E-Commerce Letters 电子商务评论, 2026, 15(6), 335-345
Published Online June 2026 in Hans. https://www.hanspub.org/journal/ecl
https://doi.org/10.12677/ecl.2026.156641

Hans汉斯

基于强化学习与动态异构图的电商跨域商品序列推荐方法

王则林[1]，王　祥[1]，赵书娴[1,2*]，程　实[1]

[1]南通大学人工智能与计算机学院，江苏　南通
[2]江苏汇环环保科技有限公司，江苏　南通

收稿日期：2026年3月14日；录用日期：2026年3月26日；发布日期：2026年6月16日

摘　要

本文面向电商领域跨域商品推荐场景，针对目标域商品交互数据稀疏、用户长期偏好难以有效建模与优化的问题，在动态异构图表征与多视图对齐学习的基础上，将推荐任务建模为带迁移代价约束的受限马尔可夫决策过程，提出离线强化学习方法DHGCRL。模型以源域局部图、目标域局部图与跨域全局异构图生成的三视图表征构造状态，并结合候选集约束与SlateQ列表价值近似实现Top-K商品列表决策。奖励函数由目标域商品反馈与跨域对齐相似度塑形项共同构成，以增强迁移表征对目标域商品偏好的刻画能力。同时，采用推荐列表表征与源域稳定兴趣表征之间的余弦距离定义迁移代价，通过联合回报价值与代价价值形成修正价值，从而指导策略学习并满足代价阈值约束。为提升离线训练的稳定性，进一步引入保守价值学习与行为一致性正则。实验结果表明，DHGCRL在HR和NDCG等指标上均优于多种序列推荐与离线强化学习基线方法。

关键词

跨域序列推荐，动态异构图，离线强化学习，迁移代价约束

A Reinforcement Learning and Dynamic Heterogeneous Graph-Based Method for Cross-Domain Product Sequential Recommendation in E-Commerce

Zelin Wang[1], Xiang Wang[1], Shuxian Zhao[1,2*], Shi Cheng[1]

[1]School of Artificial Intelligence and Computer Science, Nantong University, Nantong Jiangsu
[2]Jiangsu Huihuan Environmental Protection Technology Co., Ltd., Nantong Jiangsu

[*]通讯作者。

文章引用: 王则林, 王祥, 赵书娴, 程实. 基于强化学习与动态异构图的电商跨域商品序列推荐方法[J]. 电子商务评论, 2026, 15(6): 335-345. DOI: 10.12677/ecl.2026.156641

Received: March 14, 2026; accepted: March 26, 2026; published: June 16, 2026

Abstract

In the context of cross-domain product recommendation in e-commerce, this study addresses the sparsity of target-domain interaction data and the difficulty of modeling and optimizing users' long-term preferences. Building on dynamic heterogeneous graph representations and multi-view alignment learning, the recommendation process is formulated as a constrained Markov decision process with transfer cost constraints, and an offline reinforcement learning method, DHGCRL, is proposed. The model constructs the state from three-view representations derived from the source-domain local graph, target-domain local graph, and cross-domain global heterogeneous graph, and performs Top-K product list decision-making by combining candidate set constraints with Slate Q-based list value approximation. The reward function integrates target-domain feedback with a cross-domain alignment similarity shaping term to strengthen target-domain preference modeling. Meanwhile, the transfer cost is defined as the cosine distance between the recommended list representation and the stable interest representation in the source domain. By jointly modeling return value and cost value, the method learns a revised value function that guides policy optimization under a predefined cost threshold. To improve the stability of offline training, conservative value learning and behavior-consistency regularization are further introduced. Experimental results demonstrate that DHGCRL outperforms a range of sequential recommendation and offline reinforcement learning baselines in terms of HR and NDCG.

Keywords

Cross-Domain Sequential Recommendation, Dynamic Heterogeneous Graph, Offline Reinforcement Learning, Migration Cost Constraint

1. 引言

在电商平台多业务生态不断融合的背景下，用户的浏览、点击、加购、收藏与购买等行为通常分布于不同业务域，并呈现出明显的序列性与阶段性。跨域序列推荐通过利用源域信息缓解目标域数据稀疏与冷启动问题[1]，已成为提升电商推荐覆盖能力与用户体验的重要手段。然而，跨域迁移并非越强越好，源域与目标域在物品语义、交互分布与反馈机制上的差异容易导致负迁移，使推荐列表偏离用户在目标域的真实需求[2]。同时，电商推荐的优化目标具有长期性，用户满意度、复购与留存等信号往往存在延迟反馈，仅依赖单步监督学习的排序优化，难以刻画多步交互下长期收益的累积过程及迁移强度之间的权衡关系[3]。

在现有研究中，第一类工作聚焦于推荐的表征与预测能力提升，例如基于自注意力的序列建模方法通过捕获长程依赖提升 Top-K 排序性能，典型代表包括 SASRec [4]与 BERT4Rec [5]等模型。第二类工作进一步引入图表示学习来刻画用户与物品的高阶关联，并通过时间建模描述兴趣演化，涵盖动态图与异构图建模思路，为跨域迁移提供更丰富的结构与语义线索。第三类工作围绕跨域迁移与表示对齐展开，常用策略包括共享表示空间、跨域对比学习与多视图一致性约束，以提升源域知识向目标域迁移的有效

性与鲁棒性。上述方法在提升短期指标与迁移效果方面取得进展，但在长期收益优化与负迁移可控性方面仍存在不足。

为更贴近电商推荐系统的交互决策本质，强化学习推荐逐渐成为重要方向，其将推荐过程抽象为序贯决策并以长期回报为优化目标。考虑到实际系统输出为 Top-K 列表而非单一物品，列表级决策与价值估计成为关键问题[6]，SlateQ 等方法通过对列表价值进行可计算分解来缓解组合爆炸。同时，在线探索在工业场景中成本高且风险大，离线强化学习因此被广泛采用，但其面临分布外动作导致的价值过估计与策略偏移等挑战，保守价值学习与行为约束等稳定化技术成为重要研究分支[7]。此外，在跨域场景中，迁移强度的约束与安全性更加重要，引入受限决策过程与拉格朗日优化以实现约束策略学习也逐渐受到关注[8]。

基于上述研究脉络，本文面向电商领域跨域序列商品推荐任务提出离线约束强化学习模型 DHGCRL，将跨域推荐形式化为带迁移代价约束的受限马尔可夫决策过程[9]，并以 Top-K 推荐列表作为动作输出。在状态建模上，本文利用源域动态图、目标域动态图与跨域全局异构图生成的三视图对齐表征构造强化学习状态，在决策与估计上，通过候选集合约束缩小动作空间，并采用列表价值近似实现列表动作价值评估[10]，在优化目标上，以目标域反馈为主奖励并引入对齐奖励塑形增强迁移贡献，同时将推荐列表与源域稳定兴趣的偏离刻画为迁移代价，并通过拉格朗日机制联合学习回报价值与代价价值[11]，使策略在提升长期收益的同时满足代价阈值约束，在离线训练上，引入保守价值学习与行为一致性正则以提升训练稳定性与部署可靠性[12]。通过上述设计，DHGCRL 在兼顾长期收益的同时实现了迁移强度的显式可控，为电商场景下的跨域序列商品推荐提供了一种可行的策略优化框架。

2. 方法

2.1. 数据来源于预处理

本文实验统一采用 Amazon Reviews 2023 评论数据集作为数据来源，利用其覆盖多品类子域且记录完整时间戳的特点构建跨域序列推荐评测环境。具体选取 Video Games、Movies and TV 以及 Toys and Games 三个领域数据用于跨域任务设置，并在相同口径下复用同一套数据与预处理流程，以保证不同实验结果具备可比性与可复现性。

在数据预处理方面，首先对原始评论记录进行字段抽取，仅保留用户标识、物品标识、评分信息与交互时间戳等与推荐建模直接相关的内容，去除文本描述等冗余信息，随后依据时间戳对每位用户的交互行为进行排序，构建时间一致的行为序列。在反馈定义与过滤规则上，本文将评分行为视为隐式正反馈信号，用以刻画用户对物品的兴趣强度。为降低极端稀疏用户与物品对训练稳定性的影响，并缓解冷启动带来的噪声干扰，本文进一步删除在对应数据集中交互次数少于 5 次的用户与物品记录，从而提升序列有效性与连续性。在跨域样本构建时，为确保迁移关系可建模，仅保留在多个领域中均存在交互记录的重叠用户，并分别抽取其在不同领域内的交互序列用于后续跨域建模。在评估阶段，为避免对全量物品进行排序带来的计算开销，本文采用负采样构建候选集合，对每个用户在其真实交互物品之外随机采样 100 个未交互物品作为负样本，并与真实物品共同组成候选集合进行 Top-K 排序评测。

2.2. 模型介绍与参数设置

为在序列交互过程中同时刻画目标域长期收益提升与跨域迁移过程的可控性，本文将跨域序列推荐过程形式化为带约束的马尔可夫决策过程(Constrained Markov Decision Process, CMDP)。该建模方式将推荐系统的多步交互视为连续决策问题，使模型能够在考虑未来反馈的条件下学习决策策略，并通过显式约束抑制潜在的负迁移与不稳定行为，从而更贴合实际推荐场景中对长期效用与系统稳定性的综合要求。

首先，将用户－系统交互过程表示为马尔可夫决策过程(MDP)：

$$\mathcal{M}=(\mathcal{S},\mathcal{A},\mathcal{P},\mathcal{R},\gamma) \tag{1}$$

其中，$\mathcal{S}$ 表示状态空间，$\mathcal{A}$ 表示动作空间，$\mathcal{P}(s'\mid s,a)$ 为状态转移概率，$\mathcal{R}(s,a)$ 为即时奖励函数，$\gamma\in(0,1)$ 为折扣因子。考虑到推荐系统的实际输出通常为 Top-K 列表，本文将动作定义为推荐列表：

$$a_t\ @A_t=\left[i_{t,1},i_{t,2},\mathrm{L}\ ,i_{t,K}\right] \tag{2}$$

相应地，策略定义为在状态 s_t 条件下输出列表动作的条件概率分布：

$$\pi(A\mid s)=\Pr(A_t=A\mid s_t=s) \tag{3}$$

在策略 π 下，交互轨迹定义为：

$$\tau=(s_0,a_0,s_1,a_1,\mathrm{L}\ ) \tag{4}$$

其概率分布满足：

$$p_\pi(\tau)=\rho_0(s_0)\prod_{t=0}^{\infty}\pi(A_t\mid s_t),\mathcal{P}(s_{t+1}\mid s_t,A_t) \tag{5}$$

其中，$\rho_0(\cdot)$ 为初始状态分布。$P(s_{t+1}\mid s_t,A_t)$ 描述用户对列表动作响应后所诱导的状态转移。由于推荐系统的真实转移机制通常不可显式获得，离线训练阶段将以日志数据中观测到的样本对该转移过程进行经验近似。

跨域序列推荐的状态需要同时反映用户在不同域的交互历史、兴趣随时间的演化以及跨域偏好的一致性。为此，DHGCRL 以动态异构图编码结果及跨域对齐表示作为状态表征的主体，将复杂的历史交互压缩为可用于决策的低维向量。记时刻 t 的状态为：

$$s_t=\left[G_u^G(t);G_u^S(t);G_u^T(t);\phi(\Delta t);\mathbf{x}_t\right] \tag{6}$$

其中 $G_u^S(t)$、$G_u^T(t)$ 分别表示用户在源域与目标域动态图异构结构下的表征，$G_u^G(t)$ 表示跨域全局视角下的对齐表征，用于刻画跨域共享偏好，$\phi(\Delta t)$ 为时间间隔编码，用于描述交互间隔带来的兴趣衰减或漂移，$\mathbf{x}_t$ 为可选则的上下文特征，例如场景、设备、会话特征等。上述状态构建保证策略在决策时能够同时利用结构信息、时序信息与跨域关联信息。

在列表动作设定下，动作空间规模随候选集合与 K 呈组合增长。若直接对 $\mathcal{A}=\{A:|A|=K\}$ 进行价值学习将面临显著的计算与估计困难。为兼顾可训练性与可部署性，本模型采用 SlateQ 的列表价值建模思想：在保持列表作为原子动作的前提下，用可计算的方式近似刻画列表的长期价值，并将其作为后续策略优化的价值评估基础。具体地，给定状态 s_t 与列表动作 A_t 定义列表动作价值函数为：

$$Q^\pi(s_t,A_t)=\mathbb{E}_{\tau\sim p_\pi}\left[\sum_{l=0}^{\infty}\gamma^l r_{t+l}\mid s_t,A_t\right] \tag{7}$$

并在 SlateQ 假设下将 $Q^\pi(s_t,A_t)$ 近似分解为由列表中物品价值项构成的可计算形式：

$$Q^\pi(s_t,A_t)\approx\sum_{k=1}^{K}w_k Q^\pi(s_t,i_{t,k}),\quad w_k\geq 0 \tag{8}$$

其中 w_k 用于刻画位置偏置或曝光权重，$Q^\pi(s_t,i)$ 为物品级价值项。该近似的核心目的在于将列表级长期价值估计转化为物品级价值估计的组合，从而显著降低价值函数学习的复杂度，并为后续在候选集合上进行可扩展的列表构造与策略更新提供基础。相应地，策略在候选集合诱导的可行动作集合 $\mathcal{A}(s_t)\subseteq\mathcal{A}$

上满足归一化约束：

$$\sum_{A\in\mathcal{A}(s_t)} \pi_\theta(A\mid s_t)=1,\quad A\in\mathcal{A}(s_t) \tag{9}$$

CMDP 建模默认交互过程满足马尔可夫性质：

$$\Pr(s_{t+1}\mid s_0,A_0,\mathrm{L}\ ,s_t,A_t)=\Pr(s_{t+1}\mid s_t,A_t) \tag{10}$$

即当前状态应尽可能包含对未来演化具有充分预测力的信息。同时，本模型关注目标域的长期效用提升，首先定义目标域即时反馈奖励 $r_t^{(T)}$，用于刻画点击、停留、转化等行为信号，并为增强稀疏反馈条件下的学习信号，DHGCRL 引入对齐塑形项 $r_t^{(\mathrm{align})}$，构成综合奖励：

$$\text{[illegible]}=r_t^{(T)}+\lambda r_t^{(\mathrm{align})},\qquad \lambda\ge 0 \tag{11}$$

塑形项用于将跨域一致偏好信息注入策略学习过程，由用户全局对齐表征与列表动作对应物品表征之间的相似度度量构造，本文采用余弦相似度定义：

$$r_t^{(\mathrm{align})}=\cos\left(G_u^g(t),G_{A_t}\right)=\frac{\left(G_u^g(t)\right)^{\top}G_{A_t}}{\left\|G_u^g(t)\right\|\left\|G_{A_t}\right\|} \tag{12}$$

其中，G_{A_t} 表示列表动作对应的嵌入，由列表内物品表示按位置权重聚合得到，从而与 SlateQ 的列表建模保持一致，该设计使奖励塑形在列表级输出，避免仅对单一物品塑形而与实际动作定义不一致。在约束建模方面，策略若仅以回报最大化为目标，可能倾向于采取过强的迁移或探索行为，带来源域体验下降、跨域干预过度或系统稳定性风险。为此引入代价函数：

$$c_t=c(s_t,A_t) \tag{13}$$

并设置代价阈值 η，将迁移强度与潜在副作用以显式约束纳入优化。定义折扣累计回报与折扣累计代价分别为：

$$G_r(\tau)=\sum_{t=0}^{\infty}\gamma^t\text{[illegible]}\qquad G_c(\tau)=\sum_{t=0}^{\infty}\gamma^t c_t \tag{14}$$

则策略 π 的期望回报与期望代价为：

$$J(\pi)=\mathbb{E}_{\tau\sim p_\pi(\tau)}\left[G_r(\tau)\right]=\mathbb{E}_{\tau\sim p_\pi(\tau)}\left[\sum_{t=0}^{\infty}\gamma^t\text{[illegible]}\right] \tag{15}$$

$$J_c(\pi)=\mathbb{E}_{\tau\sim p_\pi(\tau)}\left[G_c(\tau)\right]=\mathbb{E}_{\tau\sim p_\pi(\tau)}\left[\sum_{t=0}^{\infty}\gamma^t c_t\right] \tag{16}$$

因此，DHGCRL 的 CMDP 目标可表示为：

$$\max_{\pi} J(\pi)\quad \text{s.t.}\quad J_c(\pi)\le\eta. \tag{17}$$

为便于求解，本文采用拉格朗日松弛将约束优化转化为对偶形式。定义拉格朗日函数：

$$\mathcal{L}(\pi,\beta)=J(\pi)-\beta\left(J_c(\pi)-\eta\right),\qquad \beta\ge 0 \tag{18}$$

对应的对偶优化为：

$$\max_{\pi}\min_{\beta\ge 0}\mathcal{L}(\pi,\beta) \tag{19}$$

其中 β 为代价约束的惩罚系数，用于在回报最大化与代价控制之间进行自适应平衡。为保证方法实现过

程清晰可复现，本文将 DHGCRL 训练与优化过程中涉及的关键超参数进行统一整理，并以表 1 给出默认配置。该表主要覆盖三类参数：一是图表示与状态构造相关的结构参数，用于确定用户与物品表征的容量与信息聚合深度，二是强化学习优化相关参数，用于控制价值估计与策略更新的时间视野、收敛速度与训练规模，三是约束与稳健化项参数，用于刻画迁移代价控制强度、奖励塑形贡献以及离线训练的稳定化程度，从而形成收益提升与风险控制之间的可调平衡。

Table 1. Relevant parameter settings
表 1. 相关参数设置

类别	参数	符号	参数默认值
表征学习	嵌入维度	d	128
表征学习	GNN 层数	L	2
强化学习优化	折扣因子	γ	0.97
强化学习优化	学习率	lr	0.001
强化学习优化	Batch Size	BS	2048
强化学习优化	训练轮数	epochs	500
约束设置	代价阈值	η	1.6
约束优化	惩罚系数	β	1.1
约束优化	更新步长	α_{β}	0.01
训练稳定性	保守价值正则权重	α	0.1
训练稳定性	行为一致性正则权重	λ_{BC}	0.2
奖励塑形	塑形项权重	λ	0.1

表 1 中，嵌入维度设为 128，图神经网络层数为 2，用于生成状态向量的核心表示，折扣因子取 0.97，学习率取 0.001，批大小为 2048，训练轮数为 500，以兼顾长期收益建模与训练稳定性。迁移代价约束方面，代价阈值设为 1.6，惩罚系数初值为 1.1，更新步长为 0.01，奖励塑形权重设为 0.1，用于调节对齐塑形信号在总奖励中的占比。离线稳健化方面，保守价值约束权重取 0.1，行为一致性约束权重取 0.2，以抑制分布外动作带来的价值过估计并限制策略偏移。后续实验若无特殊说明，均采用表 1 所列的默认参数设置。

3. 模型结果与分析

在统一的数据划分与候选集合设定下，本文将 DHGCRL 与 SASRec、BERT4Rec、SlateQ、SQN、DRR、CQL、IQL 等方法进行对比，并统一采用 HR@10 与 NDCG@10 作为评价指标。实验结果如表 2 所示，其中 MV 代表跨域迁移方向为 Movies → Video，DHGCRL 在三组跨域迁移任务上均取得最优表，其中 Movies → Video 的 NDCG@10 与 HR@10 分别为 0.19524、0.35583，Video → Toys 为 0.14231、0.23144，Toys → Movies 为 0.17728、0.32493。该结果表明，在跨域序列推荐场景中，引入列表级决策以及约束式长期优化的策略学习框架能够稳定提升目标域 Top-K 排序质量与命中效果。

进一步对比离线强化学习基线可观察到，DHGCRL 相比 IQL 与 CQL 在三项任务的两项指标上均实现提升。以 IQL 为例，DHGCRL 在 Movies → Video、Video → Toys、Toys → Movies 三个任务上的 NDCG@10 与 HR@10 均有明确增益，说明在同样使用离线日志进行策略学习的条件下，DHGCRL 能学

习到更符合跨域序列推荐目标的列表策略。这类提升与模型的三点关键设计相吻合，将动作显式建模为 Top-K 列表并采用 SlateQ 列表价值近似，使价值学习与实际列表生成过程保持一致，同时通过迁移代价约束抑制不适当迁移带来的性能波动风险

为量化各模块对性能提升的贡献，本文在完整模型基础上构建了四个消融变体，包括去除迁移代价约束的 DHGCRL-NoCost、移除 SlateQ 列表价值近似的 DHGCRL-NoSlateQ、固定惩罚系数的 DHGCRL-FixBeta，以及去除奖励塑形项的 DHGCRL-NoShape。如表 3 所示，完整模型在三组任务的 HR@10 与 NDCG@10 均为最优，而四类删减模型均导致不同程度下降。

Table 2. Multi-model comparison results
表 2. 多模型对比实验结果

模型	NDCG@10 (MV)	HR@10 (MV)	NDCG@10 (VT)	HR@10 (VT)	NDCG@10 (TM)	HR@10 (TM)
SASRec	0.17873	0.32091	0.11984	0.19267	0.14538	0.27946
BERT4Rec	0.17162	0.30341	0.12427	0.19836	0.14819	0.28357
SlateQ	0.16841	0.29526	0.12178	0.19544	0.14732	0.28118
SQN	0.16493	0.28817	0.11829	0.19063	0.14246	0.27482
DRR	0.16047	0.28135	0.11516	0.18728	0.13921	0.27049
CQL	0.18663	0.33874	0.13389	0.21943	0.16824	0.30938
IQL	0.18817	0.34192	0.13618	0.22126	0.17016	0.31247
DHGCRL	0.19524	0.35583	0.14231	0.23144	0.17728	0.32493

Table 3. Ablation study results
表 3. 消融实验结果

模型	NDCG@10 (MV)	HR@10 (MV)	NDCG@10 (VT)	HR@10 (VT)	NDCG@10 (TM)	HR@10 (TM)
NoCost	0.18912	0.34526	0.13574	0.21936	0.17102	0.31344
NoSlateQ	0.18795	0.34680	0.13692	0.22518	0.17038	0.31615
FixBeta	0.19273	0.35164	0.14008	0.22811	0.17511	0.32140
NoShape	0.19108	0.35012	0.13854	0.22602	0.17384	0.31928
DHGCRL	0.19524	0.35583	0.14231	0.23144	0.17728	0.32493

如表 3 所示，其中，NoCost 在三个任务上均出现退化，说明仅以目标域回报为优化目标难以充分控制迁移强度，域间差异较大时更易引发负迁移，从而削弱最终推荐效果。FixBeta 虽优于 NoCost，但仍低于完整模型，表明固定惩罚系数只能提供基础约束，而自适应调节惩罚强度能够更有效地在长期回报提升与迁移代价控制之间取得平衡。NoSlateQ 的下降说明若不显式刻画列表动作的价值结构，策略学习难以与 Top-K 列表生成过程对齐，从而限制收益。NoShape 的下降则表明对齐塑形信号能够在离线日志中提供更充分的优化引导，帮助策略更有效利用跨域偏好一致性信息。

为考察对齐奖励塑形在离线策略学习中的贡献强度，本文在其余超参数与实验配置保持不变的条件下，对塑形项权重 λ 进行敏感性分析，并在三组跨域迁移任务上比较不同 λ 取值下的推荐效果。实验中将 λ 设为 0、0.05、0.1、0.2、0.3，其中 λ 取 0 表示不引入塑形信号，与去除塑形项的模型变体一致。

如图 1 所示，λ 的变化会直接影响最终排序表现。当 λ 从 0 逐步增大至约 0.15 附近时，HR@10 与 NDCG@10 整体呈上升趋势，说明适度的塑形信号能够在目标域反馈稀疏的离线条件下补充有效学习信

号，引导策略更稳定地利用跨域一致性信息，从而提升 Top-K 推荐质量。但当 λ 继续增大时，增益逐渐变弱并在个别任务上出现轻微回落，表明过强的塑形可能使优化更偏向一致性约束，压缩策略对目标域即时反馈的适配空间，进而影响区分能力。

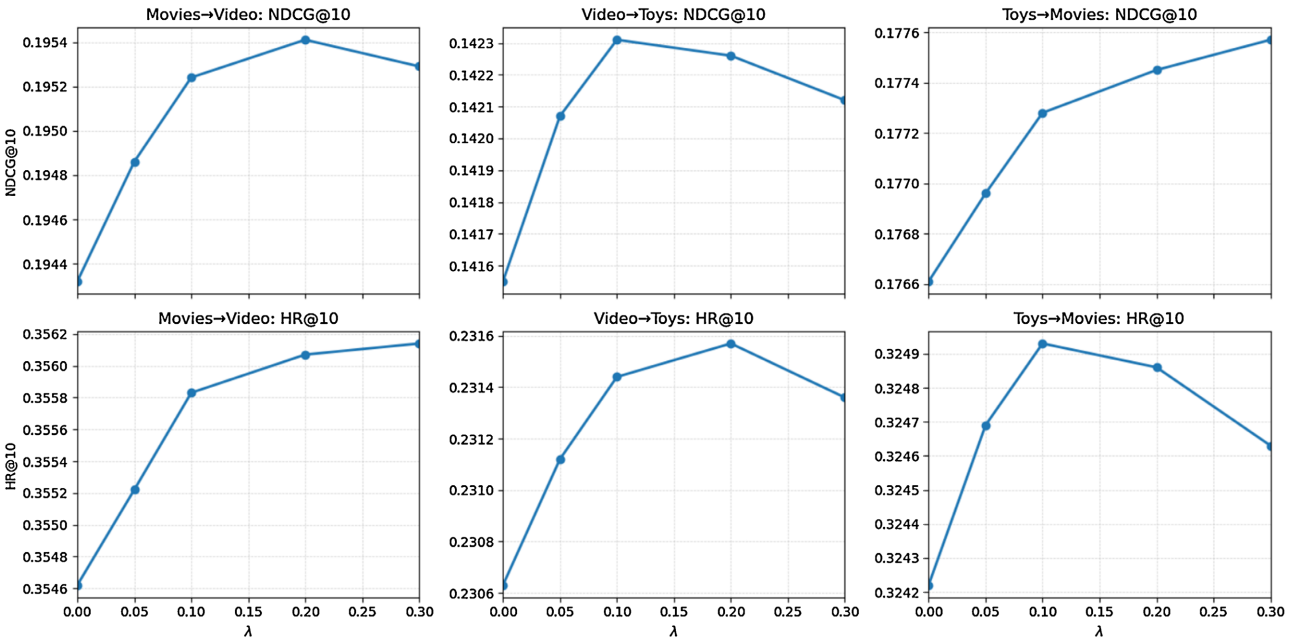

Figure 1. Experiment on hyperparameter of reward shaping term weight
图 1. 塑形项权重参数实验

为分析长期收益权衡对策略学习的影响，本文在其余设置保持一致的条件下，对折扣因子 γ 进行敏感性实验。γ 用于控制模型对未来回报的重视程度，取值越大表示越强调长期累计收益。实验中在代表性跨域任务上对 γ 取 0.92、0.94、0.96、0.97、0.98、0.99 进行对比评测，并观察 HR@10 与 NDCG@10 的变化趋势。

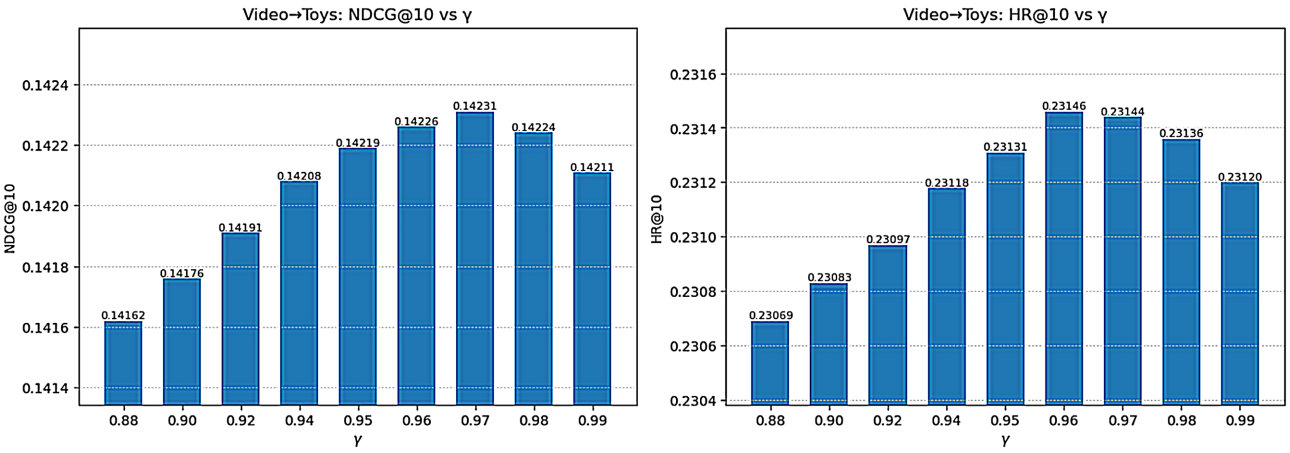

Figure 2. Experiment on hyperparameter of discount factor
图 2. 折扣因子参数实验

如图 2 所示，随着 γ 从较小值逐步增大，两项指标整体呈现先提升后趋稳的趋势，说明适度强调未

来回报能够更好地捕获跨步交互带来的收益累积，从而提升列表策略的整体质量。当γ接近 1 时，部分指标出现轻微回落，表明过度强调长期回报会引入更强的估计方差与误差累积，使离线价值学习更难稳定收敛，进而影响最终性能。值得注意的是，NDCG@10 在γ约为 0.97 附近达到相对最优，而 HR@10 在γ约为 0.96 时略占优势，反映两类指标对长期性权重的敏感程度存在差异。

为检验离线训练中两类稳定化机制的作用强度，本文以 Video → Toys 迁移任务为例，对保守价值约束权重α与行为一致性约束权重λ_{BC}进行参数敏感性分析。在其余训练配置保持不变的条件下，先选取α的三个代表性取值 0.05、0.1、0.2，再对λ_{BC}在 0、0.05、0.1、0.2、0.3 范围内进行扫描，并分别记录 NDCG@10 与 HR@10 的变化，用于衡量列表排序质量与命中效果。

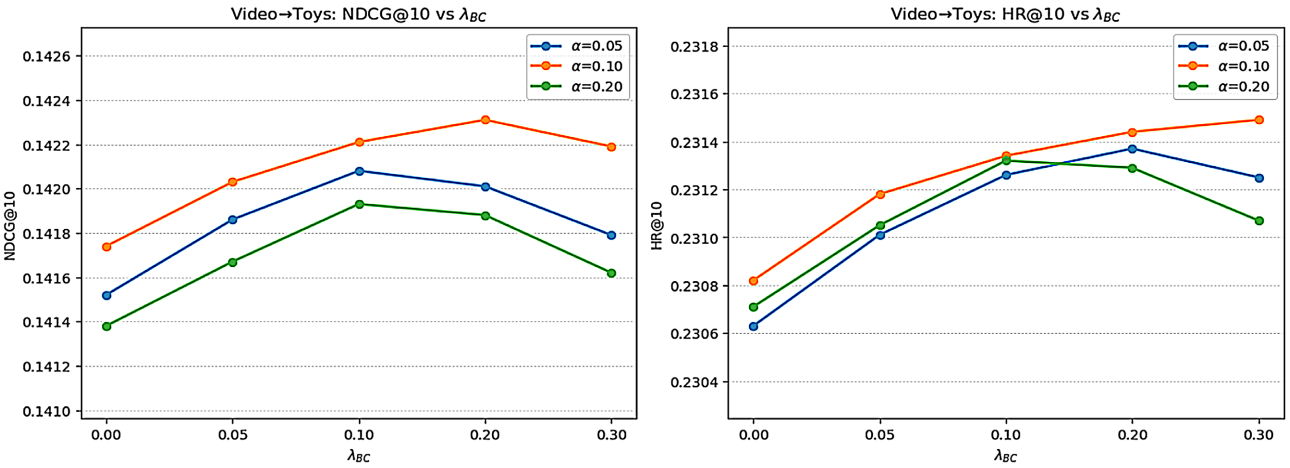

Figure 3. Experiment on hyperparameter of stabilization mechanism
图 3. 稳定机制相关参数实验

从图 3 可以观察到，当λ_{BC}取 0 时，策略更新缺少对日志行为分布的牵引，更容易向离线数据覆盖不足的动作区域偏移，因而两项指标整体偏低。随着λ_{BC}增大到适中区间，模型性能明显改善，说明适度的行为一致性约束有助于抑制策略漂移并提升训练可靠性。但当λ_{BC}继续增大时，增益逐渐减弱，甚至出现轻微回落，反映过强的行为收缩会使策略过度贴合历史行为，从而压缩进一步优化长期回报的空间。另一方面，α控制价值学习的保守化程度，α较小时对分布外高估的抑制较弱但更易获得较高指标。当α增大时，价值预测被更强压制，模型趋于保守，整体 NDCG 水平下降。

4. 总结

本文面向电商领域跨域商品推荐中目标域交互稀疏、反馈具有延迟性以及跨域迁移强度难以控制等问题，提出一种以长期收益优化为核心的离线约束强化学习框架 DHGCRL。不同于仅依赖单步监督信号的排序学习，本文将跨域商品推荐过程形式化为受限马尔可夫决策过程，以 Top-K 推荐列表作为动作输出，将长期累计回报最大化与迁移代价受限统一到同一优化目标中，从建模层面明确刻画跨域迁移带来的收益与风险权衡，为后续策略学习提供了可解释、可控的理论基础。

在方法设计上，DHGCRL 以动态异构图建模用户跨域行为结构，并结合源域动态图、目标域动态图与跨域全局异构图生成的三视图对齐表征构造强化学习状态，保证策略输入能够同时反映源域稳定兴趣、目标域即时偏好与跨域共享信息。在决策层面，为解决列表动作空间过大的问题，模型首先构建动态候选集合以约束搜索空间，再采用 SlateQ 风格的列表价值近似实现列表价值的可计算评估，从而支持可部署的列表生成机制。在优化层面，模型以目标域反馈为主奖励，并引入对齐相似度作为奖励塑形信号以

增强迁移表征对目标域决策的引导。同时以推荐列表表征与源域稳定兴趣表征的偏离定义迁移代价，通过拉格朗日乘子联合学习回报价值与代价价值，形成修正价值来驱动策略更新，使策略在提升长期收益的同时满足代价阈值约束。针对离线强化学习的分布外估计与策略偏移问题，进一步引入保守价值学习与行为一致性正则，以提升价值学习稳定性与策略可靠性。

在实验验证方面，本文基于统一的数据来源与预处理流程，在三个跨域迁移任务上采用 HR@10 与 NDCG@10 进行评估，并将 DHGCRL 与典型序列推荐模型及离线强化学习方法进行对比。结果表明，DHGCRL 在各任务上均取得最优或显著更优的 Top-K 推荐性能，说明所提出的列表级离线约束策略学习能够有效提升目标域推荐质量。消融实验进一步验证了迁移代价约束、SlateQ 列表价值近似与奖励塑形信号对性能提升的关键作用，自适应惩罚系数相较固定惩罚能够带来更稳定的收益与风险平衡。参数敏感性分析显示，塑形权重、折扣因子以及保守约束与行为约束存在合理有效区间，适度设置可在长期收益提升、迁移风险控制与离线训练稳定性之间取得更优折中，体现了方法的鲁棒性。

致　谢

本文的完成离不开许多人的支持与帮助。在论文研究与写作过程中，感谢所有在课题思路讨论、方法实现、实验设计与结果分析等方面给予指导与建议的老师、同学与朋友，感谢在资料查阅、数据整理、论文排版与修改完善过程中提供协助的相关人员，同时感谢学院与实验平台在学习与科研条件方面给予的支持。

参考文献

[1] Chen, S., Xu, Z., Pan, W., *et al*. (2024) A Survey on Cross-Domain Sequential Recommendation. *The* 33*rd International Joint Conference on Artificial Intelligence Survey Track*, Jeju, 3-9 August 2024, 7989-7998.

[2] Lin, G., Gao, C., Zheng, Y., Chang, J., Niu, Y., Song, Y., *et al*. (2024) Mixed Attention Network for Cross-Domain Sequential Recommendation. *Proceedings of the* 17*th ACM International Conference on Web Search and Data Mining*, Merida, 4-8 March 2024, 405-413. https://doi.org/10.1145/3616855.3635801

[3] Ni, R., Cai, W. and Jiang, Y. (2024) Contrastive Cross-Domain Sequential Recommendation via Emphasized Intention Features. *Neural Networks*, **179**, Article ID: 106488. https://doi.org/10.1016/j.neunet.2024.106488

[4] Kang, W. and McAuley, J. (2018) Self-Attentive Sequential Recommendation. 2018 *IEEE International Conference on Data Mining* (*ICDM*), Singapore, 17-20 November 2018, 197-206. https://doi.org/10.1109/icdm.2018.00035

[5] Sun, F., Liu, J., Wu, J., Pei, C., Lin, X., Ou, W., *et al*. (2019) BERT4Rec: Sequential Recommendation with Bidirectional Encoder Representations from Transformer. *Proceedings of the* 28*th ACM International Conference on Information and Knowledge Management*, Beijing, 3-7 November 2019, 1441-1450. https://doi.org/10.1145/3357384.3357895

[6] Zhou, D., Cai, X. and Pan, W. (2025) Contrastive Text-Enhanced Transformer for Cross-Domain Sequential Recommendation. *Proceedings of the* 31*st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Vol. 2, 4110-4119. https://doi.org/10.1145/3711896.3736893

[7] Li, W., Lin, X., Pan, W. and Ming, Z. (2024) Dynamic Stage-Aware User Interest Learning for Heterogeneous Sequential Recommendation. 18*th ACM Conference on Recommender Systems*, Bari, 14-18 October 2024, 465-474. https://doi.org/10.1145/3640457.3688103

[8] Liu, S., Cai, Q., He, Z., Sun, B., McAuley, J., Zheng, D., *et al*. (2023) Generative Flow Network for Listwise Recommendation. *Proceedings of the* 29*th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Long Beach, 6-10 August 2023, 1524-1534. https://doi.org/10.1145/3580305.3599364

[9] Chen, X., Wang, S., McAuley, J., Jannach, D. and Yao, L. (2024) On the Opportunities and Challenges of Offline Reinforcement Learning for Recommender Systems. *ACM Transactions on Information Systems*, **42**, 1-26. https://doi.org/10.1145/3661996

[10] Wang, K., Zou, Z., Zhao, M., Deng, Q., Shang, Y., Liang, Y., *et al*. (2023) RL4RS: A Real-World Dataset for Reinforcement Learning Based Recommender System. *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Taipei, 23-27 July 2023, 2935-2944. https://doi.org/10.1145/3539618.3591899

[11] Wachi, A., Shen, X. and Sui, Y. (2024) A Survey of Constraint Formulations in Safe Reinforcement Learning.

Proceedings of the 33*rd International Joint Conference on Artificial Intelligence*, Jeju, 3-9 August 2024, 8262-8271.

[12] Chen, X., Yao, L., McAuley, J., Zhou, G. and Wang, X. (2023) Deep Reinforcement Learning in Recommender Systems: A Survey and New Perspectives. *Knowledge-Based Systems*, **264**, Article ID: 110335. https://doi.org/10.1016/j.knosys.2023.110335