软件学报 ISSN 1000-9825, CODEN RUXUEW
2026,37(5):2024−2042 [doi: 10.13328/j.cnki.jos.007563] [CSTR: 32375.14.jos.007563]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

TB-Match: 融合条件扩散模型与近端策略优化的弹性时间约束运单分配方法*

廖家俊, 董宜滔, 毛嘉莉

(华东师范大学 数据科学与工程学院, 上海 200062)

通信作者: 毛嘉莉, E-mail: jlmao@dase.ecnu.edu.cn

摘　要: 在动态环境下的双边匹配问题中, 对于时间约束与多目标优化的处理机制是影响匹配效率的重要因素之一, 网络货运平台的运单分配即为此类问题的典型实例. 现有方法在处理时间约束的刚性建模和多目标冲突的权衡机制方面存在显著局限性, 难以准确刻画决策主体在约束边界附近的行为特征. 提出一种基于条件扩散模型与分层强化学习的时间约束感知匹配框架 TB-Match, 通过弹性约束量化、偏好表征学习、动态权衡优化和策略生成这 4 个协同模块实现系统性能提升. 该方法的核心贡献包括: (1) 基于条件扩散概率模型的约束弹性化表征机制, 通过渐进噪声扩散与逆向去噪过程将确定性时间边界转化为连续概率分布, 精确建模决策主体在约束临界区域的接受概率; (2) 融合动态目标权衡与近端策略优化的分层决策架构, 高层网络根据反馈信号自适应调节目标权重, 低层网络通过信任域约束实现长期累积收益最大化. 在两个大规模真实数据集上的实验结果表明, TB-Match 在匹配率指标上比现有最优方法相对提升了 17.66%, 同时在满意度等指标中均展现出显著的性能优势, 证明了该方法在复杂约束环境下的有效性和适用性.

关键词: 运单分配; 司机偏好建模; 弹性时间约束; 扩散模型; 近端策略优化

中图法分类号: TP301

中文引用格式: 廖家俊, 董宜滔, 毛嘉莉. TB-Match: 融合条件扩散模型与近端策略优化的弹性时间约束运单分配方法. 软件学报, 2026, 37(5): 2024–2042. http://www.jos.org.cn/1000-9825/7563.htm

英文引用格式: Liao JJ, Dong YT, Mao JL. TB-Match: Elastic Time-constrained Transport Order Assignment Method Integrating Conditional Diffusion Model and Proximal Policy Optimization. Ruan Jian Xue Bao/Journal of Software, 2026, 37(5): 2024–2042 (in Chinese). http://www.jos.org.cn/1000-9825/7563.htm

TB-Match: Elastic Time-constrained Transport Order Assignment Method Integrating Conditional Diffusion Model and Proximal Policy Optimization

LIAO Jia-Jun, DONG Yi-Tao, MAO Jia-Li

(School of Data Science & Engineering, East China Normal University, Shanghai 200062, China)

Abstract: In the bilateral matching problem under dynamic environments, the mechanism for handling time constraints and multi-objective optimization is one of the important factors affecting matching efficiency. The transport order assignment in online freight platforms serves as a typical instance of such problems. Existing methods exhibit significant limitations in rigid modeling of time constraints and in the trade-off mechanisms for multi-objective conflicts, making it difficult to accurately characterize the behavioral patterns of decision agents near constraint boundaries. To address these issues, this study proposes a time-constraint-aware transport order assignment framework called TB-Match. The framework consists of four collaborative modules: elastic constraint quantification, preference representation learning, dynamic objective trade-off optimization, and policy generation. The core contributions are as follows: (1) a constraint elasticity representation mechanism based on conditional diffusion probabilistic models, which converts deterministic time boundaries into continuous

* 基金项目: 国家自然科学基金 (62461146205); 海南省重点研发项目 (ZDYF2025GXJS179)
收稿时间: 2025-07-18; 修改时间: 2025-08-20; 采用时间: 2025-10-10; jos 在线出版时间: 2026-01-28
CNKI 网络首发时间: 2026-01-29

probabilistic distributions through progressive noise diffusion and reverse denoising processes, thus accurately modeling the acceptance probability of decision agents in boundary regions; (2) a hierarchical decision framework integrating dynamic objective trade-off and proximal policy optimization, where the high-level network adaptively adjusts objective weights according to feedback signals, and the low-level network maximizes long-term cumulative rewards under trust region constraints. Experimental results on two large-scale real-world logistics datasets demonstrate that TB-Match achieves a 17.66% relative improvement in matching rate compared with state-of-the-art methods. It also exhibits significant advantages in metrics such as satisfaction, verifying the effectiveness and applicability of the proposed method under complex constraint environments.

Key words: transport order assignment; driver preference modeling; elastic time constraint; diffusion model; proximal policy optimization

网络货运平台作为连接货主与车主的数字化物流中枢, 在现代物流体系中扮演着至关重要的角色, 其不仅是物流行业的关键基础设施, 更是推动产业升级的重要力量. 近年来, 伴随着电子商务的蓬勃发展和智能物流技术的快速进步, 网络货运平台在整合分散的运力资源、提升整体物流运作效率等方面展现出独特价值, 发挥着不可替代的作用. 国家层面出台的《“十四五”数字经济发展规划》等政策文件明确强调了网络货运平台的重要性, 要求通过此类平台构建智慧物流智能系统, 实现城市物流资源的全面互联互通与高效共享, 持续优化城市物流运行体系, 完善智慧物流网络建设, 这些政策导向充分彰显了网络货运平台在推动数字物流创新发展进程中的核心地位. 为积极响应国家政策号召, 切实增强平台核心竞争力, 网络货运平台亟需实现运营效率的持续优化与资源利用率的最大化, 这一目标已成为推动平台高质量发展的关键所在. 在此过程中, 提升运单分配的效能不仅被视为实现上述目标的核心突破口, 也是提升整体运作水平和市场竞争力的关键着力点.

运单分配作为网络货运平台最核心的业务流程, 本质上是一个通过智能算法将货主提交的运输需求 (即运单) 与最合适的货车司机进行精准匹配的自动化决策过程, 其优化程度在很大程度上决定了平台的运营效率、用户满意度以及市场竞争力. 大量实践表明, 高效的运单分配机制不仅能够显著提升平台对运力资源的利用效率、有效降低车辆空驶率, 还能同步增强平台在服务层面对于货主与司机双方的吸引力和用户黏性. 然而, 通过对现有研究的系统梳理可以发现, 尽管学界已在运单分配算法方面开展了大量深入探讨, 但相关研究仍存在较为明显的局限性: 在优化目标设定方面, 多数研究集中于成本相关的性能指标, 通常以最大化单车装载重量或最小化空驶距离等单一维度作为模型构建的核心导向[1,2]; 在运输可行性评估方面, 现有方法普遍采用基于平均运输时间的启发式判断策略, 即通过估算订单的平均运输时间是否满足最晚送达要求来筛选具备执行能力的运力资源参与匹配. 虽然此类方法在计算效率上具有一定优势, 但却忽略了实际运营中影响货车司机接单决策的一系列关键时间约束因素, 例如装卸货过程中的等待时间成本、按时交付的保障概率以及返程任务的时间限制等. 这些未被充分考虑的因素构成了司机评估运单可行性的核心依据, 直接影响其对运单完成时间的预期判断, 进而导致平台所生成的运单分配结果与其实际接受意愿之间可能出现较大偏差, 影响了司机对分配结果的接受概率.

近年来学术界针对当前研究中存在的局限性, 开始探索将更多影响货车司机接单决策的关键因素纳入运单分配模型的构建过程. 在运输领域, 已有学者通过分析司机的历史行为数据, 建立其对运输提货点与目的地、偏好运输路线等静态属性的偏好模型, 并基于此类偏好进行运单匹配的优化决策[1]. 然而, 这类方法在理论层面仍存在明显不足, 主要表现为对静态偏好指标的过度依赖, 未能充分考虑运输任务中具有重要影响的动态因素. 此外, 现有研究普遍采用固定时间预期或平均运输时间来评估任务的可行性, 这种简化处理忽略了实际运输场景中的复杂性和个体差异. 具体而言, 不同经验水平的司机在面对相同运输任务时可能表现出显著的行为差异, 例如经验丰富的司机往往能够通过路线优化、灵活调整作业时间等方式有效缩短实际运输时间, 而这些行为特征并未被现有模型准确捕捉, 从而导致对司机接单意愿的刻画与实际情况存在偏差. 值得注意的是, 在众包任务分配等相关研究领域, 已有成果表明工作者的接单偏好具有显著的时间动态特性, 相关研究通过引入时间敏感性等动态因素, 提高了接单率预测的准确性[3]. 尽管如此, 将此类研究方法直接迁移至网络货运平台的运单分配场景仍面临诸多挑战. 运输任务的复杂性对适应性建模提出了更高要求, 需要解决动态约束条件下的决策优化问题. 具体而言, 现有研究在运单分配算法设计方面存在两个关键性不足, 这些不足严重制约了整体匹配成功率和司机满意度的提升.

挑战 1: 运输时间约束降低了司机的接单概率. 严格的运输时间约束作为导致司机拒单的主要因素, 包含装卸

货地点的工作时间限制、最晚送达时间以及司机返程时间要求等多维约束条件. 这些约束限制了司机的可选运单范围, 而现有算法通常基于平均运输时间进行匹配决策, 未能充分考虑司机在实际运输过程中对行驶时间和停留时间的动态调整能力. 以图 1 所示情况为例, 当某运单的平均运输时间为 4 h 时, 传统算法可能判定司机无法满足时间约束而放弃分配, 但经验丰富的司机往往能够通过优化路线、提高装卸效率等方式缩短实际运输时间. 尽管时间约束在形式上表现为硬性边界, 但其实际执行具有弹性特征, 因此简单的二元判断 (能否完成) 应转变为概率化评估 (完成的可能性). 这种转变更符合现实场景中司机的行为特征, 有助于扩大可匹配运单池并提高接单率. 基于此, 本文需要解决的第 1 个挑战是如何从历史数据中构建时间约束的松弛模型, 以准确刻画司机在不同约束条件下的接单概率分布.

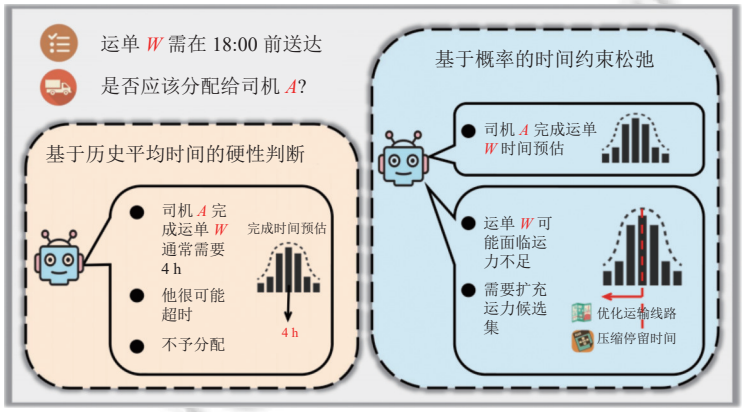

图 1 时间约束松弛示意图

挑战 2: 运单分配算法在匹配率与司机满意度之间存在目标冲突. 在运单分配过程中, 当多位司机对同一时空范围内的运单表现出相似偏好时, 现有算法通常优先考虑匹配率最大化, 例如通过对前序时段内未被成功分配的运单进行重新分配或基于需求预测模型为后续时段即将到达的车辆预先保留合适的运单. 然而, 这种做法往往以牺牲司机满意度为代价, 进而可能触发司机拒单行为的发生 (如图 2 方案 1 所示). 相反地, 若仅以司机满意度为优化目标, 则可能降低整体匹配率 (如图 2 方案 2 所示). 虽然已有研究尝试通过引入加权机制来寻求二者间的平衡, 但预设的静态权重系数难以适应物流场景中供需关系与司机偏好的动态演化特征, 同时, 未能充分考虑运单分配问题特有的序列决策属性——即当前决策时段的分配方案会通过改变系统状态 (如车辆位置、车辆载货状态等) 而影响后续决策时段的匹配效果. 这种时序关联性使得传统的帕累托最优分析方法 (主要适用于静态决策环境) 在评估全局优化效果时表现出明显的局限性. 如图 2 方案 3 所示, 当在 T_1 决策时段进行运单分配时, 若算法能够考虑当前决策对 T_2 时段匹配效果的潜在影响, 并策略性地为司机 X 分配次优而非即时最优的运单, 尽管这种决策在局部时间窗口内可能偏离瞬时最优解, 但从全局视角来看, 却能有效提升累计成功匹配数量, 同时降低总体拒单率. 因此, 本文需要解决的第 2 个挑战是如何在动态供需环境下通过考虑分配决策的时序关联性实现匹配率与司机满意度的有效权衡.

针对运输时间约束对司机接单概率的负面影响问题, 本文设计了一种基于扩散模型的时间约束表征模型. 该方法突破了传统确定性时间约束的建模局限, 通过引入渐进式噪声注入与去除的逆向学习过程, 实现了对复杂时间约束分布的准确建模. 我们将其拓展到运输时间约束的表征过程, 通过将各个时间约束边界进行概率化处理, 将原本离散且确定性的时间约束区间转化为一个连续的概率分布空间, 从而为每个相关时间点赋予具体的接受概率值. 这一策略允许模型更准确地捕捉司机在时间约束临界点附近的接单倾向, 有效扩大了潜在的有效匹配范围. 针对挑战 2 所涉及的多目标优化问题, 本文提出了一种融合动态目标权衡与近端策略优化的分配策略生成模型, 旨在解决匹配率与司机满意度之间的冲突. 其中, 动态目标权衡模块利用强化学习反馈机制设计了以拒单率作为负向惩罚项的奖励函数, 用以评估当前分配策略的表现, 并根据实时拒单情况动态调整匹配率与司机满意度在优化

目标中的相对权重. 具体地, 在高拒单率情况下增加司机满意度的权重以提升接单意愿; 而在低拒单率情形下则侧重于提高匹配率权重, 以优化整体分配效率. 此外, 为解决序列决策中的长期优化问题, 引入改进的近端策略优化(PPO) 算法, 构建了跨时间窗口的价值评估网络. 该网络通过设计具有时间折扣因子的累积奖励函数, 实现了对当前决策在未来多个时间窗口内影响的系统性评估. 算法采用裁剪策略梯度更新的技术手段, 通过约束策略更新的最大幅度, 有效避免了优化过程中的策略振荡现象, 保证了训练过程的稳定性, 最终实现包括匹配率与司机满意度在内的长期累积收益最大化.

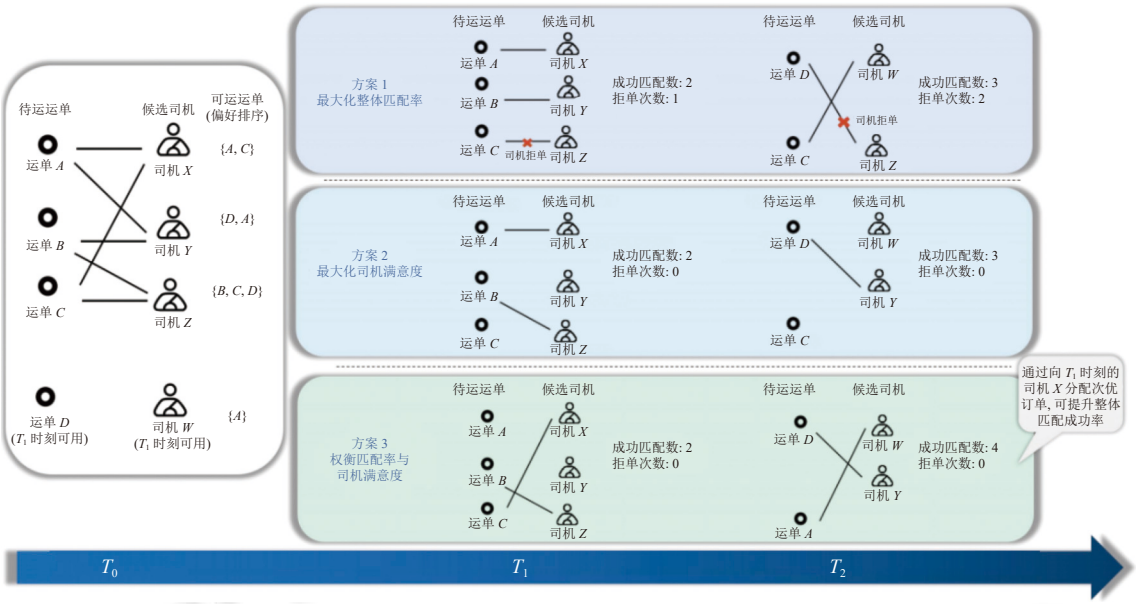

图 2　匹配率与司机满意度冲突示意图

总的来说, 本研究提出了一种时间约束感知与平衡导向匹配框架 TB-Match (time-constrained balance matching), 该框架通过 4 个相互协同的技术创新实现了运单分配问题的系统性优化. 首先, 基于条件扩散模型的弹性时间约束建模模块将刚性时间约束转化为动态弹性表征, 通过概率扩散技术构建了时间约束满足度的连续概率分布模型; 其次, 基于序列注意力机制的司机偏好表征模块通过整合历史运输记录、实时状态信息和弹性时间约束表征, 利用多头注意力机制有效捕捉特征间的非线性关联, 实现对司机接单行为的精准预测; 第三, 基于强化学习反馈的目标权衡优化模块建立了以拒单行为为反馈信号的动态调节机制, 通过在线学习算法自适应调整匹配率与满意度的权重分配; 最后, 基于近端策略优化的匹配策略生成模块构建跨时间窗口价值评估网络, 利用信任区域机制调整运单分配策略, 实现匹配率与满意度的长期累积收益最大化. 本研究的创新性贡献主要体现在以下 4 个方面.

1) 为提高网络货运平台的运单匹配率与司机满意度, 提出了一个时间约束偏好感知的运单分配框架, 称为 TB-Match, 由时间约束弹性量化、司机偏好表征、目标权衡与匹配策略生成等模块组成.

2) 针对将时间约束视为硬性边界的分配决策局限性, 设计了一种基于条件扩散模型的时间约束表征模型. 通过概率扩散机制将刚性时间约束评估 (即仅判断能否完成) 转化为考虑个体差异的连续概率评估 (即对于完成概率的量化估计), 扩展了可匹配范围.

3) 为解决匹配率与司机满意度之间的目标冲突问题, 提出了一种融合动态目标权衡机制与近端策略优化的分配策略生成模型. 该模型利用拒单行为作为反馈信号, 通过强化学习动态调节匹配率与满意度的权重, 并通过近端策略优化算法优化分配决策的长期影响, 实现全局最优的运单分配策略.

4) 基于两个真实物流数据集的实验结果表明, TB-Match 框架在运单分配性能上显著优于现有最优方法. 在日均货运量达 30 000 吨、运力规模为 1 200 辆货车的实际运营场景下, TB-Match 实现了 17.66% 的成功匹配率提

升, 显著提高了物流系统的运营效率.

1 相关工作

运单分配问题需要在满足时间约束的前提下平衡平台匹配率与司机满意度. 现有研究在运单分配算法优化与司机偏好建模方面已取得重要进展, 然而这些研究成果并不能直接应用在运单分配场景. 具体而言, 其面临的困难包含两个方面: 忽视了针对不同司机个体对运输时间约束下运输可行性的判断, 以及在匹配率与司机满意度的多目标优化时忽略了运单分配场景的时序决策属性.

1.1 运单分配与司机偏好建模

物流平台中的订单分配作为经典的组合优化问题, 其研究方法已从静态匹配发展为动态优化. 静态匹配算法 (如贪心匹配、匈牙利算法、拍卖算法和局部搜索启发式) 假设所有订单和司机信息在分配时完全已知且保持不变, 虽能保证理论最优性却难以适应实际物流场景的动态性需求[4−8]. 为应对订单和司机状态的实时变化, 在线匹配算法 (如竞争比分析框架和动态规划方法)[9,10]以及强化学习技术[11]被引入该领域, 显著提升了系统对动态环境的适应能力. 然而, 这些方法仍存在关键局限性: 一方面, 它们未能有效建模和预测司机的拒单行为, 导致分配策略的实际效果受限; 另一方面, 现有动态优化方法多聚焦于即时收益最大化, 缺乏对长期累积收益的系统性优化[12,13].

司机偏好建模作为提升订单接受率的关键技术, 其研究范式经历了从静态规则到动态学习的演进. 早期基于规则的方法通过分析历史接单记录构建静态司机画像, 虽能识别基本偏好模式却无法捕捉偏好的时空动态性[1]. 近年来, 图神经网络通过建模空间关联关系捕捉地理场景差异[14,15], 序列模型 (如 LSTM 和 Transformer) 则专注于学习偏好的时序演化规律, 结合注意力机制进一步提升了模型对时空动态特征的捕捉能力[3,16]. 然而, 当前偏好建模研究存在两个显著不足: 在建模维度上, 现有工作过度关注订单基本属性 (如货物类型、距离等) 的偏好分析, 而忽视了时间约束压力下司机决策机制的研究; 在优化目标上, 多数方法片面追求接单率最大化, 未能将偏好建模与平台整体效率提升有机结合, 导致部分运单长期滞留的系统性问题.

1.2 运输时间约束建模

当前时间约束建模研究面临的核心挑战在于准确评估司机在给定时间限制内完成运输任务的可能性. 主流方法采用基于阈值的二元判断机制, 将装货时间窗口和最晚到达时间等约束视为刚性边界, 通过比较预计运输时间与约束条件的数值关系进行可行性判定 (如直接排除超时匹配)[2,9,10,17]. 虽然这种方法计算高效, 但其固有缺陷在于: 一方面, 它忽略了不同司机在相同时间约束下可能表现出的执行能力差异; 另一方面, 它无法反映运输过程中固有的时间波动性.

为应对时间不确定性, 近期研究主要沿着两个技术路径发展: 一是基于蒙特卡洛仿真和鲁棒优化的外部环境因素建模, 重点处理交通状况等随机扰动对运输时间的影响[17]; 二是采用贝叶斯方法的动态预测框架, 通过持续更新时间估计来适应环境变化[3,15]. 这些概率模型虽然提升了时间预测的鲁棒性, 但其建模视角仍存在本质局限——将时间波动完全归因于外部环境因素, 而未能充分考虑司机个体在面对时间压力时的差异化决策行为. 这种对司机主观能动性的建模缺失, 导致现有方法难以准确预测司机的实际接单意愿, 从而制约了运单分配系统的整体效能.

物流运单分配本质上是一个多目标优化问题, 旨在同时优化匹配效率与司机满意度等相互冲突的关键指标[9,10,17]. 现有研究主要沿着两条技术路径展开: 一方面, 基于权重聚合的单目标转化方法通过线性加权的方式将多个目标合并为一个综合目标函数, 从而利用成熟的单目标优化算法进行求解, 此类方法实现简单、计算高效, 但其预设的固定权重难以适应物流市场中随时间动态变化的需求特征, 尤其在对系统目标优先级的不同要求下, 表现出明显的适应性局限[1,10,14,15,17]; 另一方面, 帕累托优化方法则通过构建非支配解集, 提供了一种无需预先设定权重的多目标决策框架, 能够在不牺牲某一目标的前提下探索不同目标之间的折中关系[18,19]. NSGA-II 和 MOEA/D 等典型帕累托优化算法已被应用于运单分配场景中, 能够生成反映匹配率与司机满意度之间权衡关系的解集. 然而, 这些方法在实际应用中仍面临显著挑战: 但其计算复杂度随规模指数增长, 难以满足大规模实时分配的秒级响应需求, 虽然部分研究[20−23]利用 MOEA/D 分解与 SMS-EMOA 超体积指标简化搜索过程降低计算开销, 但是仍存

在一些问题使其无法直接应用于运单分配场景. 一是这些方法主要针对离线优化场景, 对于运单分配这种实时决策环境仍面临状态空间爆炸问题, 无法满足在线分配的要求; 二是这些方法采用静态优化范式, 忽略了当前分配决策对司机状态 (如位置分布和满意度水平) 演变过程的影响, 从而可能引发长期分配性能的下降[9,10,16]. 针对静态优化问题, 部分研究引入了强化学习技术来建模时序决策过程, 兼顾决策的长期收益, 但仍存在奖励函数难以设计的问题. 具体而言, 单一智能体的复合奖励函数会因量纲差异导致梯度失衡, 而多智能体方案又因策略冲突引发收敛难题[11]. 近年来, 多目标强化学习 (MORL) 作为处理多目标优化问题的新兴范式, 在网约车任务分配等领域已被应用. 现有 MORL 方法主要分为两类: 基于帕累托前沿的方法 (如 multi-policy actor-critic (MPAC)) 通过维护多个策略网络同时学习帕累托最优解集, 能够在不预设权重的情况下探索目标间的权衡关系[24]; 基于标量化函数的方法则通过动态调整目标权重将多目标问题转化为单目标优化, 在保持计算效率的同时适应环境变化[25]. 然而, 这些方法在运单分配场景中并不能直接应用. 具体而言, 帕累托方法虽能生成多样化解集, 但其计算复杂度随目标数量指数增长, 难以在运单分配的复杂场景中应用; 标量化方法虽然计算高效, 但其权重调整策略多基于启发式规则, 其调整策略在运单分配场景中难以确定.

针对上述方法在动态适应性 (加权和法)、计算效率与长期最优 (帕累托优化) 以及可学习性 (强化学习) 的问题, 在本文中我们使用分层强化学习来解决这些问题, 通过将多目标权衡决策与具体分配执行分离到不同层次, 高层智能体专注于根据系统状态动态调整目标权重, 低层智能体则在给定权重下执行具体的分配决策, 这种分层结构既避免了复合奖励函数设计的困难, 又避免了多智能体方法的目标冲突, 同时通过分层分解方法有效降低了计算开销.

2 问题定义

本节对运单分配问题进行形式化定义, 其核心在于建立融合时间约束弹性量化、司机偏好建模和多目标动态权衡的智能分配决策框架.

定义 1. 运单. 给定物流运输平台, 运单 o 被形式化定义为六元组: $o=\langle loc_{\text{pick}}, loc_{\text{del}}, t_{\text{load}}^{\text{start}}, t_{\text{load}}^{\text{end}}, t_{\text{del}}^{\text{late}}, cargo\_type\rangle$, 其中, loc_{pick}, $loc_{\text{del}} \in \mathbb{R}^2$ 分别表示装货地与卸货地的地理坐标, $\left[t_{\text{load}}^{\text{start}}, t_{\text{load}}^{\text{end}}\right]$ 构成装货时间窗口, 表示司机可以在这个时间段内接货, $t_{\text{del}}^{\text{late}}$ 为合同约定的最晚送达时间, $cargo\_type$ 表示待运输货物类别. $\mathcal{O}=o_{k=1}^{|\mathcal{O}|}$, 其中 $|\mathcal{O}|$ 表示在某个时间段内需要分配的运单集合.

定义 2. 司机. 平台上参与运单分配的司机被定义为司机, 被形式化定义为三元组: $d=\langle loc_{\text{curr}}, P_d, \mathcal{C}_d\rangle$, 其中, $loc_{\text{curr}} \in \mathbb{R}^2$ 表示司机所属货车当前 GPS 坐标, $P_d \subseteq \mathbb{R}^2 \times \mathbb{R}^2$ 为司机偏好的运输线路集合 (装货地-卸货地对), $\mathcal{C}_d$ 表示该司机可承运的货物类型集合, 平台可用资源构成有限集合 $\mathcal{D}=d_j^{|\mathcal{D}|}$, 其中 $|\mathcal{D}|$ 为当前在线可调度车辆数.

定义 3. 时间约束. 时间约束是指运输过程需要遵守的时间规定, 其形式被定义为四元组: $\mathcal{T}_{o,d}=\langle t_{\text{load}}^{\text{start}}, t_{\text{load}}^{\text{end}}, t_{\text{del}}^{\text{late}}, t_{\text{return}}^{d}\rangle$, 其中, 装货时间窗口 $\left[t_{\text{load}}^{\text{start}}, t_{\text{load}}^{\text{end}}\right]$ 为闭区间, 表示司机需要在此期间至装货地进行接货; 最晚送达时间 $t_{\text{del}}^{\text{late}}$ 构成硬约束, 表示货物必须在此时间点前送达目的地; 返程时间 t_{return}^{d} 为可选约束, 即司机需要在特定时间前返回原地点.

基于上述定义, 挑战 1 可形式化为时间约束弹性的建模问题. 我们将司机的接单决策分解为两个步骤: 先评估任务可行性 (货物兼容性与时间约束满足度), 再判断司机的偏好.

定义 4. 运单接受概率. 我们定义运单接受概率 $P_{\text{accept}}(o,d)$ 为司机 d 接受运单 o 的条件概率. 该概率通过分析司机 d 的历史承运记录 $H_d=\{(o_1,a_1),(o_2,a_2),\ldots,(o_n,a_n)\}$, 其中 o_i 是历史上分配给司机 d 的运单, $a_i \in \{0,1\}$ 表示司机是否接受该运单. 形式化地, 我们将司机 d 接受运单 o 的总概率 $P_{\text{accept}}(o,d)$ 定义为两个独立概率的乘积:

$$P_{\text{accept}}(o,d)=P_{\text{feasible}}(o,d)\times P_{\text{pref}}(o,d) \tag{1}$$

其中, $P_{\text{feasible}}(o,d)$ 表示可行性概率, 它首先表明司机 d 的货车是否能够运输运单 o 中的货物, 同时在给定的时间约束 $(\mathcal{T}_o, t_{\text{return}}^{d})$ 下, 能够按时完成运单 o 的可能性, 该概率由第 3.1 节的模型计算得到. $P_{\text{pref}}(o,d)$ 表示司机的偏好概率, 它表示在任务时间可行的前提下, 司机 d 的主观偏好 (如目的地、运输时间等) 上愿意接受运单 o 的可能性. 该概率主要通过第 3.2 节的模型分析司机的历史选择记录 H_d 计算得出.

定义 5. 司机满意度. 司机在周期 T 内的满意度 $S(d)$ 定义为其接受运单的平均接受概率:

$$S(d)=\frac{1}{|A_d^T|}\sum_{o\in A_d^T}P_{\text{accept}}(o,d) \tag{2}$$

其中, $A_d^T\subseteq\mathcal{O}$ 表示 T 周期内司机实际接受的运单集合. 该度量反映司机获得偏好匹配运单的期望程度.

问题陈述 (problem statement): 给定运单集合 $\mathcal{O}$ 和货车集合 $\mathcal{D}$, 以及它们的属性和司机的历史行为数据, 目标是找到一个随时间动态调整的匹配策略 π, 该策略生成一系列匹配决策 M_t (在时间窗口 t 内的分配方案), 以最大化长期累积的成功匹配数量, 同时考虑并提升司机对所接受运单的满意度. 形式化地, 目标是优化

$$\max_{\pi}\mathbb{E}\left[\sum_t\gamma^t\left(\lambda\sum_{(o,d)\in M_t}\mathbb{I}(accepted(o,d))+(1-\lambda)\sum_{(o,d)\in M_t}P_{\text{accept}}(o,d)\right)\right] \tag{3}$$

其中, $\mathbb{I}(accepted(o,d))$ 是指示函数, 表示司机 d 是否最终接受了分配给他的运单 o. $\gamma\in(0,1]$ 为时间折扣因子. 参数 $\lambda\in[0,1]$ 用于平衡两个优化目标.

- 最大化成功匹配: 即 $\lambda\sum_{(o,d)\in M_t}\mathbb{I}(accepted(o,d))$, 最大化实际接受的运单.
- 最大化司机满意度: 即 $(1-\lambda)\sum_{(o,d)\in M_t}P_{\text{accept}}(o,d)$, 该目标旨在提升分配方案的采纳率, 通过模型预测, 优先推荐司机更可能接受的选项.

整个运单分配问题中既需要最大化成功匹配数量, 又要提升司机整体满意度. 为实现此目标, 我们需要分别构建时间可行性概率 P_{feasible} 和司机偏好概率 P_{pref} 的模型, 并在此基础上设计出最优的匹配策略 π.

3 TB-Match 技术框架

为解决当前运单分配方法忽视时间约束弹性与司机偏好的问题, 本文提出了一种基于时间约束松弛优化理论的物流运单分配优化方法 TB-Match, 如图 3 所示. 该框架包含 4 个核心模块: (1) 基于扩散模型的时间约束表征模块, 该模块首先通过 Transformer 提取司机历史交互中的上下文特征, 然后利用条件去噪扩散概率模型 (CDDPM) 将刚性的时间约束边界转化为概率分布, 实现约束弹性的精确建模; (2) 基于序列注意力机制的司机偏好融合模块, 该模块将生成的弹性时间偏好与运单特征和司机状态整合, 预测司机的接单概率; (3) 基于强化学习反馈的目标权衡优化模块, 该模块通过高层策略网络, 根据司机拒单行为动态调整匹配效率与司机满意度两个优化目标的相对权重; (4) 基于近端策略优化的匹配策略生成模块, 该模块在高层策略网络输出的权重指导下生成具体的司机-运单匹配决策, 同时通过信任区域机制确保策略的稳定更新. 上述模块整合为时间约束特征捕捉模块 (第 3.1 节) 与分配策略生成模块 (第 3.2 节).

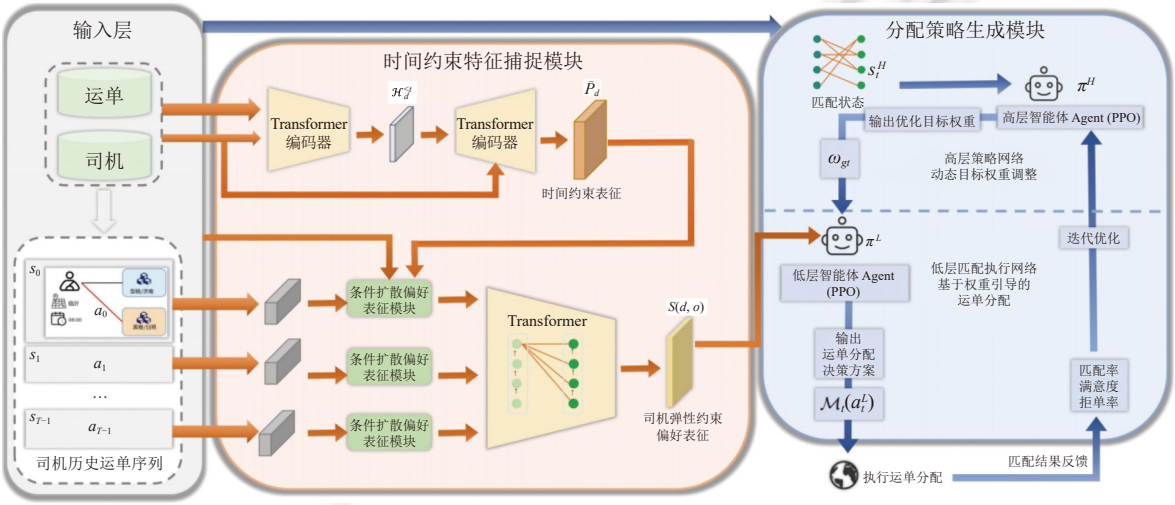

图 3 TB-Match 技术框架示意图

3.1 时间约束特征捕捉模块

本节旨在构建一个能够精确建模司机时间约束弹性偏好的表示模块. 传统方法在评估司机对时间约束的接受意愿时, 往往忽略了司机在不同时间约束条件下表现出的差异化弹性特征. 为解决这一局限性, 本节提出一种融合 Transformer 结构与条件扩散模型的新型框架, 如图 3 所示. 该框架首先利用 Transformer 从司机历史交互序列中提取深层次的上下文特征, 并结合当前订单信息, 生成司机偏好的条件均值表征; 随后, 以该条件均值为引导, 采用条件扩散模型生成能够反映司机在时间约束边界区间内决策倾向的弹性偏好概率分布; 最后, 将该概率化偏好与当前订单特征融合, 输出可用于后续匹配决策的动态偏好评分.

为实现对司机动态时间偏好的高精度建模, 本节提出一种基于 Transformer 的历史上下文感知与条件均值偏好估计方法. 具体而言, 我们将每位司机的历史交互序列 $\mathcal{H}_d=(e_1,e_2,\ldots,e_K)$ 作为输入, 其中每一条交互 $e_k=(c_k,\mathcal{O}_k^{\text{avail}},a_k)$ 包含了 k 时刻的上下文状态、可选运单集以及司机实际选择的运单. 为充分利用历史行为序列的时序依赖性和上下文丰富性来捕捉时间约束的相关特征, 我们借助 Transformer 结构多头自注意力的序列建模能力来提取多层次的行为特征, 更准确地表征时间约束对于司机运单选择的影响.

该模块采用条件 Transformer 编码器-解码器架构. 编码器 $\mathcal{T}_{\text{cond}}$ 输入包括历史行为序列的嵌入表示 $Embed(\mathcal{H}_d)$ 以及当前订单特征的嵌入 $Embed(o_{\text{curr}})$. 通过多层自注意力和前馈网络的深度融合, 模型能够有效整合历史与当前信息, 生成一个上下文向量 $\mathbf{c}(o_{\text{curr}}|\mathcal{H}_d)$. 随后, 解码器部分 $\mathcal{T}_{\text{dec}}$ 以该上下文向量和当前订单特征 $Embed(o_{\text{curr}})$ 为输入, 通过自注意力和与编码器输出的交叉注意力机制, 输出条件均值偏好表征 $\hat{\mathbf{p}}(o_{\text{curr}}|\mathcal{H}_d)$. 该表征反映了司机在当前情境下对时间约束的主观弹性预期. 具体而言, 条件均值偏好表征 $\hat{\mathbf{p}}(o_{\text{curr}}|\mathcal{H}_d)$ 中捕捉了司机在运输场景决策时的多种特征: (1) 时间敏感性特征, 表征司机在高峰期与平峰期不同的历史接单模式; (2) 距离偏好特征, 表征反映其对短途、中途、长途运单的不同偏好强度; (3) 货物类型适应性, 表征司机对不同货物类型的历史承运经验, 捕捉其在特定货物处理方面的专业能力; (4) 延误容忍度, 通过表征司机历史运单中的实际完成时间与预期时间的偏差模式, 量化其在面对潜在延误时的接受概率; (5) 返程约束敏感性, 基于司机的历史返程行为模式, 表征其对需要在特定时间返回原地的运单的偏好程度. 另外, 该部分采用 Transformer 架构的另一个好处是能够为后续的条件扩散模型简化学习难度, 使其更专注于建模数据分布的高阶不确定性和个体化弹性特征. 在本框架中, $\hat{\mathbf{p}}$ 作为弹性偏好生成的中心参考点, 使得扩散模型能够围绕其进行概率扩展, 从而更准确地反映司机在时间约束边界附近的真实决策倾向.

3.1.1 条件扩散驱动的弹性偏好概率分布生成

在获得条件均值偏好表征 $\hat{\mathbf{p}}(o_{\text{curr}}|\mathcal{H}_d)$ 后, 我们设计了一个基于条件扩散模型的时间约束表征模型来生成能够捕捉司机时间偏好弹性的概率化表征 $\mathbf{p}_0$, 如图 4 所示. 具体而言, 为了捕捉运单分配场景下司机时间偏好弹性的概率化表征, 需要解决以下技术难点: 首先, 司机在时间约束临界点附近的决策行为呈现高度非线性的概率分布特征, 需要模型能够精确刻画从“完全可接受”到“完全不可接受”之间的平滑过渡区域, 而自回归模型的序列生成机制无法有效建模这种连续概率空间中的复杂分布结构; 其次, 不同司机在相同时间压力下的偏好概率分布存在显著个体差异, 要求模型能够在保持分布多样性的同时避免模式坍塌, 而对抗生成网络在处理此类多模态分布时容易出现训练不稳定和模式坍塌问题, 导致生成的偏好表征缺乏必要的个体化差异.

为了解决以上问题, 从运单分配场景出发, 我们发现司机对于时间约束的弹性本质上是一个从确定性约束向概率性评估的渐进转化过程, 这与扩散模型通过逐步去噪实现从噪声到数据的生成机制在概念上高度契合. 基于此, 我们设计了一个条件扩散偏好表征模块, 其通过模拟可逆的噪声添加与去除过程, 能够学习到以 $\hat{\mathbf{p}}$ 为中心、覆盖司机在时间约束边界附近所有可能决策状态的复杂概率分布, 该渐进式生成过程模拟了司机在面对时间约束时的决策思维过程: 从基础时间预期出发, 根据实际运输条件进行弹性调整, 最终形成对运单的接受概率判断.

为训练该模型, 我们首先需要定义其学习目标, 即源自历史数据的真实偏好状态 $\mathbf{p}_0^{\text{true}}$. 该状态是对司机已完成运单的实际决策结果 (如接受或拒绝) 及其相关的时间弹性特征 (如相较于预估时间的提前或延迟量) 进行向量化编码的结果. 基于此, 条件扩散模型的目标便是学习一个从标准正态分布噪声出发, 在条件均值 $\hat{\mathbf{p}}$ 的引导下, 通过逐步去噪生成与真实数据 $\mathbf{p}_0^{\text{true}}$ 分布一致的偏好样本.

前向加噪过程 (forward process): 与标准的扩散模型类似, 我们定义一个固定的前向过程, 在 S 个离散时间步 (用 s 索引, 以区别于物理时间 t) 中, 逐步向真实的偏好状态 $\mathbf{p}_0^{\text{true}}$ 中添加高斯噪声, 同时将过程锚定到条件均值 $\hat{\mathbf{p}}$. 这一过程可以理解为从具体的历史运输表现逐渐向基于统计规律的平均预期收敛的过程, 其中噪声代表了司机个体差异和环境不确定性等因素的综合影响. 在任意扩散步骤 s, 从 $\mathbf{p}_{s-1}$ 到 $\mathbf{p}_s$ 的转移定义为:

$$q(\mathbf{p}_s|\mathbf{p}_{s-1},\hat{\mathbf{p}})=\mathcal{N}\left(\mathbf{p}_s;\sqrt{1-\beta_s}\mathbf{p}_{s-1}+\left(1-\sqrt{1-\beta_s}\right)\hat{\mathbf{p}},\beta_s\mathbf{I}\right) \tag{4}$$

其中, $\{\beta_s\}_{s=1}^S$ 是预设的噪声调度表. 这使得 $\mathbf{p}_s$ 可以直接从 $\mathbf{p}_0^{\text{true}}$ 和 $\hat{\mathbf{p}}$ 采样得到:

$$\mathbf{p}_s=\sqrt{\bar{\alpha}_s}\mathbf{p}_0^{\text{true}}+\left(1-\sqrt{\bar{\alpha}_s}\right)\hat{\mathbf{p}}+\sqrt{1-\bar{\alpha}_s}\epsilon \tag{5}$$

其中, $\bar{\alpha}_s=\prod_{i=1}^{s}(1-\beta_i)$ 且 $\epsilon\sim\mathcal{N}(0,\mathbf{I})$. 公式 (5) 描述了加噪过程是在真实偏好 $\mathbf{p}_0^{\text{true}}$ 和条件均值 $\hat{\mathbf{p}}$ 之间进行插值, 并叠加受控噪声. 在运输场景中, 这一过程反映了实际运输表现的生成机制: 任何具体的运输结果都可以看作是基础预期 (由历史数据和订单特征决定)、司机个体能力差异以及环境随机因素三者综合作用的产物. 当 $s\to S$ (S 表示扩散总步数) 时, $\bar{\alpha}_s\to 0$, $\mathbf{p}_S$ 的分布将主要由 $\hat{\mathbf{p}}$ 和标准差决定, 这对应了在缺乏具体历史信息时只能依靠统计规律进行预测的现实情况. 我们定义扩散过程的终点为 $p(\mathbf{p}_S|\hat{\mathbf{p}})=\mathcal{N}(\mathbf{p}_S;\hat{\mathbf{p}},\mathbf{I})$.

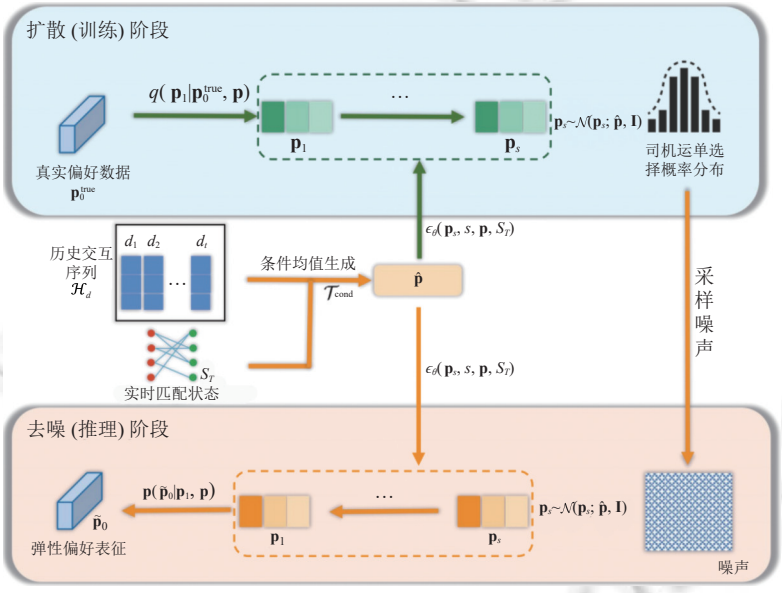

图 4 条件扩散偏好表征模块示意图

反向去噪过程 (reverse process): 反向过程的目标是训练一个网络 $\epsilon_\theta(\mathbf{p}_s,s,\hat{\mathbf{p}},o_{\text{curr}})$, 使其能够根据当前的含噪偏好表征 $\mathbf{p}_s$、扩散步数 s、条件均值偏好 $\hat{\mathbf{p}}$ 以及当前订单特征 o_{curr}, 来预测在前向过程中每一步所添加的噪声 ϵ. 在运输场景中, 这个网络学习的是如何从基础的平均预期和订单特征出发, 逐步恢复那些代表司机个体能力差异和环境随机性的细节信息, 从而生成既符合统计规律又包含合理个体化弹性的时间偏好表征. 训练目标是最小化预测噪声与真实噪声之间的均方误差:

$$\mathcal{L}_{\text{CDDPM}}(\theta)=\mathbb{E}_{s,\mathbf{p}_0^{\text{true}},\epsilon\sim\mathcal{N}(0,\mathbf{I})}\left[w(s)\|\epsilon-\epsilon_\theta(\mathbf{p}_s(\mathbf{p}_0^{\text{true}},\epsilon,s,\hat{\mathbf{p}}),s,\hat{\mathbf{p}},o_{\text{curr}})\|^2\right] \tag{6}$$

其中, $\mathbf{p}_s(\cdot)$ 是根据公式 (5) 计算得到的含噪样本, $w(s)$ 是与扩散步数相关的权重.

推理时, 我们从先验 $\mathbf{p}_S\sim\mathcal{N}(\hat{\mathbf{p}},\mathbf{I})$ 开始, 利用训练好的 ϵ_θ 迭代地执行去噪步骤, 最终生成一个弹性偏好表征样本 $\tilde{\mathbf{p}}_0$. 从概率分布的角度分析, 在时间约束宽松的情况下, 分布趋向于以司机历史偏好为中心的单峰分布; 而在时间约束紧张时, 分布呈现双峰或多峰特征, 反映了司机在“接受挑战性运单”与“选择安全运单”之间的决策分歧. 条件扩散模型生成的这种概率结构, 相比于传统的单点估计方法提供了更丰富的决策信息, 在运输场景中对应了从

基于统计规律的平均预期逐步细化为包含个体特征和环境因素的具体时间偏好表征的推理过程, 使得模型能够为每个具体的运输任务生成既合理又具有适当不确定性的弹性偏好预测.

3.1.2 偏好融合与接单概率预测

最后, 为了得到一个可用于决策的最终分数, 我们将扩散模型生成的弹性时间偏好 $\tilde{\mathbf{p}}_0$、订单的详细特征 e_o 以及司机的实时状态 c_d^{curr} 进行融合. 我们设计了一个基于注意力机制的融合网络 $\mathcal{F}_{\text{attn}}$ 来实现这一目标, 它将这三者作为序列输入, 通过多头自注意力机制捕捉它们之间的复杂交互, 最终输出一个单一的接单概率 $S(d,o)$:

$$S(d,o)=\sigma(MLP(\lambda_{\text{attn}}(Embed(\tilde{\mathbf{p}}_0),Embed(e_o),Embed(c_d^{\text{curr}})))) \tag{7}$$

其中, $\sigma(\cdot)$ 是 Sigmoid 函数. 该模块的训练由标准的二元交叉熵损失 $\mathcal{L}_S$ 监督, 其标签 $y\in\{0,1\}$ 来自司机的历史接单记录. 所得到的 $S(d,o)$ 不仅量化了司机的接单意愿, 也作为其对该分配满意度的度量, 作为后续决策模块的输入.

3.2 分配策略生成模块

在获取了司机的弹性时间约束偏好表征后, 接下来需要解决的核心问题是如何基于该表征实现运单的优化分配. 在本节中我们将使用一个基于分层强化学习的决策框架来解决该问题. 具体而言, 该框架包含一个高层策略网络和一个低层匹配执行网络, 两者均采用近端策略优化 (PPO) 算法进行训练. 高层网络负责在宏观层面动态调整优化目标的权重 (即匹配效率与司机满意度之间的相对重要性), 而低层网络则在高层策略的指导下执行具体的运单分配决策. 通过这两个网络的协同作用, 框架旨在学习一个能够最大化长期累积综合回报的分配策略, 该回报综合考量了匹配效率与司机满意度两个优化目标.

3.2.1 高层策略网络: 多目标权重动态调整

高层策略网络旨在应对匹配率与司机满意度之间的权衡问题, 通过学习机制动态调整系统在不同时间窗口内对这两个核心优化目标的侧重程度.

● 状态 (state s_t^H): 高层网络在每个决策周期 t 开始时观测一个高层状态 s_t^H, 该状态全面表征了当前的整体匹配环境. 其构成如下: $s_t^H=\left(\mathcal{O}_{\text{total},t},\mathcal{D}_{\text{avail},t},\bar{S}_{\text{avg},t-1},R_{\text{reject},t-1},\mathbf{z}_{\text{VI},t}\right)$. 其中, $\mathcal{O}_{\text{total},t}$ 表示当前时间窗口内待分配的运单总量; $\mathcal{D}_{\text{avail},t}$ 表示当前可用的司机数量; $\bar{S}_{\text{avg},t-1}$: 表示上一周期内已匹配运单的平均司机满意度; $R_{\text{reject},t-1}$ 表示上一周期内的司机平均拒单率. $\mathbf{z}_{\text{VI},t}$ 表示由变分推断网络生成的对当前司机-订单潜在匹配质量的低维表征. 为有效从这些可能蕴含复杂时空依赖的状态特征中提取信息, 高层策略网络采用一个基于时空注意力机制的状态编码器, 对输入状态 s_t^H 进行深度特征提取, 生成更为稠密的状态表征 $\mathbf{s}_t^H$.

其中, 变分推断网络 $\mathbf{z}_{\text{VI},t}$ 由编码器网络 $q_\phi(\mathbf{z}|\mathcal{O}_{\text{total},t},\mathcal{D}_{\text{avail},t})$ 与解码器网络 $p_\theta(\mathcal{O}_{\text{total},t},\mathcal{D}_{\text{avail},t}|\mathbf{z})$ 所组成, 其损失函数结合了重构损失与 KL 散度正则化项: $\mathcal{L}_{\text{VI}}=\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})]-\beta\cdot KL(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))$, 其中 $\beta=0.1$ 为 KL 权重, $p(\mathbf{z})=\mathcal{N}(0,\mathbf{I})$ 为标准正态先验. 该网络通过学习运单-司机匹配空间的潜在结构, 为高层策略网络提供当前匹配环境的稠密表征 $\mathbf{s}_t^H$.

● 动作 (action a_t^H): 基于状态表征 $\mathbf{s}_t^H$, 智能体输出一个动作 $a_t^H=\omega_t$, 其中 $\omega_t\in[0,1]$ 是一个动态调整的平衡因子. 该因子 ω_t 用于决定在当前决策周期内, 低层匹配执行网络在优化其奖励时, 应给予“最大化匹配数量”和“最大化司机满意度”这两个目标的相对权重. 例如, ω_t 趋近于 1 时, 侧重于提升匹配数量; ω_t 趋近于 0 时, 则更侧重于司机满意度.

● 奖励 (reward R_t^H): 高层策略网络的奖励 R_t^H 在低层网络根据其指导参数 ω_t 完成一个批次的运单分配任务之后进行更新, 直接反映了所选平衡因子 ω_t 在该周期内的实际效果. 其形式化定义如下:

$$R_t^H=\lambda_{\text{rate}}\cdot MR_t+\lambda_{\text{sat}}\cdot DS_t-\lambda_{\text{rej}}\cdot RR_t \tag{8}$$

其中, MR_t 是当前批次实际达成的匹配率, DS_t 是该批次所有成功匹配运单的平均司机满意度 (基于 $S(d,o)$ 计算), RR_t 是该批次的司机拒单率. λ_{rate}、λ_{sat}、λ_{rej} 是预设的超参数, 用以调整各组成部分对整体奖励的贡献. 此奖励设计使司机拒单行为作为一种负向反馈, 直接影响高层策略的优化方向.

● 策略与价值函数: 高层网络的策略 $\pi^H(\omega_t|\mathbf{s}_t^H)$ 和价值函数 $V^H(\mathbf{s}_t^H)$ 均由神经网络参数化, 并通过 PPO 算法进行优化.

3.2.2 低层执行网络: 权重引导的运单分配策略

低层匹配执行网络负责在接收到高层策略网络输出的平衡因子 ω_t 后, 生成具体的司机与运单的分配策略.

● 状态 (state s_t^L): 低层网络在每个匹配决策时刻 t (通常对应一个更细的时间粒度或一个分配批次) 观测的状态 s_t^L 包含当前待决策的微观环境信息: $s_t^L=\left(\mathcal{D}_{\text{curr}},\mathcal{O}_{\text{curr}},\{S(d,o)\}_{\forall d\in\mathcal{D}_{\text{curr}},o\in\mathcal{O}_{\text{curr}}},\omega_t\right)$. 其中, $\mathcal{D}_{\text{curr}}$ 表示当前时间窗口内所有可用司机的集合, 包含其各自的实时位置、能力、历史行为等特征; $\mathcal{O}_{\text{curr}}$ 表示当前待分配的运单集合, 包含其各自的起讫点、货物类型、时间窗口等特征; $\{S(d,o)\}$ 为前文司机偏好模块输出的每个可用司机 d 对每个待分配运单 o 的偏好评分 (即接单概率 $P_{\text{accept}}(o,d)$); ω_t 表示由高层策略网络传递的当前平衡因子. 这些信息都通过一个 Transformer 编码器转化为低层状态表征 $\mathbf{s}_t^L$.

● 动作 (action a_t^L): 基于状态表征 $\mathbf{s}_t^L$, 智能体输出一个动作 $a_t^L=\mathcal{M}_t$, 其中 $\mathcal{M}_t=\{(o,d),\ldots\}$ 是一组具体的司机-运单匹配对. 该匹配决策需满足基本约束, 如一个运单仅能分配给一个司机, 一个司机在同一时段内通常只承接一个运单. PPO 算法通过计算策略 $\pi^L(\mathcal{M}_t|\mathbf{s}_t^L)$ 来生成这些匹配决策.

● 奖励 (reward R_t^L): 低层网络的奖励 R_t^L 用于评估当前批次匹配决策 $a_t^L=\mathcal{M}_t$ 的质量, 并直接受到高层传递的平衡因子 ω_t 的影响. 对于每一个最终被司机接受的成功匹配 $(o,d)\in\mathcal{M}_t$, 其对低层奖励的贡献 $r_t^L(o,d|\omega_t)$ 定义为:

$$r_t^L(o,d|\omega_t)=\omega_t\cdot\mathbb{I}(accepted(o,d))+(1-\omega_t)\cdot S(d,o)\cdot\mathbb{I}(accepted(o,d)) \tag{9}$$

其中, $\mathbb{I}(accepted(o,d))$ 是指示函数, 当司机 d 最终接受运单 o 时为 1, 否则为 0. $S(d,o)$ 是司机 d 对运单 o 的偏好评分. 则低层网络在当前批次的总奖励为所有成功匹配贡献之和: $R_t^L=\sum_{(o,d)\in\mathcal{M}_t}r_t^L(o,d|\omega_t)$ 低层网络的目标是学习一个策略 π^L 来最大化该累积奖励.

● 策略与价值函数: 低层网络的策略 $\pi^L(\mathcal{M}_t|\mathbf{s}_t^L)$ 和价值函数 $V^L(\mathbf{s}_t^L)$ 同样由神经网络参数化, 并通过 PPO 算法进行优化.

3.2.3 分层网络协同训练机制

为确保高层策略网络与低层匹配执行网络能够有效协作, 我们为两个网络的训练设计了一种协同训练的方法. 在每个训练 epoch 中, 首先固定低层网络参数, 使用高层网络收集的经验数据更新高层策略; 然后固定高层网络参数, 基于高层输出的权重指导更新低层策略. 这种交替训练策略避免了两个网络同时更新可能导致的训练不稳定问题. 此外, 引入经验回放机制, 高层和低层网络分别维护独立的经验池 (容量分别为 10 000 和 50 000), 通过随机采样历史经验进行离线策略更新, 提高样本利用效率并增强训练稳定性.

具体而言, 训练过程采用迭代方式进行. 在每个训练迭代 (episode) 中, 有如下步骤.

● Step 1. 高层策略网络根据当前高层状态 s_t^H (经其状态编码器处理) 选择一个平衡因子 $a_t^H=\omega_t$.

● Step 2. 该平衡因子 ω_t 被传递给低层匹配执行网络, 并作为其状态 s_t^L 的一部分.

● Step 3. 低层网络依据其当前策略 $\pi^L(a_t^L|s_t^L)$ 生成具体的匹配决策 $a_t^L=\mathcal{M}_t$.

● Step 4. 仿真环境或真实系统执行匹配决策后, 反馈该批次匹配的即时结果, 包括每个匹配对 (o,d) 是否被司机接受以及司机对该运单的偏好评分 $S(d,o)$.

● Step 5. 根据公式 (9) 计算低层网络的奖励 $R_t^L=\sum r_t^L(o,d|\omega_t)$, 并用于更新低层网络的策略 π^L 和价值函数 V^L 的参数.

● Step 6. 在低层网络完成一批次任务 (或一个完整的高层决策周期) 后, 系统计算该周期的整体性能指标, 如实际匹配率、平均司机满意度及拒单率.

● Step 7. 根据公式 (8) 计算高层策略网络的奖励 R_t^H, 并用于更新高层网络的策略 π^H 和价值函数 V^H 的参数.

整个过程产生的经验数据 $(s_t^H,a_t^H,R_t^H,s_{t+1}^H)$ 和 $(s_t^L,a_t^L,R_t^L,s_{t+1}^L)$ 将分别存入各自的经验回放池中, 用于 PPO 算法的后续离线策略更新.

3.3 端到端优化与梯度流控制

TB-Match 框架通过一个组合式的多组件损失函数进行端到端优化, 该函数旨在协同训练偏好表示模块 (第 4.1 节) 与分层匹配决策智能体 (第 4.2 节). 本节将详细阐述该目标函数, 并解析模块间的信息流与梯度控制机制.

3.3.1 总体损失函数

框架的总损失 $\mathcal{L}_{\text{total}}$ 由偏好表示损失、分层决策损失以及辅助损失这 3 部分组成:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{pref}}(\mathcal{L}_{\text{CDDPM}} + \mathcal{L}_S) + \lambda_{\text{rl}}\left(\mathcal{L}_{\text{RL}}^H + \mathcal{L}_{\text{RL}}^L\right) + \mathcal{L}_{\text{aux}} \tag{10}$$

其中, 各项定义如下.

● 偏好表示损失 ($\mathcal{L}_{\text{CDDPM}} + \mathcal{L}_S$): 这部分负责训练偏好表示模块. $\mathcal{L}_{\text{CDDPM}}$ (公式 (6)) 训练扩散模型以学习时间弹性的条件分布; $\mathcal{L}_S$ 是用于训练最终接单概率预测网络的二元交叉熵损失. λ_{pref} 是其权重.

● 分层决策损失 ($\mathcal{L}_{\text{RL}}^H + \mathcal{L}_{\text{RL}}^L$): 这部分负责训练分层强化学习智能体. $\mathcal{L}_{\text{RL}}^H$ 和 $\mathcal{L}_{\text{RL}}^L$ 分别是高层和低层策略网络基于 PPO 算法的损失函数, 每个损失都包含策略裁剪项、价值函数项和熵奖励项. λ_{rl} 是其权重.

● 辅助损失 ($\mathcal{L}_{\text{aux}}$): 包括所有网络参数的 L2 正则化项, 以防止过拟合.

3.3.2 梯度传播路径与训练稳定性机制

在训练中仍需要解决的一个问题是, 强化学习信号 (奖励) 如何反向传播以优化上游的偏好表示网络. 为此, 我们对模型整体的梯度流进行了梳理, 并采用梯度隔离策略进一步提升模型训练时的稳定性.

① 信息流路径: 偏好表示模块输出的接单概率 $S(d,o)$ 是连接表示学习与决策学习的关键桥梁. 具体而言, 低层策略的奖励 R_t^L 直接依赖于 $S(d,o)$ (见公式 (9)), 而 $S(d,o)$ 又是偏好表示网络 (参数为 θ_{pref}) 的可微输出. 因此, 存在一条从 RL 损失到偏好表示网络的梯度路径:

$$\frac{\partial \mathcal{L}_{\text{RL}}}{\partial \theta_{\text{pref}}} = \frac{\partial \mathcal{L}_{\text{RL}}}{\partial \hat{A}_t^L} \frac{\partial \hat{A}_t^L}{\partial R_t^L} \frac{\partial R_t^L}{\partial S(d,o)} \frac{\partial S(d,o)}{\partial \theta_{\text{pref}}} \tag{11}$$

其中, $\hat{A}_t^L$ 是低层策略的优势函数. 此梯度流使得智能体在决策时, 不仅会优化其分配策略, 还会通过奖励信号, 隐式地指导偏好表示网络生成更有利于获得高奖励的偏好分数.

② 训练稳定性控制: 完全端到端的训练可能因 RL 信号的高方差而导致不稳定. 为解决此问题, 我们基于多目标强化学习中的训练稳定性分析设计了一种梯度隔离策略. 具体而言, 当 $S(d,o)$ 同时作为状态输入和奖励计算的组成部分时, 会形成循环依赖的梯度路径, 导致训练过程中的非平稳性问题[26].

为提升训练中的稳定性, 当 $S(d,o)$ 作为 RL 智能体状态 (state) 的一部分输入时, 将其视为常数 (即 “detach()”操作), 切断其梯度回传, 从而避免扩散噪声在策略梯度中被放大而引起的非平稳性. 仅当 $S(d,o)$ 用于计算奖励 (reward) 时, 该梯度将被保留, 确保奖励信号中的梯度仍能有效回传, 使偏好表示网络能够根据其输出对最终决策的贡献进行优化. 这种选择性梯度传播策略在理论上等价于在偏好学习和策略学习之间建立了一个“软解耦”机制, 既保持了两个模块间的信息交互, 又避免了训练不稳定和目标冲突问题.

3.4 计算复杂度分析

为了评估 TB-Match 在实际部署中的可行性, 我们分析了其在推理阶段 (即生成一次批处理分配决策) 的计算复杂度.

(1) 偏好表示生成: 对于一个包含 $|\mathcal{D}|$ 个司机和 $|\mathcal{O}|$ 个订单的批次, 最坏情况是为所有可能的 $|\mathcal{D}|\times|\mathcal{O}|$ 对计算偏好.

● Transformer: 对于每个司机, 需要处理其长度为 K 的历史序列, 复杂度为 $O(|\mathcal{D}|\cdot K\cdot d^2)$, 其中 d 是模型隐藏维度. 此步骤可预先计算并缓存.

● 条件扩散模型: 对每个候选对 (d,o), CDDPM 采样需要 S 步, 每步网络计算复杂度为 C_{denoiser}. 总复杂度为 $O(|\mathcal{D}||\mathcal{O}|\cdot S\cdot C_{\text{denoiser}})$.

(2) 匹配决策生成

● 高层策略网络: 其状态维度固定, 计算开销可忽略, 为 $O(1)$.

● 低层策略网络: 需要评估所有候选对的分数, 复杂度为 $O(|\mathcal{D}||\mathcal{O}|\cdot C_{\text{policy}})$, 其中 C_{policy} 是低层策略网络的单次

前向传播开销.

总体而言, TB-Match 的推理复杂度主要由待分配货车和运单的规模决定, 约为 $O(|\mathcal{D}||\mathcal{O}|)$, 与贪心算法 (greedy matching, GM) 在同一数量级, 但具有更大的常数因子. 然而, 它远优于需要构建和求解二分图的匈牙利算法 (Hungarian algorithm, HA), 后者的复杂度为 $O\left(\max(|\mathcal{D}|,|\mathcal{O}|)^3\right)$. 在实际应用中, 可以通过引入候选集生成等剪枝策略 (例如, 仅考虑地理上邻近的司机-订单对), 将复杂度从 $O(|\mathcal{D}||\mathcal{O}|)$ 降低到 $O\left(|\mathcal{D}|\cdot\bar{k}\right)$, 其中 $\bar{k}$ 是每个司机的平均候选订单数, 以更高效地处理大规模的实时运单分配任务.

4 实 验

本实验部分将解决以下问题.

RQ1: 我们提出的模型与其他现有模型相比表现如何?

RQ2: 不同模型组件的效果如何?

RQ3: 超参数如何影响模型性能?

RQ4: 我们提出的模型在鲁棒性指标方面的表现如何?

4.1 实验设置

4.1.1 实验数据集描述

● RS 数据集. 该数据集由一家钢铁公司提供, 包含货运账单和司机信息. 整个数据集包含 173 000 条运输任务记录, 涉及 1 200 多名司机的信息, 时间跨度为两个月 (2020 年 7–8 月). 在实际实验中, 平均每日运单数量为 2 883 条, 每 15 min 划分为一个批次, 平均每批次分配涉及 150 个运单与 80 名可用司机. 每条记录包括执行运输任务的司机 ID、运输目的地 (经度、纬度及其所属行政区域)、运输开始和结束的日期与时间、完成运输所需的时间以及运输距离. 该数据集中的运输任务都具有相同的出发点, 我们将使用该数据集测试模型在只有一个出发点时的性能.

● DT-CARGO 数据集[27]. DT-CARGO 是慕尼黑工业大学于 2021 年秋季至 2022 年春季从 54 辆货车收集的开源数据集. 由于该数据集不包含具体的任务信息, 我们将货车从驶离工业区到下一个工业区的行程视为一次运输任务, 并为每次运输任务随机分配运输重量. 每次行程的第 1 个和最后一个标记分别被标记为运输的出发点和目的地.

4.1.2 基线模型设置

为全面评估所提出 TB-Match 框架的有效性与先进性, 本文选取了 6 种具有代表性的基线方法进行对比实验, 涵盖了传统启发式、经典组合优化、元启发式算法、图神经网络及强化学习等不同技术范式. 具体如下.

● 贪婪匹配 (GM): 作为运单分配领域的经典启发式算法, GM 采用局部最优策略, 在每个决策时刻将待分配运单优先分配给当前接单概率最高的可用司机.

● 匈牙利算法 (HA): 作为二分图最大权匹配的经典组合优化算法, HA 通过构建司机-运单二分图并求解最大权重匹配, 能够在静态假设下保证全局最优解.

● 遗传算法 (genetic algorithm, GA)[28]: 作为元启发式优化方法的代表, GA 通过模拟生物进化过程求解带时间约束的任务分配问题. 该方法采用染色体编码表示分配方案, 通过选择、交叉、变异等遗传操作在解空间中进行全局搜索.

● 时间加权偏好感知任务分配 (temporal-weighted preference-aware task assignment, TPTA)[3]: TPTA 基于协同过滤理论构建工作者偏好模型, 通过分析历史交互数据挖掘工作者的任务偏好模式, 并结合时间衰减因子捕捉偏好的动态演化特征.

● 融合用户偏好学习任务分配 (fused user preference learning for task assignment, FUPTA)[29]: FUPTA 采用图神经网络架构建模工作者-任务交互关系, 通过图卷积操作聚合邻域信息学习工作者的深层偏好表征.

● 偏好强化学习 (multi-agent preference Transformer, MAPT)[11]: MAPT 是一个多智能体偏好学习框架, 利用级

联 Transformer 架构来捕捉时序与协同依赖关系, 从而训练出能更契合用户偏好的智能体奖励函数.

4.1.3 评估指标

为反映各方法在实际物流运单分配场景下的综合表现, 本文使用如下评估指标.

● 匹配率 (*MR*): 在评估周期 T 内成功匹配的运单占总运单数的比例, 即:

$$MR = \frac{|\{(o,d) \in M_T : \mathbb{I}(accepted(o,d)) = 1\}|}{|\mathcal{O}|},$$

其中, M_T 表示周期 T 内的分配方案集合, $\mathbb{I}(accepted(o,d))$ 为指示函数, 当司机 d 接受运单 o 时为 1, 否则为 0. 该指标直接反映平台运力资源的利用效率.

● 司机满意度 (*DS*): 量化已分配运单与司机偏好的契合度, 其定义为所有成功匹配运单的平均接受概率:

$$DS = \frac{1}{|\{(o,d) \in M_T : \mathbb{I}(accepted(o,d)) = 1\}|} \sum_{(o,d)\in M_T} P_{\text{accept}}(o,d) \cdot \mathbb{I}(accepted(o,d)),$$

其中, $P_{\text{accept}}(o,d)$ 为司机 d 对运单 o 的接受概率 (由公式 (7) 计算得出). 该指标反映算法在追求整体效率的同时对司机个人偏好的兼顾程度.

4.1.4 实现细节

我们将每个数据集按照时间顺序划分为训练集 (占前 75% 的数据) 与测试集 (占后 25% 的数据). 在模型架构方面, 偏好表示模块中的条件扩散模型包含 4 个残差块, 隐藏维度为 256. 分层匹配决策框架中的高层与低层策略网络均采用包含 6 层、8 个注意力头、隐藏维度为 512 的 Transformer 编码器结构. 对于训练超参数, 我们为所有网络均采用 Adam 优化器, 其中偏好表示模块的学习率设为 1×10^{-4}, 匹配决策模块的学习率设为 5×10^{-5}, 批大小 (batch size) 统一为 256. 强化学习部分采用近端策略优化 (PPO) 算法进行训练, 其折扣因子 γ 为 0.95, 广义优势估计 (GAE) 的 λ 参数为 0.95, PPO 的裁剪范围 ϵ_{clip} 为 0.2, 并设置熵正则化系数为 0.01 以鼓励探索. 在高层策略网络的奖励函数中, 各项目权重设置为: $\lambda_{\text{rate}} = 1.0$, $\lambda_{\text{sat}} = 0.8$, $\lambda_{\text{rej}} = 1.5$. 所有对比基线模型的参数均按照其原始论文或官方实现进行设置, 以确保公平比较. 具体而言, 遗传算法 (GA) 设置种群规模为 100, 交叉概率为 0.8, 变异概率为 0.1, 迭代次数为 500 代; TPTA 模型中的时间衰减因子 α 设为 0.1, 邻域大小 k 设为 20, 学习率为 1×10^{-3}; FUPTA 采用 3 层图卷积网络, 隐藏维度为 128, dropout 率为 0.2, 学习率为 5×10^{-4}; MAPT 方法使用 PPO 算法训练多智能体系统, 学习率为 5×10^{-4}, 折扣因子为 0.99, 经验回放缓冲区大小为 10 000. 所有深度学习基线均采用相同的数据预处理流程. 对于缺乏内置偏好感知机制的传统分配算法, 则采用本文提出的司机偏好建模方法来获取偏好信息.

4.2 总体性能对比 (RQ1)

表 1 展示了我们提出的 TB-Match 模型与 6 种基线方法在 RS 数据集和 DT-CARGO 数据集上的总体性能比较. 实验从匹配率 (*MR*) 和司机满意度 (*DS*) 两个维度上评估了各方法的性能表现. 结果如表 1 所示. 最佳结果以粗体突出显示.

表 1 在两个数据集上的总体性能比较

方法	RS数据集		DT-CARGO数据集	
	MR	*DS*	*MR*	*DS*
GM	0.582	0.651	0.663	0.612
HA	0.615	0.704	0.692	0.685
GA[28]	0.643	0.732	0.721	0.713
TPTA[3]	0.661	0.768	0.739	0.742
FUPTA[29]	0.675	0.785	0.752	0.764
MAPT[11]	0.691	0.812	0.760	0.793
TB-Match	**0.813**	**0.857**	**0.845**	**0.836**

从匹配率指标来看, TB-Match 在 RS 数据集上实现了 0.813 的匹配率, 与最佳基线模型 MAPT 的 0.691 相比, 提升了 17.66%, 这表明我们的模型能够更有效地利用可用运力资源, 减少运单积压. 在 DT-CARGO 数据集上, TB-

Match 同样取得了 0.845 的匹配率, 相比于 MAPT 提升了 11.18%. 司机满意度指标进一步证实了 TB-Match 的优势. 在 RS 数据集上, 模型达到了 0.857 的满意度得分, 比 MAPT 相对提升了 5.5%. 在 DT-CARGO 数据集上, 满意度得分为 0.836, 相比于 MAPT 提升了 5.4%, 说明模型在保证整体效率的同时能够更好地满足司机需求. 相比于传统的静态匹配方法和基于强化学习的动态方法, TB-Match 实现了匹配率和司机满意度的协同优化, 验证了我们提出方法的有效性和实用价值.

4.3 消融实验 (RQ2)

为了细致评估我们提出的 TB-Match 框架中核心组件的各自贡献, 我们在 RS 数据集上进行了一项全面的消融研究. 我们将完整 TB-Match 模型的性能与几个变体进行了比较, 每个变体都系统地省略或简化了一个关键的架构元素. 该实验的结果详见表 2.

表 2 TB-Match 及其变体在 RS 数据集上消融实验的性能比较

框架模块	模型变体	*MR*	*DS*
—	完整模型 (TB-Match)	**0.813**	**0.857**
偏好表示变体 (第3.1节)	去除条件扩散模型	0.762	0.831
	去除序列注意力机制	0.775	0.840
	去除时间约束捕捉模块	0.748	0.754
匹配决策框架变体 (第3.2节)	去除动态目标权衡模块	0.794	0.846
	去除匹配策略生成网络	0.769	0.826

条件扩散模型作为时间约束弹性建模的核心组件, 其重要性在消融实验中得到了充分验证. 去除条件扩散模型导致了显著的性能下降, *MR* 从 0.813 降至 0.762, 相对降幅为 6.4%, *DS* 从 0.857 降至 0.831, 相对降幅为 3.1%. 这一结果充分证实了将时间约束转化为概率化弹性表征的必要性. 传统的二元判断方法 (能否完成) 忽略了司机在时间约束边界附近的适应能力, 而条件扩散模型通过学习时间约束的概率分布, 能够更准确地捕捉司机对不同时间约束下的运单选择概率, 从而扩大了有效匹配的范围.

序列注意力机制的移除同样造成了性能退化, *MR* 相对下降 4.7%, *DS* 相对下降 2%. 这一结果强调了历史行为序列信息在司机偏好建模中的关键作用. 简单的 MLP 方法虽然能够处理当前运单特征, 但无法有效捕捉司机偏好的时序演化特征和上下文依赖关系. 序列注意力机制通过对历史运单序列的动态加权, 能够识别出与当前决策最相关的历史行为模式, 从而提升偏好预测的准确性.

时间约束捕捉模块的重要性在消融实验中表现最为突出, 其移除导致 *MR* 相对下降 8% (从 0.813 降至 0.748), *DS* 下降 12.1% (从 0.857 降至 0.754). 这一结果表明, 显式的时间约束建模是整个框架性能的基础. 没有时间约束捕捉模块, 模型无法理解司机在不同时间压力下的行为差异, 导致分配决策与司机实际接受意愿之间出现较大偏差.

动态目标权衡模块的消融实验揭示了其在平衡匹配率与司机满意度方面的重要作用. 当该模块被固定权重策略替代时, *MR* 相对下降 2.4% (从 0.813 降至 0.794), *DS* 相对下降 1.3% (从 0.857 降至 0.846). 这一结果验证了动态权重调整机制的有效性. 固定权重策略无法适应不同时间段内供需关系的变化, 而动态权衡机制能够根据实时的拒单反馈调整优化目标的相对重要性, 在高拒单率情况下增加司机满意度的权重, 在低拒单率情况下侧重提高匹配率, 从而实现两个目标的自适应平衡.

匹配网络的替换实验进一步证实了基于 PPO 的强化学习方法在序列决策优化中的优势. 将 PPO 替换为动态规划算法导致 *MR* 相对下降 5.5% (从 0.813 降至 0.769), *DS* 相对下降 3.7% (从 0.857 降至 0.826). 这意味着动态规划算法虽然能够在给定状态下找到局部最优解, 但缺乏对长期累积收益的考虑; 而 PPO 通过价值函数网络和策略网络的协同优化, 能够学习到跨时间窗口的最优分配策略, 在当前决策中考虑对未来匹配效果的潜在影响.

4.4 超参数实验 (RQ3)

TB-Match 的性能受多种超参数影响. 我们进行了一系列实验来分析它们的敏感性并确定合适的取值. 在这些

分析中, 主要的评估指标是 RS 数据集上的匹配率 (*MR*), 除非另有说明, 其他超参数均保持其默认最优值 (如第 5.1.4 节所述或通过先前调整确定).

4.4.1 核心架构与强化学习参数

我们首先研究了与模型架构和强化学习智能体相关的关键参数. 针对表 3 所展示的结果 (括号内数值代表参数值对应的匹配率 (*MR*)), 我们对各参数的敏感度进行分析.

表 3 核心架构与强化学习参数对匹配率 (*MR*) 的敏感性分析

参数	测试值及对应的 *MR*				
	值 1	值 2	最优/默认值	值 4	值 5
规划范围	2 (0.774)	4 (0.783)	8 (0.813)	10 (0.805)	12 (0.791)
PPO 裁剪 (ϵ_{clip})	0.1 (0.780)	0.15 (0.792)	0.2 (0.813)	0.25 (0.801)	0.3 (0.786)
折扣因子 (γ)	0.90 (0.771)	0.93 (0.794)	0.95 (0.813)	0.97 (0.785)	0.99 (0.778)

● 规划范围: 随着规划范围从 2 增加到 8, 智能体的性能得到改善, 表明更长远的预见性对决策制定有益. 规划范围为 8 时可获得最佳 *MR* 值 0.813. 将其进一步扩展到 10 时则表现出轻微下降, 这可能是由于学习复杂度增加或价值估计的方差增大所致. 因此, 为 TB-Match 选择了规划范围 8.

● PPO 裁剪 (ϵ_{clip}): 此参数约束策略更新步长. PPO 实现中常用的 ϵ_{clip} 值 0.2 获得了最高的 *MR* (0.813). 较小的值 (例如 0.1) 会减慢学习速度, 而较大的值 (例如 0.3) 可能导致过于激进的更新和不稳定性.

● 折扣因子 (γ): 此参数决定未来奖励的现值. γ 为 0.95 时提供了最佳 *MR* (0.813), 在重视即时结果和未来结果之间取得了平衡. 较低的值使智能体更短视, 而非常高的值在未来奖励高度不确定或遥远时会使学习变得困难.

4.4.2 学习过程中的超参数

我们还对学习过程的超参数进行了分析. 如表 4 所示, 对于偏好表示模块, 学习率 1×10^{-4} 是最优的. 对于匹配决策模块, 5×10^{-5} 产生了最佳结果. 这些值在收敛速度和稳定性之间提供了良好的权衡. 显著较高的学习率会导致不稳定, 而较低的学习率则导致收敛过慢.

表 4 学习过程参数对匹配率 (*MR*) 的敏感性分析

参数	测试值及对应的 *MR*				
	值 1	值 2	最优/默认值	值 4	值 5
偏好模块学习率	5E–5 (0.804)	8E–5 (0.809)	1E–4 (0.813)	2E–4 (0.804)	5E–4 (0.799)
匹配模块学习率	1E–5 (0.803)	3E–5 (0.808)	5E–5 (0.813)	8E–5 (0.805)	1E–4 (0.801)

4.5 鲁棒性实验 (RQ4)

除了最优条件外, 一个实用的匹配系统还必须能稳健应对各种现实世界中的运营挑战. 本节评估 TB-Match 在几种此类场景下的性能: 波动的市场动态 (需求高峰、司机短缺)、不均衡的资源分配 (地理失衡)、数据质量问题 (稀疏性、输入噪声) 以及可扩展性需求. 除非另有说明, 所有实验均在 RS 数据集上进行, 并将性能与关键基线方法进行比较.

4.5.1 市场波动下的韧性

现实世界的物流平台经常经历不可预测的供需变化. 我们模拟了两种常见场景. 1) 需求高峰: 运单量比平均水平增加 50%, 测试系统处理需求突然激增的能力. 2) 运力短缺: 可用运力数量减少 30%, 评估在运力供给受限情况下的性能.

后文表 5 总结了各种方法的性能保持情况. 性能保持率定义为挑战条件下指标值除以正常条件下指标值, 以百分比表示. 例如, 87.5% *MR* 表示模型保留了其原始匹配率的 87.5%. 最优保持率以粗体标出.

在需求高峰下, TB-Match 保持了其原始匹配率 (*MR*) 的 87.5%, 优于次优基线 MAPT (83.2%). 同时 TB-Match 也最有效地保持了司机满意度 (*DS*) (92.1% 的保持率). 这表明即使在需求很高的情况下, 它也能有效地分配运力.

在司机短缺的情况下, TB-Match 也表现出了优秀的鲁棒性, 保持了 83.2% 的 *MR*, 而 MAPT 为 79.5%. 同时司机满意度的保持率为 89.3%, 这表明 TB-Match 即使在司机较少的情况下也能更好地利用可用司机运力并维持服务水平. 基于 PPO 的长期价值估计可能有助于在资源受限的环境中做出更具战略性的分配.

4.5.2 运单分布失衡下的性能

物流需求通常在空间上是倾斜的. 我们通过将 70% 的运单集中在 30% 的区域, 而司机最初分布更均匀的方式模拟了运单分布失衡. 表 6 显示了 *MR* 的性能保持率以及空间均衡性 (SB, 一种基于基尼系数的指标, 越低越好, 此处采用得分 1 – Gini 的保持率) 和长期稳定性 (LTS, 即 *MR* 随时间变化的方差) 的绝对值. 最佳保持率以粗体标出.

表 5 市场波动鲁棒性 (%)

方法	需求高峰 (+50%运单)		司机短缺 (–30%司机)	
	*MR*保持率	*DS*保持率	*MR*保持率	*DS*保持率
GM	68.3	73.2	65.7	71.8
HA	72.1	78.5	69.4	75.3
GA	75.6	81.7	72.8	78.9
TPTA	78.3	84.2	75.1	81.4
FUPTA	80.7	86.5	77.3	83.8
MAPT	83.2	88.9	79.5	85.7
TB-Match	**87.5**	**92.1**	**83.2**	**89.3**

表 6 运单分布失衡鲁棒性 (%)

方法	*MR* 保持率	SB (1–Gini) 保持率	LTS 保持率
GM	71.2	58.3	63.5
HA	74.8	62.1	67.2
GA	77.5	65.7	70.8
TPTA	80.2	68.4	73.5
FUPTA	82.7	71.2	76.1
MAPT	84.9	73.8	78.4
TB-Match	**88.3**	**78.5**	**82.7**

在运单分布失衡条件下, TB-Match 保持了 88.3% 的 *MR*, 显著优于 MAPT (84.9%). 更重要的是, 它保留了 78.5% 的空间均衡性得分和 82.7% 的长期稳定性, 表明其在管理不均匀分布和引导司机前往需求热点而不过度空驶方面的有效性, 这可能得益于时空注意力协调机制和 PPO 智能体学到的长期价值.

4.5.3 数据稀疏性的影响 (冷启动司机)

新司机通常历史数据有限, 构成冷启动挑战. 我们通过将随机选定的司机子集 (占总司机数量的 10%、20% 和 30%) 的历史交互数据减少 75% 来模拟这种情况. 表 7 显示了 *MR* 相对于全数据性能的绝对下降百分点.

表 7 新司机数量占比鲁棒性 (%)

方法	10% 新司机	20% 新司机	30% 新司机
TPTA	−1.8	−3.5	−5.5
FUPTA	−1.6	−3.2	−5.1
MAPT	−1.4	−2.9	−4.6
TB-Match	**−0.9**	**−1.8**	**−2.9**

TB-Match 对数据稀疏的情况仍具有一定的韧性. 当 30% 的司机为新司机时, TB-Match 的 *MR* 仅下降了 2.9 个百分点, 而 MAPT 的 *MR* 下降了 4.6 个百分点. 这种鲁棒性源于: (1) 扩散模型能够学习时间约束偏好的普适模式, 这些模式不仅依赖于个体司机数据; (2) PPO 框架能够从可用的聚合数据中学习更广泛的市场动态和司机原型, 然后将其应用于冷启动司机. 序列注意力机制虽然受益于丰富的历史数据, 但仍可以利用新司机的通用特征.

4.5.4 系统规模的可扩展性

为了评估系统规模可扩展性, 我们在规模不断增加的合成数据集上测试了 TB-Match, 其数据模拟自 RS 数据集 (运单数从 5 万到 50 万, 司机数量按比例缩放). 我们测量了每批次的平均计算消耗时间和 *MR*.

表 8 表示不同数据集规模下的平均每批次决策时间与匹配率 (*MR*), 其最佳结果以粗体标出. 由结果可见, TB-Match 具有良好的可扩展性. 其每批次决策时间随数据集规模呈次线性增长, 这归因于其神经网络组件中的高效批处理以及 PPO 智能体在每个状态下固定复杂度的策略评估. 虽然 MAPT 略快, 但 TB-Match 在所有规模上都保持着显著更高的 *MR*. 例如, 在 50 万运单时, TB-Match 的 *MR* 为 91.2%, 计算消耗时间为 5 096 ms/批次, 而 MAPT 在 4 125 ms/批次下实现 87.5% 的 *MR*. 基于匈牙利算法的运单分配方法 (HA) 是一种组合方法, 表现出较差的时间

可扩展性, 对于大规模的运单分配任务消耗了较长的时间. TB-Match 稳定的 *MR* 表明其学习到的策略泛化良好, 并且该框架可以处理分配规模的大幅增加, 而不会产生过高的计算开销或匹配质量的显著损失.

表 8　可扩展性分析

数据集规模 (运单数) (k)	TB-Match		MAPT		HA	
	时间 (ms/批次)	*MR* (%)	时间 (ms/批次)	*MR* (%)	时间 (ms/批次)	*MR* (%)
50 (基线)	817	**92.3**	**624**	89.1	4573	81.5
100	1489	**92.1**	**1106**	88.7	8581	81.0
200	2270	**91.8**	**1958**	88.2	15402	80.2
500	5096	**91.2**	**4125**	87.5	33957	78.5

5 总　结

本文提出的 TB-Match 方法解决了网络货运平台运单分配中的两个关键问题: 刚性时间约束的弹性化建模以及匹配率与司机满意度的多目标优化. 该方法通过 4 个协同工作的技术模块实现了系统性能的提升: 基于条件扩散模型的概率化时间约束评估、融合多维特征的司机偏好表征、强化学习驱动的动态权重调节以及近端策略优化的长期决策生成. 在日均 30 000 吨货运量的实际运营场景中, TB-Match 方法实现了 17.66% 的匹配率提升, 验证了其在实际应用中的有效性. 未来的研究工作将从多个方向展开进一步拓展: 首先, 通过整合天气、路况等动态外部因素增强模型的环境适应性; 其次, 开发适用于冷启动场景的迁移学习策略, 优化新司机和新区域的分配效率; 最后, 探索可持续的智能分配范式, 推动物流平台向更高效、更智能的方向发展. 这些研究方向将进一步强化智能物流系统的鲁棒性和普适性.

References

[1] Ge Y, Xiong H, Tuzhilin A, Xiao KL, Gruteser M, Pazzani M. An energy-efficient mobile recommender system. In: Proc. of the 16th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. Washington: ACM, 2010. 899–908. [doi: 10.1145/1835804.1835918]

[2] Stiglic M, Agatz N, Savelsbergh M, Gradisar M. The benefits of meeting points in ride-sharing systems. Transportation Research Part B: Methodological, 2015, 82: 36–53. [doi: 10.1016/j.trb.2015.07.025]

[3] Zhao Y, Xia JF, Liu GF, Su H, Lian DF, Shang S, Zheng K. Preference-aware task assignment in spatial crowdsourcing. In: Proc. of the 33rd AAAI Conf. on Artificial Intelligence. Honolulu: AAAI, 2019. 2629–2636. [doi: 10.1609/aaai.v33i01.33012629]

[4] Kuhn HW. The Hungarian method for the assignment problem. Naval Research Logistics Quarterly, 1955, 2(1–2): 83–97. [doi: 10.1002/nav.3800020109]

[5] Jonker R, Volgenant A. A shortest augmenting path algorithm for dense and sparse linear assignment problems. Computing, 1987, 38(4): 325–340. [doi: 10.1007/BF02278710]

[6] Burkard RE, Dell'Amico M, Martello S. Assignment Problems. Philadelphia: SIAM, 2009.

[7] Bertsekas DP. The auction algorithm: A distributed relaxation method for the assignment problem. Annals of Operations Research, 1988, 14(1): 105–123. [doi: 10.1007/BF02186476]

[8] Pisinger D, Ropke S. A general heuristic for vehicle routing problems. Computers & Operations Research, 2007, 34(8): 2403–2435. [doi: 10.1016/j.cor.2005.09.012]

[9] Lee S, Boomsma TK. An approximate dynamic programming algorithm for short-term electric vehicle fleet operation under uncertainty. Applied Energy, 2022, 325: 119793. [doi: 10.1016/j.apenergy.2022.119793]

[10] Yan CW, Zhu HL, Korolko N, Woodard D. Dynamic pricing and matching in ride-hailing platforms. Naval Research Logistics (NRL), 2020, 67(8): 705–724. [doi: 10.1002/nav.21872]

[11] Zhu TC, Qiu Y, Zhou HY, Li JX. Decoding global preferences: Temporal and cooperative dependency modeling in multi-agent preference-based reinforcement learning. In: Proc. of the 38th AAAI Conf. on Artificial Intelligence. Vancouver: AAAI, 2024. 17202–17210. [doi: 10.1609/aaai.v38i15.29666]

[12] Zhen XH, Long J, Cai ZP. A survey of order dispatch policy based on online ride-hailing services. Computer Engineering and Science, 2020, 42(7): 1267–1275 (in Chinese with English abstract). [doi: 10.3969/j.issn.1007-130X.2020.07.016]

[13] Tong YX, Yuan Y, Cheng YR, Chen L, Wang GR. Survey on spatiotemporal crowdsourced data management techniques. Ruan Jian Xue Bao/Journal of Software, 2017, 28(1): 35–58 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5140.htm [doi: 10.13328/j.cnki.jos.005140]

[14] Li YG, Yu R, Shahabi C, Liu Y. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In: Proc. of the 6th Int'l Conf. on Learning Representations. Vancouver: OpenReview.net, 2018.

[15] Xu Z, Li ZX, Guan QW, Zhang DS, Li Q, Nan JX, Liu CY, Bian W, Ye JP. Large-scale order dispatch in on-demand ride-hailing platforms: A learning and planning approach. In: Proc. of the 24th ACM SIGKDD Int'l Conf. on Knowledge Discovery & Data Mining. London: ACM, 2018. 905–913. [doi: 10.1145/3219819.3219824]

[16] Xie Y, Wang YH, Li KL, Zhou X, Liu Z, Li KQ. Satisfaction-aware task assignment in spatial crowdsourcing. Information Sciences, 2023, 622: 512–535. [doi: 10.1016/j.ins.2022.11.081]

[17] Bertsimas D, Jaillet P, Martin S. Online vehicle routing: The edge of optimization in large-scale applications. Operations Research, 2019, 67(1): 143–162. [doi: 10.1287/opre.2018.1763]

[18] Deb K, Pratap A, Agarwal S, Meyarivan T. A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Trans. on Evolutionary Computation, 2002, 6(2): 182–197. [doi: 10.1109/4235.996017]

[19] Li BD, Li JL, Tang K, Yao X. Many-objective evolutionary algorithms: A survey. ACM Computing Surveys (CSUR), 2015, 48(1): 13. [doi: 10.1145/2792984]

[20] Qi YT, Ma XL, Liu F, Jiao LC, Sun JY, Wu JS. MOEA/D with adaptive weight adjustment. Evolutionary Computation, 2014, 22(2): 231–264. [doi: 10.1162/EVCO_a_00109]

[21] Dong ZM, Wang XP, Tang LX. MOEA/D with a self-adaptive weight vector adjustment strategy based on chain segmentation. Information Sciences, 2020, 521: 209–230. [doi: 10.1016/j.ins.2020.02.056]

[22] Beume N, Naujoks B, Emmerich M. SMS-EMOA: Multiobjective selection based on dominated hypervolume. European Journal of Operational Research, 2007, 181(3): 1653–1669. [doi: 10.1016/j.ejor.2006.08.008]

[23] Zheng W, Doerr B. Runtime analysis of the SMS-EMOA for many-objective optimization. In: Proc. of the 2024 AAAI Conf. on Artificial Intelligence. Vancouver: AAAI, 2024. 20874–20882. [doi: 10.1609/aaai.v38i18.30077]

[24] Lyu GD, Cheung WC, Teo CP, Wang H. Multi-objective stochastic optimization: A case of real-time matching in ride-sourcing markets. Manufacturing and Service Operations Management, 2024, 26(2): 500–518. [doi: 10.1287/msom.2020.0247]

[25] Zhou F, Lu CF, Tang XC, Zhang F, Qin ZW, Ye JP, Zhu HT. Multi-objective distributional reinforcement learning for large-scale order dispatching. In: Proc. of the 2021 IEEE Int'l Conf. on Data Mining. Auckland: IEEE, 2021. 1541–1546. [doi: 10.1109/ICDM51629.2021.00202]

[26] Teh YW, Bapst V, Czarnecki WM, Quan J, Kirkpatrick J, Hadsell R, Heess N, Pascanu R. Distral: Robust multitask reinforcement learning. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 4499–4509.

[27] Balke G, Adenaw L. Heavy commercial vehicles' mobility: Dataset of trucks' anonymized recorded driving and operation (DT-CARGO). Data in Brief, 2023, 48: 109246. [doi: 10.1016/j.dib.2023.109246]

[28] Li X, Zhang XL. Multi-task allocation under time constraints in mobile crowdsensing. IEEE Trans. on Mobile Computing, 2021, 20(4): 1494–1510. [doi: 10.1109/TMC.2019.2962457]

[29] Ma Y, Ma L, Gao XF, Chen GH. Fused user preference learning for task assignment in mobile crowdsourcing. In: Proc. of the 21st Int'l Conf. on Service-oriented Computing. Rome: Springer, 2023. 227–241. [DOI: 10.1007/978-3-031-48424-7_17]

附中文参考文献

[12] 郑小红, 龙军, 蔡志平. 关于网约车订单分配策略的综述. 计算机工程与科学, 2020, 42(7): 1267–1275. [doi: 10.3969/j.issn.1007-130X.2020.07.016]

[13] 童咏昕, 袁野, 成雨蓉, 陈雷, 王国仁. 时空众包数据管理技术研究综述. 软件学报, 2017, 28(1): 35–58. http://www.jos.org.cn/1000-9825/5140.htm [doi: 10.13328/j.cnki.jos.005140]

作者简介

廖家俊, 博士生, 主要研究领域为数据驱动的物流智能决策.

董宜滔, 博士生, CCF 学生会员, 主要研究领域为物流决策优化, 强化学习.

毛嘉莉, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为智能决策理论与技术, 数据治理.