分类号：　　密　　级：公开
学　号：20232108023　　单位代码：10759

石河子大学

硕士学位论文

基于元学习和因果强化的短视频推荐方法研究及实践

学位申请人　王梦笛
指导教师　刘长征　教授

申请学位类别　专业硕士
专业名称　电子信息
研究领域　计算机技术
所在学院　信息科学与技术学院

中国·新疆·石河子
2026 年 5 月

分类号：　　密　　级：公开

学　号：20232108023　　单位代码：10759

石河子大学

硕 士 学 位 论 文

基于元学习和因果强化的短视频推荐方法研究及实践

学 位 申 请 人	王梦笛
指 导 教 师	刘长征　教授
申 请 学 位 类 别	专业硕士
专 业 名 称	电子信息
研 究 领 域	计算机技术
所 在 学 院	信息科学与技术学院

中国·新疆·石河子

2026 年 5 月

Research and Implementation of Short Video Recommendation Methods Based on Meta-Learning and Causal Reinforcement Learning

A Dissertation Submitted to

Shihezi University

In Partial Fulfillment of the Requirements

for the Degree of

Master of Engineering

By

Wang Meng-di
(Electronic Information)

Dissertation Supervisor:Prof.Liu Chang-zheng

May,2026

石河子大学学位论文独创性声明及使用授权声明

学位论文独创性声明

本人所呈交的学位论文是在我导师的指导下进行的研究工作及取得的研究成果。据我所知，除文中已经注明引用的内容外，本论文不包含其他个人已经发表或撰写过的研究成果。对本章的研究做出重要贡献的个人和集体，均已在文中作了明确的说明并表示谢意。

研究生签名：王梦笛　　时间：　2026 年 5 月 25 日

使用授权声明

本人完全了解石河子大学有关保留、使用学位论文的规定，学校有权保留学位论文并向国家主管部门或指定机构送交论文的电子版和纸质版。有权将学位论文在学校图书馆保存并允许被查阅。有权自行或许可他人将学位论文编入有关数据库提供检索服务。有权将学位论文的标题和摘要汇编出版。保密的学位论文在解密后适用本规定。

研究生签名：王梦笛　　时间：　2026 年 5 月 25 日

导师签名：刘长江　　时间：　2026 年 5 月 25 日

摘要

近年来，移动互联网推动抖音、快手短视频平台快速发展，已成为公众娱乐的重要渠道。用户规模与内容数量的爆发式增长加剧了信息过载问题，推荐系统虽然有效提升了内容分发效率，但仍面临挑战：一是针对新用户缺乏历史行为导致的冷启动问题；二是对历史偏好的过度依赖易引发内容同质化与“信息茧房”，损害了内容多样性与用户长期体验。为解决上述问题，本文主要开展了以下研究：

（1）针对新用户交互数据稀疏导致的冷启动问题，提出了一种基于多层次粗细粒度偏好增强的元学习推荐方法（MCFPEM）。该方法在元学习框架下，从“评分级”和“类别级”两个层面构建了多层次用户偏好画像，设计粗细粒度偏好融合机制，并引入相似用户群体信息以增强特征表征抗噪性。在此基础上，构建了增强型偏好特定适配器，将多层次用户表示映射至模型参数空间，通过门控机制对预测网络参数进行动态调制，实现模型对不同用户任务的快速个性化适应。实验结果表明，相较于最优基线模型，MCFPEM 在 MovieLens 数据集上的 MAE 降低了 0.41%、NDCG@5 提升了 0.60%；在 Douban Book 数据集上的 MAE 降低了 0.50%、NDCG@5 提升了 0.50%。

（2）针对老用户交互数据丰富但推荐结果易陷入“信息茧房”、影响内容多样性与用户长期体验的问题，提出了一种基于因果偏好建模与离线强化学习的推荐方法（CPMORL）。在用户模型预学习阶段，以 DeepFM 为骨干网络，集成对比学习与数据增强策略以提升用户-项目特征表示的判别能力，并结合因果推断方法实现对用户真实偏好的无偏估计；在强化学习策略优化阶段，基于 PPO 算法构建离线强化学习框架，引入反事实分层预测机制与成本敏感奖励函数作为保守性约束，将内容同质化风险显式转化为惩罚项，引导策略在离线数据分布支持范围内进行多样性动作采样，抑制 OOD 奖励高估问题，实现推荐准确性与长效多样性的平衡。实验结果表明，在衡量交互式推荐系统长期价值的最核心指标累积奖励R_{tra}上，CPMORL 在 KuaishouRec 与 KuaiRand_Pure 数据集上相较于各自最优基线分别提升了 33.54%和 6.39%。

（3）为了验证上述两种推荐方法在实际应用场景中的可行性与有效性，将 MCFPEM 与 CPMORL 方法集成，设计并实现了一套完整的短视频推荐系统。系统包括用户端的注册登录、个性化推荐、分类浏览、搜索与行为交互，以及管理员端的用户管理、视频管理等核心功能。系统根据用户交互数据规模自动识别用户类型，并动态调用相应的推荐策略，以实现针对新老用户的差异化推荐服务。通过系统功能测试与运行验证，结果表明该系统能够稳定运行并提供高效的个性化推荐服务，整体性能良好，达到了预期设计目标。

关键词：短视频推荐；元学习；因果强化学习；冷启动；信息茧房

Abstract

In recent years, the rapid development of short video platforms such as Douyin and Kuaishou, driven by the mobile internet, has made them important channels for public entertainment. The explosive growth in user scale and content quantity has exacerbated the problem of information overload. Although recommendation systems have effectively improved content distribution efficiency, they still face challenges: first, the cold-start problem caused by the lack of historical behavior data for new users; second, the over-reliance on historical preferences, which can lead to content homogenization and the "information cocoon," thereby damaging content diversity and the long-term user experience. To address these issues, this thesis primarily conducts the following research:

(1) To address the cold-start problem caused by the sparsity of interaction data for new users, a meta-learning recommendation method based on multi-level coarse- and fine-grained enhanced preferences (MCFPEM) is proposed. Within the meta-learning framework, this method constructs multi-level user preference profiles from both "rating-level" and "category-level" perspectives, designs a coarse- and fine-grained preference fusion mechanism, and incorporates information from similar user groups to enhance the robustness against noise in feature representations. Building upon this, an enhanced preference-specific adapter is constructed to map the multi-level user representations into the model parameter space. By dynamically modulating the predictive network parameters via a gating mechanism, it enables the model to achieve rapid personalized adaptation to diverse user tasks. Experimental results demonstrate that, compared with the optimal baseline models, MCFPEM reduces the MAE by 0.41% and improves the NDCG@5 by 0.60% on the MovieLens dataset; furthermore, on the Douban Book dataset, the MAE is reduced by 0.50% and the NDCG@5 is improved by 0.50%.

(2) To address the issue that existing users with abundant interaction data are prone to being trapped in an "information cocoon"—which negatively impacts content diversity and long-term user experience—a recommendation method based on Causal Preference Modeling and Offline Reinforcement Learning (CPMORL) is proposed. In the user model pre-learning stage, adopting DeepFM as the backbone network, contrastive learning and data augmentation strategies are integrated to enhance the discriminative capability of user-item feature representations. Coupled with causal inference techniques, it achieves an unbiased estimation of users' true preferences. In the reinforcement learning strategy optimization stage, an offline reinforcement learning framework is constructed based on the PPO algorithm. A counterfactual stratified prediction mechanism and a cost-sensitive reward function are introduced as conservative constraints, explicitly transforming the risk of content homogenization into a penalty term. This mechanism guides the policy to conduct diverse action sampling within the support set of the offline data distribution, thereby

effectively suppressing the Out-of-Distribution (OOD) reward overestimation problem and achieving a balance between recommendation accuracy and long-term diversity. Experimental results demonstrate that, in terms of cumulative reward (R_{tra})—the most critical metric for evaluating the long-term value of interactive recommender systems—CPMORL achieves improvements of 33.54% and 6.39% over the best-performing baselines on the KuaishouRec and KuaiRand_Pure datasets, respectively.

(3) To verify the feasibility and effectiveness of the two proposed recommendation methods in practical application scenarios, we integrate the MCFPEM and CPMORL methods to design and implement a comprehensive short video recommendation system. The system's core functions encompass registration and login, personalized recommendation feeds, category browsing, search, and behavioral interactions on the user end, as well as user and video management on the administrator end. Based on the volume of user interaction data, the system automatically identifies user types and dynamically invokes the corresponding recommendation strategies, thereby delivering differentiated recommendation services for new and existing users. Functional testing and operational validation demonstrate that the system operates stably, provides highly efficient personalized recommendations, and exhibits excellent overall performance, successfully achieving the anticipated design objectives.

Key words: Short video recommendation; Meta-learning; Causal reinforcement learning; Cold start; Information cocoon

目　录

第 1 章 绪论

1.1 研究背景及意义

1.1.1 研究背景

随着互联网技术的飞速迭代更新，用户对网络资源的个性化需求日益增长，但与之相伴的推荐服务质量问题也逐渐凸显。互联网规模和数据井喷式的扩张使得准确定位目标数据愈加困难，用户在获取有价值信息和进行决策时付出的成本不断提高，最终对用户体验造成负面影响。推荐系统作为过滤信息的有效工具，通过分析用户过往的交互行为数据来刻画其兴趣特征，从而预测用户对特定项目的偏好程度，并生成其可能感兴趣的 Top-K 候选列表，从而实现对冗余信息的筛选与个性化内容推荐，有效缓解信息过载[1]。目前，推荐系统已经广泛应用于多个领域，如电商、社交网络、新闻推送、短视频等。例如在电商平台上，推荐系统能够为用户推荐感兴趣的商品，提高购买转化率和用户满意度；在新闻推荐中，推荐系统可结合用户历史阅读记录，推送符合其兴趣的个性化内容[2]。短视频作为新兴的内容形态，凭借内容丰富多样、短时长、互动性强等特点，迅速成为广大网民日常生活中重要的娱乐与信息获取方式。

在短视频领域中，推荐方法的研究更是至关重要。由于短视频内容形式复杂且数据规模庞大，系统需要在极短时间内精准筛选出符合用户兴趣的内容。尽管经典的协同过滤和内容推荐方法通过分析历史交互与物品属性，在传统领域表现尚可，但在处理短视频时却显得力不从心。为了解决这些问题，研究者们一直探索各种短视频推荐算法，通过综合分析用户行为模式、个体特征以及内容属性等多维信息，提升推荐结果的准确性与个性化水平[3]。人工智能的快速发展为短视频推荐提供了新的技术支撑，借助深度学习和大数据分析，系统可以从用户行为、兴趣特征及视频属性等多个角度进行建模，实现大规模内容的快速筛选与个性化推送。

1.1.2 研究意义

随着数字化发展不断深化并与互联网技术紧密结合，短视频推荐系统的研究具有重要的理论价值与现实意义。在理论层面，对该类系统的设计与改进反映了信息检索、机

器学习与人工智能等领域的进一步探索与发展。针对用户行为建模与内容理解等问题的深入研究，对丰富和发展个性化推荐的理论和方法具有重要的科学意义和应用价值。实践上，借助深度学习和协同过滤等方法，推荐系统能快速识别用户兴趣，实现内容的精准推送，有效缓解信息过载带来的决策压力，提高用户活跃度与平台粘性，从而构建更加丰富多彩、沉浸式的短视频内容环境。然而，在实际应用场景中，系统往往面临着大量新注册用户或低活跃度用户，这些用户仅有极少甚至没有任何交互记录。这种因数据极度稀疏导致模型难以准确刻画用户意图，形成典型的冷启动问题。同时，对于具有充足行为数据的老用户，推荐算法过度依赖历史兴趣进行精准匹配，容易强化既有偏好路径，进而加剧“信息茧房”现象，限制内容多样性与用户视野拓展。因此，设计并实现一个兼顾“新用户冷启动”与“老用户去茧房”的个性化短视频推荐系统，不仅是提升推荐效果与用户体验的关键技术路径，也是实现推荐系统可持续发展的重要研究方向。

1.2 国内外研究现状

推荐算法的研究可追溯至上世纪 90 年代，随着互联网技术的迅猛发展与数据规模的指数级增长，推荐系统逐渐成为互联网应用的重要组成部分。推荐算法经历了从基于内容与协同过滤的传统方法，到融合深度学习、图模型、强化学习以及大型语言模型等先进技术的不断演进和完善，以更好地适应复杂多变的应用场景与多样化用户需求。在这一发展过程中，研究重点逐渐由简单的特征匹配转向高维表示学习与序列决策优化，更加关注用户长期兴趣、结构化关系信息以及因果影响的建模能力。特别是在在数据稀疏、长期收益优化等问题日益突出的背景下，元学习、因果推断和离线强化学习逐渐成为学术界与工业界关注的研究焦点。考虑到上述技术方向与本文研究内容密切相关，后续将围绕元学习、因果推断及离线强化学习领域的代表性推荐方法研究现状进行介绍。

1.2.1 基于元学习的推荐方法

元学习的核心思想是“学会学习”，即利用大量历史任务的经验学习一种通用的初始化参数或优化策略，使模型以很少的样本即可快速适应新任务[4]。这一特性为缓解推荐系统中的数据稀疏与冷启动问题提供了新思路。在基于优化的元学习框架中，Finn 等人[5]提出的模型无关元学习 MAML 框架应用最为广泛。该方法通过学习具有良好泛化能力的初始化参数，使模型在面对新任务时仅需少量梯度更新即可完成快速适应。MAML 由内外双层循环构成，以任务作为训练数据的单位，内循环使用梯度下降最小化损失得到每个任务的局部最优参数来梯度更新初始化参数 θ[1]。虽然 MAML 在多个领

域表现出色，然而，MAML 需要计算二阶梯度，训练开销较高，且统一初始化在复杂任务分布下可能限制模型的泛化能力。为降低计算复杂度，在基于 MAML 的思想上 Nichol 等人[6]进一步提出 FOMAML 算法，通过一阶梯度近似替代二阶梯度计算，从而显著提升训练效率，但也在一定程度上损失了模型的适应能力。随着元学习方法在推荐系统中的不断发展，研究者逐渐将其应用于用户冷启动问题。Bharadhwaj[7]提出一个基于 MAML 的推荐框架 MetaCS 缓解用户冷启动，通过构建基于 MAML 的元学习任务来模拟新用户推荐场景，并利用用户评分数据构建二分类兴趣标签，从而实现对新用户偏好的快速建模。然而，该方法在构建元任务时采用固定数量的历史交互样本，与真实推荐场景中用户行为分布存在一定差异。在利用历史交互物品评估用户偏好的基础上，Wei 等人[8]提出了一种 MAML 的学习范式 MetaCF，将元学习机制与协同过滤模型相结合，并通过动态子图采样构建元学习任务，使模型能够在训练阶段学习适应新用户推荐的能力。Lee 等人[9]提出了一种基于元学习的推荐系统模型 MELU，通过利用少量用户消费记录快速预测用户兴趣，从而有效缓解用户冷启动问题。针对推荐系统中用户反馈分布不均衡的问题，Kim 等人[10]提出了一种新的基于梯度的元学习顺序推荐框架 MELO，通过设计用户特定的自适应损失函数，使模型能够更好地捕捉用户评分分布的不平衡特征，从而提高推荐效果。

传统 MAML 采用的逐任务（task-by-task）孤立学习范式容易导致感受野受限和局部最优。针对这一局限，Du 等人[11]的 CCML 框架提出了一种跨任务协作元学习策略，通过协同任务采样模块筛选出相关且有用的任务，并采用双层跨任务元训练策略利用多任务间的协同知识来增强用户建模，从而提升冷启动推荐的精度与模型鲁棒性。Guo 等人[12]提出的 MACRec 模型在异构学术网络上设计了多视图任务构造器，通过语义级和任务级的元学习器自适应，解决了极度稀疏网络下的冷启动难题。在更广泛的应用方向上，元学习逐渐与深度特征交互和序列建模相结合。Wang 等人[13]提出了一种用于预热冷启动新物品的点击率预测的 EmerG 模型。该方法利用超网络和图神经网络生成特定于物品的特征交互图，并设计专门的元学习策略在不同物品的 CTR 预测任务间优化参数，大幅降低了长尾冷启动物品过拟合的风险。在序列推荐方向，Wang 等人[14]创新性地提出了基于反事实任务增强的元学习方法，通过干预用户历史生成反事实序列，联合优化真实与反事实任务损失。Wang 等人[15]提出了一种名为 BOOML 的新颖框架，利用元学习促成共享知识在相似优化任务之间的传递，从而加速多目标贝叶斯优化的收敛速度。为了缓解多目标推荐中普遍存在的梯度冲突问题，模型在元学习的外层循环中引入了“正交梯度下降策略”，避免了冲突梯度的无效更新，从而取得了更好的模型性能。更具突破性的是，随着 LLM 在推荐系统的全面普及，Zhao 等人[16]将元学习用于大模型参数高效的提示微调。通过一阶和二阶元优化策略学习代表用户行为先验的软提示嵌入，该框架在消费级 GPU 上实现了毫秒级的快速适应，为大模型时代的零历史用户冷启动

个性化推荐开辟了极具潜力的全新范式。虽然现有基于元学习的推荐方法（如 MetaCS、MELU 等）在实现模型参数的快速适配与缓解冷启动问题上取得了显著成效，但这些方法大多依赖单一层面的特征交互，在面临短视频冷启动阶段极端稀疏和高噪的数据环境时，缺乏对高阶宏观语义与微观细粒度行为特征的联合建模。这种单一视角的建模方式导致模型在极端冷启动下的抗噪能力与表示鲁棒性不足，难以精准刻画新用户的真实意图，进而限制了元学习器生成个性化参数的质量。

1.2.2 基于因果推断的推荐方法

因果推断作为一种统计学分析手段，致力于以更科学、规范的方式揭示变量间存在的因果关联[17]。因果推断为推荐系统提供了一种强大的工具，使得能够通过审查推荐结果的生成过程来确定偏差的根本原因，并通过因果推荐建模减轻偏差的影响，提高推荐结果的准确性和公平性。因此，推荐系统中的研究人员越来越多地利用因果推断来提升性能，解决数据偏见、处理缺失和噪声，并超越单纯的准确性[18]。Swaminathan 等人[19]在 2017 年首次在推荐系统中引入因果关系这个概念，去解决推荐系统中存在的偏差问题，通过将推荐问题建模为一个因果推断问题，提出了一种新的评估方法，可以更准确地估计推荐系统的性能，从而减少因推荐偏差带来的误差。随后，研究者开始关注不同类型的行为偏差。例如，Zhang 等人[20]研究了音乐流媒体推荐系统中的注意力偏差问题，指出用户自动播放行为可能导致系统错误地学习用户偏好。为此，该研究提出了一种基于反事实学习的推荐方法，通过神经决斗老虎机算法区分用户真实反馈与自动播放产生的伪反馈，从而有效缓解注意力偏差问题。针对长视频因时长获得更多曝光的“时长偏差”，Zheng 等人[21]对时长偏差进行了深入分析和验证，提出了新的无偏评估指标“观看时长增益”和解决方案来减轻时长偏差，更好地平衡推荐性能。随后，Zhan 等人[22]利用因果图分析视频时长作为混杂变量对推荐结果的影响，并通过数据分割与直接建模的方法消除时长偏差。Wang 等人[23]利用因果建模与表示解耦方法分析用户偏好转移问题，但该方法在处理噪声行为方面仍存在一定局限。Bin 等人[24]提出了一种结合因果推断与数据增强的多行为推荐去偏框架 CIDA，通过“流行度感知噪声加权”和“模拟非热门项目增强”两个模块缓解流行度偏差，并通过对比学习增强模型对无偏特征的学习能力。另一类研究关注推荐系统中的混杂效应。Sato 等人[25]通过样本重新加权的方法处理用户属性与项目属性带来的混杂因素，并利用个体处理效应（ITE）估计推荐行为的因果影响。

近年来，随着因果学习与深度学习技术的发展，因果推断在推荐系统中的应用进一步深化。Luo 等人[26]对因果推荐方法进行了系统综述，从统一的因果理论视角对当前研究进行系统性梳理，提出以潜在结果框架、结构因果模型、反事实为核心的理论驱动分

类法，为复杂场景下的因果推荐提供了统一视角的理论框架。在因果建模方面，Zan 等人[27]提出一种因果感知社交推荐模型，将推荐问题建模为多处理因果推断问题，并利用社交网络结构建模用户之间的同质性关系，从而实现对混杂变量的有效建模。Song 等人[28]提出一种基于反事实推理的双塔会话推荐模型，通过构建反事实因果框架同时建模用户长期兴趣与短期会话行为，从而减少虚假相关关系对推荐结果的影响。在序列推荐与大语言模型的深度结合上，反事实与倾向匹配技术展现了极大的潜力。针对序列推荐系统仅能利用“已交互物品”而忽略大量“已曝光未交互”数据的缺陷，Zhao 等人[29]提出了一种面向序列推荐的、基于系统曝光的反事实增强方法 CaseRec，引入了离线强化学习与系统曝光的反事实增强技术，通过 Transformer 用户模拟器重构反事实曝光序列，极大拓宽了对用户潜在兴趣的探索边界。Yu 等人[30]提出了 LDPE 框架，将大语言模型捕获的丰富语义与因果推断协同，从用户侧和物品侧同时估计包含语义信息的“时间感知双重倾向得分”，打破了传统方法仅关注物品端曝光偏差的局限。

现有基于因果推断的推荐方法在剥离流行度、时长与曝光等偏差，还原用户真实的无偏偏好方面展现出了强大能力。然而，这些研究大多局限于静态的、单步的 CTR 预测或排序任务，缺乏对多轮动态交互中用户长期体验演变的关注。因果推断虽能“去偏”，却难以单独胜任长期的序列决策优化任务，这导致其在面对真实短视频场景中随时间累积的“信息茧房”效应时，无法从策略下发的机制上对内容同质化趋势进行持续、长效的干预。

1.2.3 基于离线强化学习的推荐方法

批量强化学习[31]（Batch Reinforcement Learning, BRL）是强化学习领域的一个重要分支，其核心思想是在固定数据集上学习策略，而无需与环境进行在线交互。随着实际应用中利用历史数据训练强化学习模型的需求不断增加，Levine 等人[32]在总结 BRL 研究的基础上，系统地提出了离线强化学习（Offline Reinforcement Learning, Offline RL）。由此明确指出了该学习面临的“分布偏移”等核心挑战，并提出了初步的解决思路[33]。Fujimoto 等人[34]受到变分自动编码器 VAE 架构的启发，提出批量约束 Q 学习算法，该方法通过生成模型学习行为策略分布，并限制策略仅在数据支持区域内进行动作选择，从而有效减少外推误差。在序列推荐场景下，为了解决应用离线强化学习所带来的分布偏移问题，Chen 等人[35]提出利用倾向得分来执行离线策略校正，但这些方法在估计的倾向得分中存在很大的差异。另一类方法采用基于模型的强化学习方法，通过构建环境模型来模拟用户行为[36]，利用模拟交互数据训练推荐策略。

随着离线强化学习技术的快速发展，研究者开始探索更加高效和稳定的推荐算法。在将离线强化学习重构为条件序列建模方面，研究者彻底改变了传统的基于时间差分更

新易产生自举误差的问题。Chen 等人[37]提出 EDT4Rec 模型，该框架引入了最大熵探索策略与奖励重标记技术，有效克服了传统决策 Transformer 缺乏“拼接次优轨迹”能力的致命弱点，挖掘出了大量子优数据中的价值。为了平衡推荐中的多目标冲突，Wang 等人[38]提出一种面向奖励驱动推荐的多目标决策 Transformer 框架 MODT4R，将不同目标的期望收益作为条件变量输入，通过监督学习的范式在准确率、多样性和新颖性之间实现了平滑的动态权衡。最为前沿的趋势是利用大语言模型（LLM）充当环境模拟器，以从根本上弥补离线数据的固有限制。Wang 等人[39]提出了一个名为 LE 的框架，利用大语言模型极强的理解能力，将其微调并用作离线强化学习的“交互环境与奖励模型”。它通过合成高质量的用户状态、奖励和扩充积极动作，大幅减少了离线强化学习对庞大真实交互数据的依赖。紧接着，Zhang 等人[40]进一步推出了大模型驱动的高效用户模拟器，为离线 RL 代理提供了保真度极高的多轮因果交互测试平台。结合因果推断，Wang 等人[41]提出策略引导的因果表示学习 PGCR 方法，通过策略引导的因果特征选择筛选并保留对奖励有因果影响的分量，再用编码器学习只聚焦因果相关信息的紧凑状态表示，该方法大幅提升了离线学习的稳定性和鲁棒性。

离线强化学习为推荐系统提供了一种低风险、低成本的长效决策优化框架，在缓解由分布偏移引起的分布外（Out-of-Distribution, OOD）动作外推误差问题上取得了重要进展。然而，现有多数离线强化学习推荐模型（如 ROLeR 等）往往单纯以最大化累积点击或交互时长作为奖励目标，缺乏对同质化推荐行为的显式约束。这使得智能体为了规避对 OOD 动作的价值高估风险，极易倾向于过度推荐离线支持集中高频、安全但高度同质化的项目。这种缺乏多样性引导的保守性倾向，极易使推荐策略退化为“安全但单一”的局部最优，非但未能打破既有偏好路径，反而加速了“信息茧房”的形成，严重损害了用户的长期满意度。

1.2.4 研究评述

综上所述，针对推荐方法的研究，研究者们不断探索新的方法和模型。近年来，元学习、因果推断以及离线强化学习在推荐系统领域取得了显著进展，为解决数据稀疏、偏差消除与长期收益优化等问题提供了重要思路。然而，在真实的短视频推荐系统中，用户行为呈现出明显的阶段性特征，不同生命周期阶段的用户面临的核心挑战截然不同。现有相关工作在应对复杂场景的具体痛点时，仍存在针对性不足的问题，具体表现在以下两个方面：

（1）对于新用户群体（冷启动挑战），现有元学习方法在极端稀疏数据下缺乏多层次特征刻画，导致泛化受限。尽管元学习能够通过参数快速适配实现模型迁移，但以 MetaCS、MELU 为代表的方法在构建元任务时，往往仅依赖单一层面的特征进行用户