hongtu Zhu、Jieping Ye 和 Rui Song。基于强化学习框架的 A/B 测试动态因果效应评估。美国统计学会会刊，118(543):2059–2071，2023b。Chengchun Shi、Jin Zhu、Ye Shen、Shikai Luo、Hongtu Zhu 和 Rui Song。混杂马尔可夫决策过程下的离策略置信区间估计。美国统计学会会刊，119(545):273–284，2024年。Tang Shi 和 Kazuya Shide。结合多维特征集成，对 LSTM、GRU 和 Transformer 模型进行比较分析，以预测建筑成本。《亚洲建筑与建筑工程杂志》，第 1–16 页，2025年。Ke Sun、Linglong Kong、Hongtu Zhu 和 Chengchun Shi。ARMA-Design：部分可观测时间序列实验中用于 A/B 测试的最优处理组分配策略。*arXiv* 预印本 *arXiv:2408.05342*，2024年。Richard S Sutton 和 Andrew G Barto。强化学习：导论。麻省理工学院出版社，2018年。X. Tang、Z. Qin、F. Zhang、Z. Wang、Z. Xu、Y. Ma、H. Zhu 和 J. Ye。基于深度价值网络的多司机订单调度方法。载于第*25*届*ACM SIGKDD*知识发现与数据挖掘大会（*KDD'19*），25:1780–1790，2019年。Ziyang Tang、Yiheng Duan、Stephanie Zhang 和 Lihong Li。利用强化学习估计长期处理效应的方法。*arXiv* 预印本 *arXiv:2210.07536*，2022年。Philip S Thomas、Georgios Theocharous 和 Mohammad Ghavamzadeh。高置信度离策略评估。载于第二十九届*AAAI*人工智能大会，2015年。Anastasios Tsiatis。半参数理论与缺失数据。施普林格科学与商业媒体，2007年。Masatoshi Uehara、Chengchun Shi 和 Nathan Kallus。强化学习中离策略评估综述。*arXiv* 预印本 *arXiv:2212.06355*，2022年。Johan Ugander、Brian Karrer、Lars Backstrom 和 Jon Kleinberg。图聚类随机化：面向多个总体的网络暴露。载于第*19*届*ACM SIGKDD*国际知识发现与数据挖掘大会论文集，第 329–337 页，2013年。Mark J van der Laan。自适应组序贯设计的构建与分析。2008年。Hado Van Hasselt、Arthur Guez 和 David Silver。基于双重 Q 学习的深度强化学习。载于*AAAI*人工智能大会论文集，第 30 卷，2016年。Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。注意力就是一切。神经信息处理系统进展，30，2017年。Davide Viviano。网络干扰下的实验设计。*arXiv* 预印本 *arXiv:2003.08421*，2020年。Davide Viviano、Lihua Lei、Guido Imbens、Brian Karrer、Okke Schrijvers 和 Liang Shi。因果聚类：网络干扰下聚类实验的设计。*arXiv* 预印本 *arXiv:2310.14983*，2023年。Stefan Wager。因果推断：一种统计学习方法，2024年。Runzhe Wan、Sheng Zhang、Chengchun Shi、Shikai Luo 和 Rui Song。用于订单调度强化学习的模式迁移学习。*arXiv* 预印本 *arXiv:2105.13218*，2021年。Runzhe Wan、Branislav Kveton 和 Rui Song。通过安全探索实现高效的策略评估与比较。载于国际机器学习会议，第 22491–22511 页。PMLR，2022年。

Runzhe Wan、Yu Liu、James McQueen、Doug Hains 和 Rui Song。实验平台与强化学习相遇：用于持续监测的贝叶斯序贯决策。载于第*29*届*ACM SIGKDD*知识发现与数据挖掘大会论文集，第5016–5027页，2023年。Jinjuan Wang、Qianglin Wen、Yu Zhang、Xiaodong Yan 和 Chengchun Shi。用于 A/B 测试的双臂老虎机框架。*arXiv*预印本 *arXiv:2507.18118*，2025年。Ian Waudby-Smith、Lili Wu、Aaditya Ramdas、Nikos Karampatziakis 和 Paul Mineiro。上下文老虎机的随时有效离策略推断。*ACM/IMS*数据科学期刊，1(3):1–42，2024年。Waverly Wei、Xinwei Ma 和 Jingshen Wang。公平自适应实验。神经信息处理系统进展，36:19157–19169，2023年。Waverly Wei、Xinwei Ma 和 Jingshen Wang。通过自适应实验学习处理效应异质性。《英国皇家统计学会会刊*B*辑：统计方法学》，第qkaf006页，2025年。Justin Weltz、Tanner Fiez、Alexander Volfovsky、Eric Laber、Blake Mason、Lalit Jain 等。异方差实验设计。神经信息处理系统进展，36，2024年。Qianglin Wen、Chengchun Shi、Ying Yang、Niansheng Tang 和 Hongtu Zhu。揭示切换实验中残留效应与奖励自相关之间的相互作用。载于第*42*届国际机器学习大会，2025年。Chien-Fu Jeff Wu 等。回归分析中的刀切法、自助法及其他重抽样方法。《统计年鉴》，14(4):1261–1295，1986年。Xiangkun Wu、Ting Li、Gholamali Aminian、Armin Behnamnia、Hamid R Rabiee 和 Chengchun Shi。用于策略评估的悲观数据整合。载于第*39*届神经信息处理系统年会，2025年。Tengyang Xie、Yifei Ma 和 Yu-Xiang Wang。利用边际化重要性采样实现强化学习的最优离策略评估。神经信息处理系统进展，32，2019年。Ruoxuan Xiong。基于历史数据模拟与优化的自动化实验设计。*SSRN 5126080*，2025年。Ruoxuan Xiong、Susan Athey、Mohsen Bayati 和 Guido Imbens。交错式发布的最优实验设计。*Management Science*，70(8):5317–5336，2024a。Ruoxuan Xiong、Alex Chin 和 Sean J Taylor。数据驱动的切换实验：理论权衡与经验贝叶斯设计。*arXiv*预印本 *arXiv:2406.06768*，2024b。Zhe Xu 、 Zhixin Li 、 Qingwen Guan 、 Dingshui Zhang、Qiang Li、Junxiao Nan、Chunyang Liu、Wei Bian 和 Jieping Ye。网约车平台的大规模订单派发：一种学习与规划方法。载于第*24*届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集，第905–913页，2018。Fanny Yang、Aaditya Ramdas、Kevin Jamieson 和 Martin Wainwright。具有在线错误发现率控制的多臂 A/B 测试框架。载于第*31*届神经信息处理系统国际会议论文集，第5959–5968页，2017年。Yue Yin 和 Julie Zhou。使用二阶最小二乘估计量的回归模型最优设计。《统计学新纪元》，第1841–1856页，2017年。Haoyang Yu、Wei Ma 和 Hanzhong Liu。具有溢出效应和残留效应的极小极大最优设计。*arXiv*预印本 *arXiv:2501.14602*，2025年。Ailing Zeng、Muxi Chen、Lei Zhang 和 Qiang Xu。Transformer是否适用于时间序列预测？载于*AAAI*人工智能大会论文集，第37卷，第11121–11128页，2023年。

Ruohan Zhan、Shichao Han、Yuchen Hu 和 Zhenling Jiang。推荐系统干扰下的处理效应估计：一种结构化神经网络方法。*arXiv*预印本 *arXiv:2406.14380*，2024年。

Baqun Zhang、Anastasios A Tsiatis、Eric B Laber 和 Marie Davidian。针对序贯治疗决策中最优动态治疗方案的稳健估计。*Biometrika*，100(3):681–694，2013。

Li-Xin Zhang、Feifang Hu、Siu Hung Cheung 和 Wai Sum Chan。协变量调整的响应自适应设计的渐近性质。《统计年鉴》，35(3):1166–1182，2007。

Qining Zhang、Tanner Fiez、Yi Liu 和 Wenyang Liu。固定预算下带验证的多指标自适应实验设计。*arXiv* 预印本 *arXiv:2506.03062*，2025a。

Yu Zhang、Shanshan Zhao、Bokui Wan、Jinjuan Wang 和 Xiaodong Yan。通过最大概率驱动的双臂老虎机进行策略性 A/B 测试。载于第四十二届国际机器学习会议，2025b。

Zhiheng Zhang 和 Zichen Wang. 网络干扰下兼顾估计与遗憾权衡的在线实验设计。*arXiv* 预印本 *arXiv:2412.03727*，2024。

Jinglong Zhao. 从优化视角看用于因果推断的实验设计。载于运筹学教程：为更美好世界作出更明智的决策，第 146–188 页。INFORMS，2024a。

Jinglong Zhao. 因果聚类的一种简单表述。可见 *SSRN 5008213*，2024b。

Wanying Zhao、Wei Ma、Fan Wang 和 Feifang Hu. 将协变量信息纳入精准医学自适应临床试验。制药统计学，21(1):176–195，2022。

Fan Zhou、Shikai Luo、Xiaohu Qie、Jieping Ye 和 Hongtu Zhu. 动态供需系统的基于图的均衡指标及其在网约车平台中的应用。美国统计学会会刊，116(536):1688–1699，2021。

Hongyi Zhou, Josiah P Hanna, Jin Zhu, Ying Yang 和 Chengchun Shi。揭示在离策略评估中使用估计的历史依赖行为策略进行重要性抽样的悖论。在第*42*届国际机器学习大会，2025。

Quan Zhou, Philip A Ernst, Kari Lock Morgan, Donald B Rubin 和 Anru Zhang。序贯再随机化。*Biometrika*，105(3):745–752，2018。

Jin Zhu, Jingyi Li, Hongyi Zhou, Yinan Lin, Zhenhua Lin 和 Chengchun Shi。平衡空间实验设计中的干扰与相关性：一种因果图割方法。在第*42*届国际机器学习大会，2025。

ShengyuZhu、IgnavierNg 和 ZhitangChen。利用强化学习进行因果发现。*arXiv* 预印本 *arXiv:1906.04477*，2019年。

大语言模型（LLM）使用声明

我们在撰写本文时使用了 **GPT-5 Pro** 作为辅助工具。其用途仅限于：

- **写作辅助**：纠正语法、重组句子并提高稿件的可读性（约占最终文本的 5–8%）。
- **代码优化**：为改进 Python 仿真脚本提供建议（例如张量运算、在图形处理器上进行并行化，以及改进可视化效果；约涉及全部代码的 10–15%）。
- **技术表述清晰度**：检查定义和公式的不同表述方式，以便更清晰地呈现。

大语言模型**未**用于：

- 生成新颖的研究构想,
- 构建数学证明,
- 设计实验，或
- 解读结果。

所有由大语言模型生成的建议均经过作者的仔细审查、核实和大幅修改。所有文本、代码和结论的最终责任完全由作者承担。

A 附录

A.1 定理 1 的证明

证明。令 H_{t-1} 表示截至时间 $t-1$的观测-动作-结果三元组集合。考虑满足以下三个条件的数据生成分布 $\{\mathcal{P}_t\}_t$：

条件 1（条件均值独立性假设（CMIA））。对于任意 t，给定 O_t、A_t 时，Y_t的条件均值独立于 H_{t-1}。

条件 2（条件独立性假设（CIA））。对于任意 t，给定 H_t 时，O_{t+1}的条件分布仅取决于 $\{O_j\}_{j\le t}$。

条件 3（历史依赖条件方差比（HCVR））。令 σ_t^2（H_t，O_t，A_t）表示给定 O_t、A_t 和 H_t 时 Y_t 的条件方差。则对于每个 t，σ_t（H_t，O_t，A_t）几乎必然为正，并且该比值

$$\frac{\sigma_t(H_t,O_t,1)}{\sigma_t(H_t,O_t,-1)}$$

取决于 H_t 和 O_t 中的所有变量。也就是说，这些变量的任何子集都不足以完全确定该比值。

构造数据生成过程 $\{\mathcal{P}_t\}_t$ 以满足这些假设很容易 s.

根据 Tsiatis (2007)；Kallus & Uehara (2020)，在 CMIA 和 CIA 下，双稳健估计量可以表示为以下形式：

$$\widehat{\text{ATE}} = \frac{1}{T}\sum_{t=1}^{T}\left[\mu_t(O_t,1)-\mu_t(O_t,-1)+\mathbb{I}(A_t=1)\frac{Y_t-\mu_t(O_t,1)}{\pi_t(1|H_t,O_t)}-\mathbb{I}(A_t=-1)\frac{Y_t-\mu_t(O_t,-1)}{\pi_t(-1|H_t,O_t)}\right],$$

其中 $\mu_t(o,\ a)\ =\ \mathbb{E}(Y_t|O_t\ =\ o,\ \ A_t\ =\ a)$。由 CMIA 可知，残差 $Y_t\ -\mu_t(O_t,\ A_t)$ 不相关，因此

$$\begin{aligned}
\text{Var}(\widehat{\text{ATE}}) &= \frac{1}{T^2}\sum_{t_1=1}^{T}\sum_{t_2=1}^{T}\left[\text{Cov}\left(\mu_{t_1}(O_{t_1},1)-\mu_{t_1}(-1,O_{t_1}),\mu_{t_2}(1,O_{t_2})-\mu_{t_2}(O_{t_2},-1)\right)\right]\\
&\quad+\frac{1}{T^2}\sum_{t=1}^{T}\left[\mathbb{E}\left(\mathbb{I}(A_t=1)\frac{Y_t-\mu_t(O_t,1)}{\pi_t(1|H_t,O_t)}\right)^2+\mathbb{E}\left(\mathbb{I}(A_t=-1)\frac{Y_t-\mu_t(O_t,-1)}{\pi_t(-1|H_t,O_t)}\right)^2\right]\\
&= \frac{1}{T^2}\sum_{t_1=1}^{T}\sum_{t_2=1}^{T}\left[\text{Cov}\left(\mu_{t_1}(O_{t_1},1)-\mu_{t_1}(O_{t_1},-1),\mu_{t_2}(O_{t_2},1)-\mu_{t_2}(O_{t_2},-1)\right)\right]\\
&\quad+\frac{1}{T^2}\sum_{t_1=1}^{T}\left[\mathbb{E}[\frac{\sigma^2(H_t,O_t,1)}{\pi_t(1|H_t,O_t)}]+\mathbb{E}[\frac{\sigma^2(H_t,O_t,-1)}{\pi_t(-1|H_t,O_t)}]\right].
\end{aligned} \tag{4}$$

通过关于分配概率最小化上述平均处理效应估计量的条件方差，我们可以得到每个时间 t处的最优分配 π_t^{opt}，其表达式为

$$\pi_t^{opt}(1|H_t,O_t)=\frac{\sigma_t(H_t,O_t,1)}{\sigma_t(H_t,O_t,1)+\sigma_t(H_t,O_t,-1)}.$$

当右端同时依赖于 O_t 和 H_t时，最优分配策略也同样如此。在HCVR下，对于任意 t，最优策略 π_t^{opt} 依赖于 O_t 和 H_t中的所有变量。
为了完成证明，我们证明 $\{\pi_t^{opt}\}_t$ 的唯一性。鉴于其唯一性，任何在任意时间步都不依赖任何历史变量的处理分配策略（其中 π_t ）都不可能是最优的。由此证明完成。

对于任意策略 π，由公式 4 可得

$$\mathrm{Var}(\pi) - \mathrm{Var}(\pi^{opt}) = \frac{1}{T^2}\sum_{t=1}^{T}\mathbb{E}\big[\phi_t\big(\pi_t; H_t, O_t\big)\big]\,, \tag{5}$$

其中

$$\phi_t(\pi; h, o) := \frac{\sigma_t^2(h,o,1)}{\pi} + \frac{\sigma_t^2(h,o,-1)}{1-\pi}, \qquad 0 < \pi < 1.$$

经过一些计算，我们得到

$$\phi_t\big(\pi_t^{\mathrm{opt}}; h, o\big) = \{\sigma_t(h,o,1) + \sigma_t(h,o,-1)\}^2.$$

由此可知，对于任意 $0 < \pi < 1$,

$$\begin{aligned}\Delta_t(\pi; h, o) &:= \phi_t(\pi; h, o) - \phi_t(\pi_t^{\mathrm{opt}}; h, o) \\ &= \frac{\big(\sigma_t(h,o,1)(1-\pi) - \sigma_t(h,o,-1)\,\pi\big)^2}{\pi(1-\pi)} \geq 0.\end{aligned} \tag{6}$$

鉴于 σ_t 在条件 3 中为正，当且仅当 $\pi = \pi_t^{opt}$ 时等式成立。

根据方程 5 和方程 6，对于任意分配规则 π，我们得到

$$\mathrm{Var}\{\widehat{\mathrm{ATE}}(\pi)\} - \mathrm{Var}\{\widehat{\mathrm{ATE}}(\pi^{\mathrm{opt}})\} = \frac{1}{T^2}\sum_{t=1}^{T}\mathbb{E}\big[\Delta_t\big(\pi_t(1 \mid H_t, O_t); H_t, O_t\big)\big] \;\geq\; 0,$$

其中，对于任意 t，当且仅当 $\pi_t = \pi_t^{opt}$ 时不等式成立。至此，最优策略唯一性的证明完成。

□

A.2 仿真实验详情。

本节将详细介绍模拟环境的状态转移规范，并报告更多实验结果。

合成模拟器（续）。在每一天 $i \in \{1, \ldots, n\}$ 开始时，从标准正态分布 $O_{i,\,1} \sim \mathcal{N}(0,\ I_2)$ 中抽取初始状态 $O_{i,\,1} \in \mathbb{R}^2$。此外，无论采用何种实验设计方法，不同天的初始状态 $\{O_{i,1}\}_{i=1}^n$ 都相互独立。对于每一天 i 和每个时间步 m，结果 $Y_{i,\,m}$ 和下一状态 $O_{i,\,m}$ 按照以下带有高斯噪声的线性转移方程演化：

$$Y_{i,m} = \alpha_m + \beta_m^\top O_{i,m} + \gamma_m A_{i,m} + \varepsilon_{i,m}^Y, \qquad \varepsilon_{i,m}^Y \sim \mathcal{N}(0, \sigma_y^2), \quad m = 1, \ldots, M, \tag{7}$$

$$O_{i,m} = \phi_m + \Phi_m O_{i,m-1} + \Gamma_m A_{i,m-1} + \varepsilon_{i,m}^O, \quad \varepsilon_{i,m}^O \sim \mathcal{N}(0, \sigma_o^2 I_2), \quad m = 2, \ldots, M. \tag{8}$$

其中，$\alpha_m \in \mathbb{R}$ 表示截距项，$\beta_m \in \mathbb{R}^2$ 是结果方程中状态 $O_{i、\,m}$ 对应的系数向量，$\Phi_m \in \mathbb{R}^{2\times 2}$ 是状态转移矩阵。结合 $\gamma_m \in \mathbb{R}$、$\phi_m \in \mathbb{R}^2$和 $\Gamma_m \in \mathbb{R}^2$，这些参数刻画了状态和动作对结果及状态转移的直接效应和间接效应。假设噪声项 $\{\varepsilon_{i,m}^Y, \varepsilon_{i,m}^O\}$ 在不同时间 m 和个体 i之间相互独立。

按照 Luo 等人（2024）的做法，真实的平均处理效应可定义为式（10）。生成的面板数据 $\{O_{i,m}, A_{i,m}, Y_{i,m}\colon\ 1 \leq i \leq n,\ \ 1 \leq m \leq M\}$ 可以展平为单条时间序列轨迹$\{O_t, A_t, Y_t\}t = 1^T$，其中 $t = (i-1)M + m$。结合我们对状态和奖励的定义，

可得 $\{S_t, A_t, R_t\}t = 1^T$。下一小节将介绍类似的过程。我们考虑$n \in \{30, 35, 40, 45\}$天和 $M = 4$。我们考虑三种设置。

设置（i）。

$$\alpha_m = 0, \qquad \beta_m = (0.6, 0.2)^\top, \qquad \gamma_m = 0.2, \qquad \text{for all } m = 1, \ldots, M,$$

and

$$\phi_m = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \qquad \Phi_m = \begin{bmatrix} 0.5 & 0.1 \\ 0.0 & 0.6 \end{bmatrix}, \qquad \Gamma_m = \begin{bmatrix} 0.1 \\ 0.05 \end{bmatrix}, \qquad \text{for all } m = 2, \ldots, M.$$

标准差设为 $\sigma_y = \sigma_o = 0.2$。（ii）。设置与（i）类似，但

$$\Phi_m = \begin{bmatrix} 0.6 & 0.2 \\ 0.5 & 0.6 \end{bmatrix},$$

用于考察我们的方法在不同动态下的表现，以及 $\sigma_y = \sigma_o = 0.3$。（iii）设置与（i）类似，但 $\sigma_y = \sigma_o = 0.3$。（iv）设置与（iii）类似，但$\beta_m = (0.3, 0.1)^\top$。

图 6 报告了设置（iii）-（iv）下的均方误差。结果表明，所提出的**TRL**通常比其他方法实现更低的均方误差。

算法 1基于自助法的模拟器

1：输入：真实数据 $\{(O_{i,m}, Y_{i,m}) : 1 \le i \le n,\ 1 \le m \le M\}$；调整参数 (δ_1, δ_2)；蒙特卡洛真值平均处理效应$_{\text{MC}}$；自助样本大小 n；随机种子；自助法重复次数 B。2：输出：合成时间序列轨迹 $\{(S_t^b, A_t^b, R_t^b)_{t=1}^T\}_{b=1}^B$，包含 B次重复。3：在模型方程 9 中计算最小二乘估计 $\{\widehat{\alpha}_m\}, \{\beta_m\}, \{\phi_m\}, \{\Phi_m\}$ 、处理组-效应参数 $\{\widehat{\gamma}_m\}, \{\Gamma_m\}$，以及方程 11 中的残差。4：对于 $b = 1$至 B，执行以下操作：5：从 $\{1, \ldots, N\}$中有放回地抽取 n天，并抽取 $\xi_i^b \sim \mathcal{N}(0, 1)$。6：使用基于Transformer的双重深度 Q 网络和 ε-贪心策略，在时间 t生成状态 S_t^b和动作 A_t^b，其中 $t = (i-1)M + m$。7：通过以下方式生成伪结果 $\{\widehat{Y}_{i,m}^b\}_{i,m}$ 和可观测状态 $\{\widehat{O}_{i,m}^b\}_{i,m}$：

$$\widehat{Y}_{i,m}^b = [1,\ (\widehat{O}_{i,m}^b)^\top,\ A_{i,m}^b] \begin{pmatrix} \widehat{\alpha}_m \\ \widehat{\beta}_m \\ \widehat{\gamma}_m \end{pmatrix} + \xi_i^b\, \widehat{e}_{i,m},$$

$$\widehat{O}_{i,m+1}^b = [\widehat{\phi}_m,\ \widehat{\Phi}_m,\ \widehat{\Gamma}_m] \begin{pmatrix} 1 \\ \widehat{O}_{i,m}^b \\ A_{i,m}^b \end{pmatrix} + \xi_i^b\, \widehat{E}_{i,m}.$$

8: 计算 $\{\text{平均处理效应}_{\text{SB}}^b\}_b$ 通过普通最小二乘法，并计算奖励 R_t^b 使用公式12。9: **结束 for**

基于真实数据的模拟器（续）。我们使用一个时间跨度为2019年5月17日至2019年6月25日、以一小时时间间隔记录的数据集，每天得到 $M = 24$个时间单位。依照Wen等人（2025，算法5）基于自助法的模拟流程，我们构建一个模拟器以生成合成实验数据。具体而言，我们对奖励函数和下一状态的期望值均应用线性模型，从而得到如下一组线性假设：

$$\begin{cases} \mu_t(O_m, A_m) = \alpha_m + O_m^\top \beta_m + \gamma_m A_m, \\ \mathbb{E}(O_{m+1} | A_m, O_m) = \phi_m + \Phi_m O_m + \Gamma_m A_m, \end{cases} \tag{9}$$

其中，α_m 和 γ_m 为实值，β_m，ϕ_m，且 Γ_m 为 $\mathbb{R}^d$中的向量，而 $\Phi_m \in \mathbb{R}^{d \times d}$。根据 Luo 等人（2024）所概述的模型方程 9，平均处理效应可表示为

$$\frac{2}{M} \sum_{m=1}^{M} \gamma_m + \frac{2}{M} \sum_{m=2}^{M} \beta_m^\top \Big[\sum_{k=1}^{m-1} (\Phi_{m-1} \Phi_{m-2} \ldots \Phi_{k+1}) \Gamma_k \Big], \tag{10}$$

其中，若 $m-1<k+1$，则将乘积 $\Phi_{m-1}\ldots\Phi_{k+1}$ 视为单位矩阵。方程 10 右侧的第一项表示动作对即时奖励的直接效应，而后一项则考虑了先前动作的延迟效应或残留效应。方程 10 为使用普通最小二乘法（OLS）回归来获得相关估计量提供了依据，随后将这些估计量代入方程 10，以计算最终的平均处理效应估计量。由此$\widehat{\quad}$ $\widehat{\quad}$ $\widehat{\quad}$得到估计量 $\{\widehat{\alpha}_m\}_m$、$\{\beta_m\}_m$、$\{\phi_m\}_m$ 和 $\{\Phi_m\}_m$。然而，$\{\gamma_t\}_m$ 和 $\{\Gamma_m\}_t$ 仍不可识别，因为 $A_m=0$几乎必然成立。然后，我们根据这些估计量计算奖励回归模型和可观测状态回归模型中的残差，如下所示：

$$\widehat{e}_{i,m}=Y_{i,m}-\widehat{\alpha}_m-O_{i,m}^{\top}\widehat{\beta}_m,\quad \widehat{E}_{i,m}=O_{i,m+1}-\widehat{\phi}_m-\widehat{\Phi}_t O_{i,m}. \tag{11}$$

为了生成处理效应大小各异的模拟数据，我们引入处理效应比参数 λ，并手动设置处理效应参数 $\widehat{\gamma}_m=\delta_1\times(\sum_i Y_{i,\ m}/N)$ 和$\widehat{\Gamma}_m=\delta_2\times(\sum_i O_{i,\ m}/N)$。处理效应比实质上对应于平均处理效应与基线策略平均回报的比值。处理效应比取值为 $\{2.5\%,\ 5\%\}$。

最后，为了创建跨度为 n 天的合成数据集，我们将其展平为一条以 $t=1$、.....、T 为索引的单一时间序列轨迹，其中 $T=nM$和 $t=(i-1)M+m$分别对应 $i\in[n]$和$m\in[M]$。状态由截至时间点 t 的完整历史确定，动作则根据基于Transformer的DDQN及epsilon-贪婪策略生成。然后，我们抽取独立同分布的标准高斯噪声 $\{\xi_i\}_{i=1}^n$。对于第 i 天，我们均匀抽取一个整数 $I\in\{1,\ldots,N\}$，将初始观测设为$O_{I,\ 1}$，并根据公式 9，利用估计得到的$\widehat{\quad}$ $\widehat{\quad}$ $\widehat{\quad}$ $\widehat{\quad}$ $\{\widehat{\alpha}_m\}_m$、$\{\beta_m\}_m$、$\{\phi_m\}_m$、$\{\Phi_m\}_m$，指定的 $\{\widehat{\gamma}_m\}_m$和 $\{\Gamma_m\}_m$，以及分别由 $\{\xi_i\widehat{e}_{i,m}:\ 1\le m\le M\}$ 和 $\{\xi_i E_{i,m}:\ 1\le m\le M\}$给出的误差残差$\widehat{\quad}$，生成奖励和状态。这确保合成数据的误差协方差结构与真实数据集的误差协方差结构十分相似。根据模拟数据，我们使用普通最小二乘法在每天结束时估计平均处理效应。根据公式 3 中的定义，随后计算非零代理奖励

$$R_t=\begin{cases}-\alpha^{\,n-i}\big[\widehat{\mathrm{ATE}}(t)-\mathrm{ATE}_{\mathrm{mc}}\big]^2, & \text{if } t=iM,\\ 0 & \text{otherwise,}\end{cases} \tag{12}$$

其中，蒙特卡洛真值平均处理效应$_{\mathrm{mc}}$ 可以通过分别在对照策略和新策略下生成大型数据集来预先估计。在时间范围 T内，我们收集完整轨迹 $\{S_t,A_t,R_t\}_{t=1}^T$，并用其训练基于Transformer的DDQN。自助法辅助程序的概要见算法 1。

图 7 报告了真实处理效应比为 2.5% 时的均方误差。与所有基线方法相比，所提出的 TRL 的平均均方误差最低，且平均均方误差随 n 增加而降低。

公共调度模拟器（续）。 在本例中，我们参照 Xu 等人（2018），对环境进行更详细的描述。我们考察司机与订单在持续 20 个时间步的 9×9 空间网格中的交互。司机每个时间步只能垂直或水平移动一个网格，而订单只能派发给曼哈顿距离不超过 2 的司机。如果订单长时间未分配给任何司机，就会被取消。取消时间服从均值为 2.5、标准差为 0.5 的截断高斯分布，其在时间轴上的取值范围为 0 到 3。为生成模拟早高峰和晚高峰的真实交通模式，我们分别对住宅区和工作区建模，并使用双分量高斯混合分布对订单的起始位置进行采样。随后，将这些位置截断为时空网格内的整数。订单目的地和司机初始位置则从网格上的离散均匀分布中随机采样。高斯混合的参数如下。

$$\pi^{(1)}=1/3,\pi^{(2)}=2/3,\mu^{(1)}=[3,3,2],\mu^{(2)}=[6,6,14],\sigma^{(1)}=[2,2,2],\sigma^{(2)}=[2,2,2],$$

这三个维度分别对应空间中的水平坐标、垂直坐标和时间坐标。

这里，我们将系统的汇总统计量，即每个时间步的活跃订单数和活跃司机数，视为一个二维状态变量 O。活跃订单数是指满足可行性约束（例如尚未过期）的订单数；活跃司机数是指处于空闲状态、因而可以接受新订单的司机数。动作 A（调度决策）有两个取值：1 和 -1，其中 $A=1$对应基于马尔可夫决策过程的策略，$A=-1$ 对应基于距离的策略。具体而言，

1 表示基于*MDP*的动作，而 -1 表示基于距离的动作。基于MDP的动作对应一种派单策略，该策略在考虑当前匹配的未来收益的基础上，选择订单以最大化司机长期预期收入之和。相比之下，基于距离的动作对应于选择一种分配方案，使时间 m的所有匹配的接客距离总和最小。结果 Y_m 定义为时间 m所有已匹配订单的总收入。我们首先为每位司机和每个订单生成信息，然后应用Sinkhorn算法，获得使整体分配目标最大化的最优匹配。基于这些生成的大规模数据，我们学习与每位司机动作相关的长期价值函数 V。随后，我们使用学得的价值函数 V构建用于订单分配的基于MDP的派单策略。基于学得的基于MDP的策略和基于距离的策略，我们生成大量样本。利用这些样本，我们通过神经网络模型拟合状态转移动态。

$$\hat{\varepsilon}^{Y}_{i,m} = Y_{i,m} - \hat{Y}_{i,m}, \quad \hat{Y}_{i,m} = \hat{f}_t(O_{i,m}, A_{i,m}), \quad m = 1, \ldots, M, \tag{13}$$

$$\hat{\varepsilon}^{O}_{i,m} = O_{i,m} - \hat{O}_{i,m}, \quad \hat{O}_{i,m} = \hat{g}_t(O_{i,m-1}, A_{i,m-1}), \quad m = 2, \ldots, M. \tag{14}$$

其中 f_m 和 $\hat{}g_m$ 表示通过最小化平方误差损失训练得到的神经网络预测器。我们在每个时间步计算所有样本的残差，并估计其均值和方差。随后，我们假设残差服从具有相应均值和方差的高斯分布。基于此，我们构建一个模拟器，在其中，神经网络 f_m 和 $\hat{}g_m$ 的预测值会受到高斯噪声的扰动，并对输出进行取整和截断，以更好地模拟现实场景（活跃司机和活跃订单的数量均为整数）。利用该模拟器，我们在 all$+1$ 和 all-1 动作序列下生成 20，000 个蒙特卡洛样本，以获得真实值 ATE。对于实验设计方法，我们的 TRL 估计量采用最小二乘时序差分（LSTD）方法估计 ATE。这与前两个实验有所不同，因为当前环境并非线性，因此需要使用泛化性更强的 LSTD 方法。

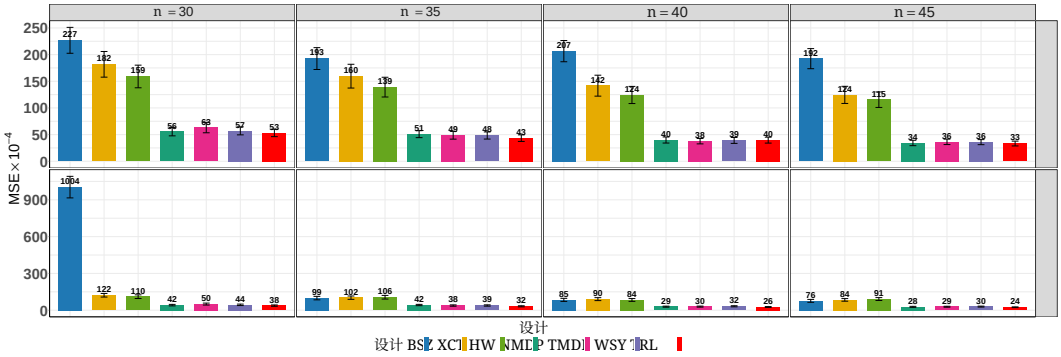


图 6：合成模拟器中的设置 (iii) 和 (iv) 下，MSE 的经验均值柱状图，含 $M = 4$。

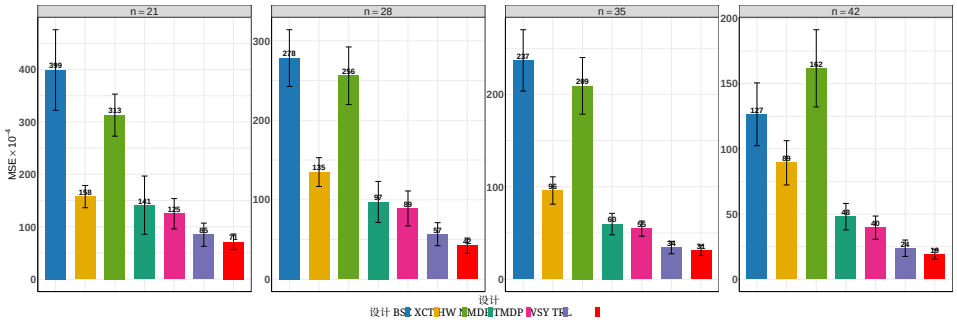


图 7：在包含 $M = 12$ 的基于真实数据的模拟中，不同设计下的经验均方误差柱状图；新策略带来了 2.5% 的性能提升。

表 1：TRL、长短期记忆网络与不同注意力窗口大小之间的经验均方误差。

TRL	LSTM	窗口大小 6	窗口大小 12	窗口大小 24
0.002750	0.002975	0.003002	0.002915	0.02871

B 附加实验

本节将介绍更多实验，以进一步支持我们的方法并提供更全面的实证评估，包括消融研究、对超参数和模拟器敏感性及鲁棒性的分析、对泛化性和计算复杂度的评估、对实际模拟器可用性的评估、基于真实 A/A 数据的实证相关性分析，以及在不同自相关水平下开展的敏感性实验。

B.1 不同注意力窗口大小下 LSTM 模型的消融实验

我们在设置（i）下进行实验，将我们的全历史 Transformer（TRL）与两项关键消融方案进行比较：（1）训练预算相同、基于 LSTM 的双重深度 Q 网络；（2）注意力窗口受限的 Transformer 变体。

表1的结果表明，基于 LSTM 的序列建模已显著优于 TMDP 等传统方法（均方误差：0.0036）。我们的全历史 Transformer（TRL）进一步提升了性能，均方误差低于 LSTM。此外，我们观察到，对于 Transformer 变体，增大注意力窗口大小始终能够提升性能，这表明利用更长距离的时间依赖关系具有优势。

B.2 预热方案的消融实验

我们的仿真协议包含一个“预热”期，在此期间，前几天的奖励设为零，使学习仅在均方误差估计变得更加稳定后才开始。我们使用基于真实数据的模拟器（$n,\ M$）$=$(35，12) 和 5% 的策略提升，针对预热长度为 2、5、7 和 14 天的情况进行了消融实验。如表2所示，较短的预热期会导致明显更高的均方误差，表明嘈杂的早期奖励会削弱学习效果。7 天预热期的表现最佳（而非 14 天），在消除不稳定的早期估计与保留充足训练数据之间取得了有效平衡。

表2：不同设计下的均方误差（$\times 10^{-4}$）

设计	**BSZ**	**HW**	**XCT**	**NMDP**	**TMDP**
MSE	260.0	224.0	110.0	66.2	59.9
设计	**WSY**	**TRL（预热=2）**	**TRL（预热=5）**	**TRL（预热=7）**	**TRL（预热=14）**
MSE	46.5	42.6	40.1	**37.9**	39.2

注：400 次模拟中的均方误差（$\times 10^{-4}$）。采用 7 天预热的 TRL 实现了最低均方误差，证实稀疏奖励（更长的预热期）优于密集奖励（更短的预热期）。最佳预热时长为 7 天，而不是 14 天。

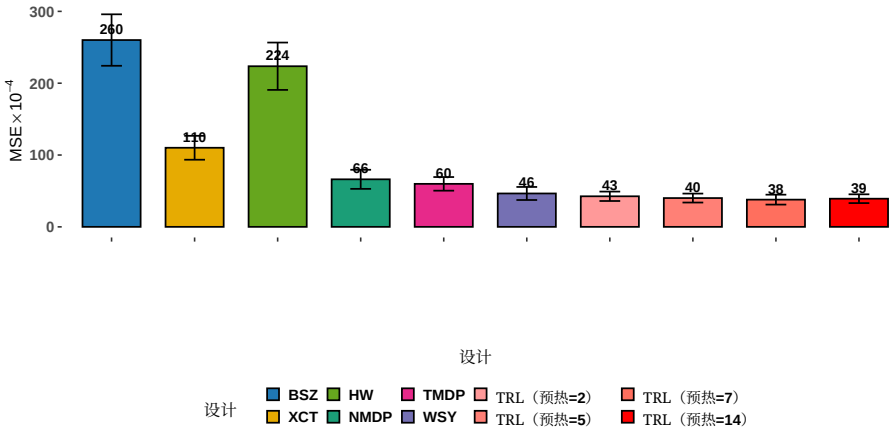


图 8：不同预热时长下的均方误差（$\times 10^{-4}$）。较短的预热时间（2、5 天）$\rightarrow$ 均方误差较高，验证了稀疏奖励。

B.3 模拟器的敏感性与鲁棒性

为评估算法对模拟器设定错误的鲁棒性，我们基于第 4 节合成模拟器中的设置（i）进行了一系列额外实验。回顾一下，在没有任何设定错误的情况下，TRL 的均方误差为 0.0027，而最佳基线 TMDP 的均方误差为 0.0036。随后，我们引入了五种类型的偏移来对模拟器进行设定错误处理：

- **噪声偏移**。 我们提高公式 8 中的噪声水平，从 $\sigma_y = \sigma_o = 0.2$ 增至 0。25，使均方误差达到 **0.0028**。
- **协变量偏移**。 公式 8 中的初始协变量$O_{i,\ 1}$ 原本服从二元标准正态分布。我们将均值平移至 0。1 和 0。3，得到的均方误差分别为 **0.0028** 和**0.0030**。
- **转移偏移**。 我们通过将方程 8 中的 2×2 转移矩阵 Φ_m 的最后一个元素从 0.60 改为 0.55，对其进行修改，得到的均方误差为**0.0030**。
- **效应系数偏移**。 我们将方程 7 中的 β 效应系数从 0.20 增加到 0.25，得到的均方误差为 **0.0030**。
- **奖励偏移**。 我们在方程 7 中向 Y_t $+0.1$ 添加一个常数，以模拟奖励偏移，得到的均方误差为**0.0029**。

在全部五种情境中，与未发生设定错误的原始设置相比，TRL 的性能仅略有下降，并且始终优于最佳基线。这些结果表明，即使模拟器与真实环境存在中等程度的偏差，TRL 仍能保持出色的性能。

B.4 超参数的敏感性与鲁棒性

我们的方法对超参数的鲁棒性至关重要。我们采用合成模拟器中的设置（i），以（$M = 4$, $n = 30$）作为基础配置，每次改变一个超参数，以创建不同的设置。表 3 报告了在我们的设计下各估计量的均方误差，结果表明，随着超参数变化，我们的方法性能存在一定波动，但所有结果仍接近论文正文中报告的均方误差 **0.0027** 。与此同时，其性能仍明显优于最佳基线 TMDP，后者的均方误差为 **0.0036**。我们得出结论：在一系列关键超参数取值范围内，我们的方法性能稳定，并始终优于最佳基线。

表 3：不同超参数下的经验均方误差。

超参数	**值 1**	**值 2（正文）**	**值 3**
折扣因子 (α)	0.0030 ($\alpha = 0.9$)	**0.0027**($\alpha = 0.8$)	0.0027 ($\alpha = 0.7$)
Transformer 宽度	0.0027 (32)	**0.0027** (64)	0.0030 (128)
目标网络更新率	0.0027 (0.004)	**0.0027** (0.005)	0.0029 (0.006)
探索率 (ϵ)	0.0029 ($\epsilon = 0.03$)	**0.0027**($\epsilon = 0.10$)	0.0029 ($\epsilon = 0.15$)
最佳基线 (TMDP)		**0.0036**	

B.5 大时间跨度

为评估所提出方法在更长时间范围 T 下的性能，我们在合成仿真的设置（i）中开展了额外实验。我们固定 $n = 30$，并令 $M \in \{4, 8, 12, 16\}$，其中 $T = nM$。

表4报告了在较大 T 值下不同方法的结果。随着时间范围增加，我们的方法始终优于所有竞争基线，展现了其对较长实验时间跨度的鲁棒性。

表4：不同时间范围下 400 次重复试验的经验均方误差（$\times 10^{-4}$）。

方法	**T=120**	**T=240**	**T=360**	**T=480**
BSZ	320	253	213	182
XCT	173	296	240	171
HW	98	317	256	173
NMDP	37	47.3	70.4	77.8
TMDP	36	45.3	70.1	77.9
WSY	39	55.7	64.7	71.2
TRL	27	41	55.2	69

B.6 计算复杂度

我们报告了实验中的**实测运行时间**。在设置（i）的合成模拟器中，我们固定 $n = 30$ 并改变 $M \in \{4, 8, 12, 16\}$，记录在 NVIDIA 2080 GPU 上的运行时间。在此配置下，$n = 30$ 和 $M = 4$ 的实验耗时略多于 3 小时，见表5。不过，策略训练完成后，为新的 A/B 测试生成实验设计非常迅速：在我们的实验中，部署时间与现有基线方法相当，通常不到一分钟。尽管我们的方法训练起来更耗时，但其估计准确度高于竞争方法，在实际应用中可能具有很高的价值。

表5：不同时间范围下的训练时间（1000 个训练轮次）T。

时间范围 T	运行时间（小时）
120	3:34:50.529
240	12:18:55.589
360	26:39:10.679
480	45:53:16.918

B.7 实际应用中的模拟器可用性评估

当不存在模拟器时，我们继续使用设置（*i*）在合成模拟器中进行实验，并探索在 TRL 的表现显著优于简单切换实验之前所需的样本量（天数）以及持续时间。根据表6，我们发现，当预热期长度约为9天时，其表现与切换实验相当；达到11天时，则明显胜过切换实验。因此，所需的9天预热期相对于30天的实验时长而言较短。

表6：不同预热天数对经验均方误差的影响

切换回退	第=5天	第=7天	第=9天	第=11天
0.003855	0.004419	0.00439	**0.003899**	**0.003704**

B.8 真实数据的经验相关性分析

为验证真实数据中的高非平稳性和长时域依赖，我们提供了图表，展示主文所述的基于真实数据的模拟器中需求与供给的经验自相关和交叉相关函数，其中 n = 35、M = 12，且新策略带来了 5% 的性能提升。

我们在图9a和图9b中展示了 GMV、订单请求数量和在线总时长的自相关函数及跨时点相关性，并在图9c中展示了订单请求与司机在线总时长之间的交叉相关性。图9a显示，三个变量均呈现较强的滞后1期和滞后2期正自相关，随后在滞后3至5期出现负自相关。图9b显示，订单请求与司机在线总时长之间存在非零且以正相关为主的交叉相关性。此外，图9c中的大多数跨时点相关性也为正。这些动态违反了 TMDP/NMDP 所依赖的马尔可夫性和平稳性假设，而这两种方法难以捕捉此类非平稳的长时域依赖。相比之下，TRL 利用完整的历史轨迹，自适应地建模不断演变的关系。

B.9 不同相关性水平下的敏感性实验

为验证 TRL 在不同相关强度下的优势，我们首先通过以下系数系统地缩放供需交叉相关性：$\{0.25、0.5、1.0、1.5\}$，其中 1.0 对应主设置，具体考虑

$$\mathbb{E}[O_{m+1,1} \mid A_m, O_m] = \Phi_{m,11} O_{m,1} + (\phi_{\text{coef}} \Phi_{m,12}) O_{m,2} + \Gamma_{m,1} A_m,$$

$$\mathbb{E}[O_{m+1,2} \mid A_m, O_m] = \Phi_{m,22} O_{m,2} + (\phi_{\text{coef}} \Phi_{m,21}) O_{m,1} + \Gamma_{m,2} A_m.$$

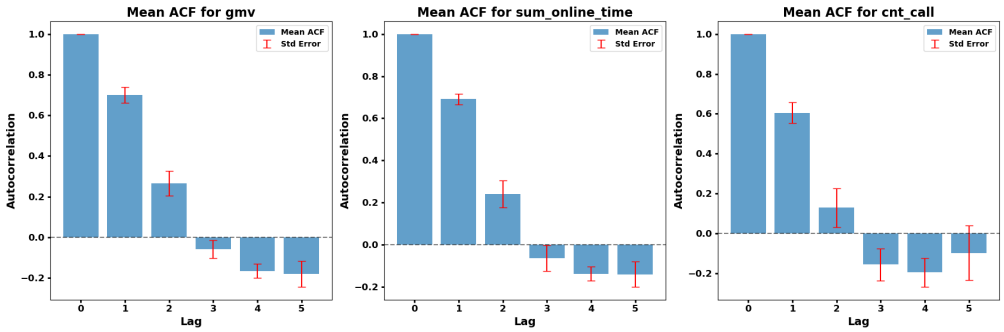
随着相关性增强，TRL 相对于 TMDP/NMDP 的均方误差**有所下降**（图 10 和表 7），这表明其优势恰恰在时间依赖性最强且最非平稳时增大，从而证实 TRL 在历史上下文最为重要的环境中表现出色。

其次，我们通过调整算法 1 中结果残差之间的相关强度$\{\hat{e}_{i,m}: i = 1, \ldots, n,\ m = 1, \ldots, M\}$进行一项实验。令 $\hat{\Sigma}_e = \frac{1}{n} \sum_{i=1}^{n} (\hat{\mathbf{e}}_i - \bar{\hat{\mathbf{e}}})(\hat{\mathbf{e}}_i - \bar{\hat{\mathbf{e}}})^\top, \hat{\mathbf{e}}_i = (\hat{e}_{i,1}, \ldots, \hat{e}_{i,M})^\top$，并将 $\hat{\Sigma}_e = L_e L_e^\top$写作 $D_e = \text{diag}(\hat{\Sigma}_e)$ 和 $R_{e,0} = D_e^{-1/2} \hat{\Sigma}_e D_e^{-1/2}$，其中后者表示经验相关矩阵。

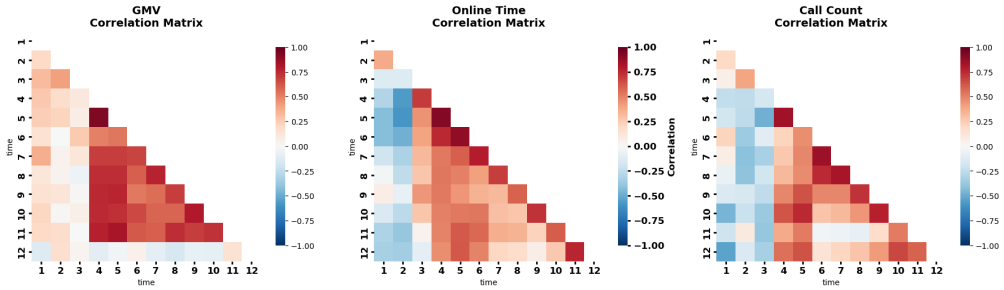
为了在保持边际方差固定的同时改变相依性，我们构造一个单参数族$\{R_e(\rho): \ \rho \in [-1, 1]\}$ 在独立性、经验结构以及相关性更强的目标$R_{e,1}$ 之间插值（对于 $m \neq m'$，我们取$(R_{e,1})_{mm'} = 0.6$ ）。具体而言，

$$R_e(\rho) = \begin{cases} (1+\rho)\, R_{e,0} - \rho\, I_M, & \rho \in [-1, 0], \\ (1-\rho)\, R_{e,0} + \rho\, R_{e,1}, & \rho \in [0, 1], \end{cases}$$

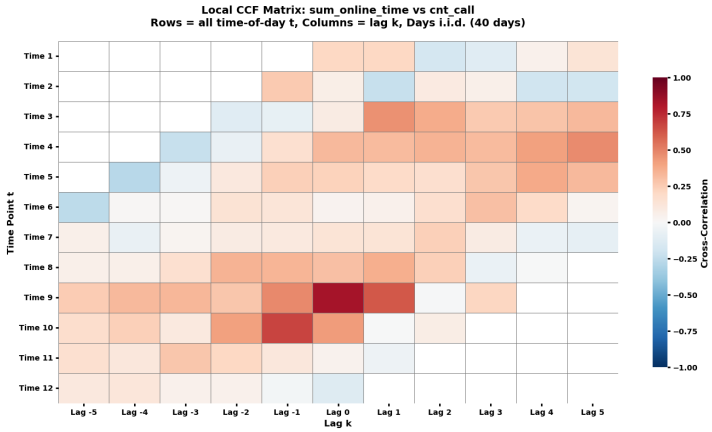
从而使得 $R_e(-1) = I_M$， $R_e(0) = R_e$，0，且 $R_e(1) = R_e$，1。随后我们设定 $C_e(\rho) = D_e^{1/2} R_e(\rho) D_e^{1/2}$， $C_e(\rho) = L_e(\rho) L_e(\rho)^\top$， $A_e(\rho) = L_e(\rho) L_e^{-1}$，并定义变换后的残差 $\tilde{\mathbf{e}}_i(\rho) = A_e(\rho) \hat{\mathbf{e}}_i$， $i = 1$，....., n. 由构造可知，对所有 ρ，$\text{Var}(\tilde{\mathbf{e}}_i(\rho)) = C_e(\rho)$且$\text{diag} C_e(\rho) = \text{diag}(\Sigma_e)$。因此，边际方差得以保持，而时间相关性在 $\rho \to -1$ 时减弱，在 $\rho \to 1$时增强。于是，在算法 1中用 $\tilde{\mathbf{e}}_i(\rho)$替换$\hat{\mathbf{e}}_i$，便得到以 ρ为索引的一族自助法模拟器，我们用其考察估计量的均方误差如何随相关性强度变化。在我们的实验中，我们考虑 $\rho \in \{-0.8, -0.4, 0, 0.4, 0.8\}$。实证结果（图11和表8）显示出显著的时间持续性、较强的跨时点相关性以及复杂的超前—滞后结构——在这些条件下，相较于TMDP和NMDP所依据的马尔可夫假设，TRL的全历史建模具有明显优势。



(a) *GMV* 与状态变量的自相关（滞后 0 至 5）。



(b) *GMV*、在线时长总和和呼叫次数的跨时间相关矩阵。



(c) 需求（在线时长总和）与供给（呼叫次数）在不同时间点（t=1 至 12）之间的互相关。

图9：真实数据中时间依赖性与跨序列依赖性的实证证据。

表7：按方法和需求–供给相关性划分的均方误差（$\times 10^{-4}$）（ϕ_{coef}）

ϕ_{coef}	NMDP	TMDP	TRL
0.25	58.4	56.5	**31.6**
0.50	60.3	56.9	**32.3**
1.00	66.2	59.9	**39.2**
1.50	75.8	66.4	**40.3**

表8：按方法和残差时间相关性划分的均方误差（$\times 10^{-4}$）（ρ）

ρ	NMDP	TMDP	TRL
-0.8	4.88	4.10	**2.91**
-0.4	5.78	5.19	**3.47**
0.0	6.62	5.99	**3.92**
0.4	7.99	7.25	**4.79**
0.8	9.47	8.26	**5.58**

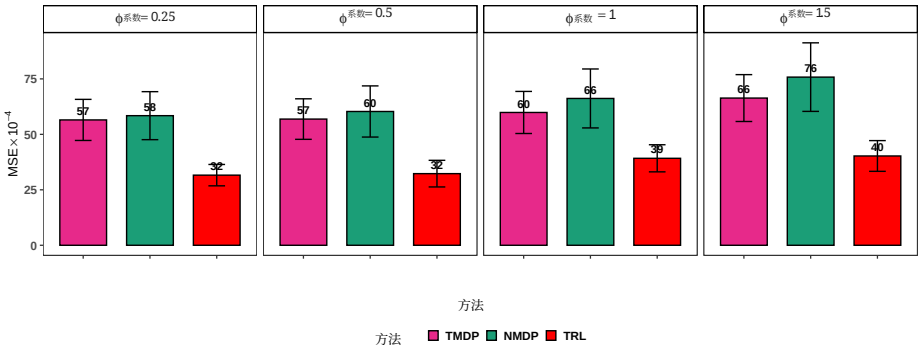


图 10：不同需求–供给相关结构（由 Φ 参数化）下的经验均方误差（$\times 10^{-4}$）。

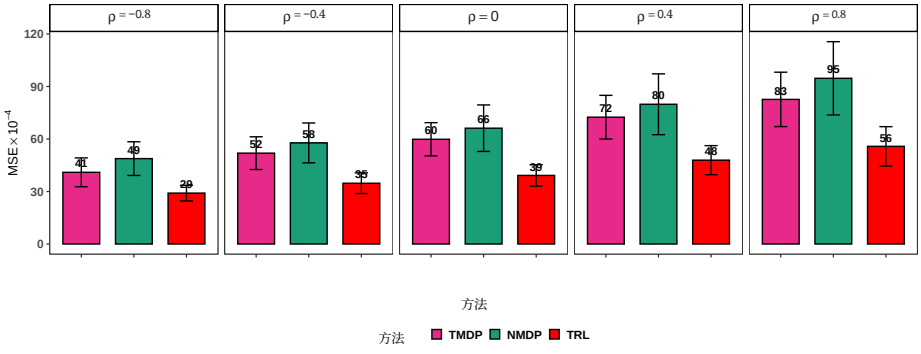


图11：不同残差相关水平（由ρ参数化）下的经验均方误差（$\times 10^{-4}$）。