主动引导　　交互式推荐中物品侧公平性的策略　　n

Chongjun Xia[1,2] , **Xiaoyu Shi**[1] , **Hong Xie**[3]
Xianzhi Wang[2] , **Yun Lu**[1] 和 **Mingsheng Shang**[11]
中国科学院[2]悉尼科技大学[3]中国科学技术大学
a260813501@gmail.com

arXiv:2603.03094v1 [cs.IR] 3 Mar 2026

摘要

物品侧公平性对于确保长尾物品在交互式推荐系统中获得公平曝光至关重要。现有方法通过将长尾物品直接纳入推荐结果来提升其曝光度。这会导致用户偏好与推荐的长尾物品之间出现错位，从而阻碍用户长期参与度，并降低推荐效果。我们的目标是采用主动公平引导策略，在交互式推荐过程中主动引导用户偏好转向长尾物品，同时维持用户满意度。为此，我们提出 HRL4PFG，这是一种交互式推荐框架，利用分层强化学习逐步引导用户偏好转向长尾物品。HRL4PFG 通过宏观层过程，根据多步反馈生成公平性引导目标；并通过微观层过程，依据这些目标和不断变化的用户偏好实时微调推荐结果。大量实验表明，与交互式推荐环境中的最先进方法相比，HRL4PFG 在累积交互奖励和最大用户交互长度方面取得了更大幅度的提升。

1 引言

交互式推荐系统是现代在线平台的重要工具，可实时提供个性化内容、推荐相关物品并提升用户参与度 [He 等人， 2016]。这些系统会跟踪并建模不断变化的用户偏好，从电子商务到直播等不同领域，提升用户体验 [Quan 等人， 2023]。尽管推荐系统效果显著，但仍面临流行度偏差带来的公平性挑战 [Abdollahpouri等人， 2019]，：少数热门物品吸引了大部分关注，而许多不太热门（即长尾）物品则曝光不足，导致其对用户的可见性降低。

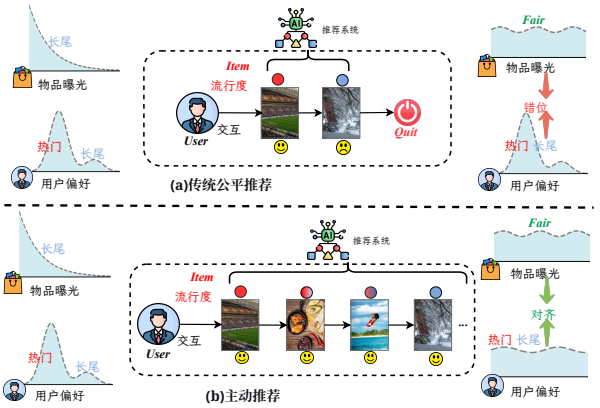

图 1：两种物品侧公平性感知推荐的示意图。

在本研究中，我们关注物品侧公平性，并将其定义为在不同物品之间更公平地分配曝光机会。这一公平性概念在内容平台中尤为重要，因为创作者、卖家或信息源都是利益相关者，其效用与物品曝光直接相关 [Wang 等， 2023]。在这一视角下，先前的研究主要致力于缓解长尾物品在曝光方面的不利处境 [Beutel等， 2019]。这些研究利用因果推断技术来减轻流行度偏差 [Wei 等， 2021], 在模型训练过程中加入公平性感知正则化项 [Abdollahpouri等，2017], 或采用重排序策略来调整推荐结果 [Zhu等， 2021]。对于动态环境，现有研究通常设计有效的奖励函数，或在强化学习（RL）中加入公平性约束，以提升长尾物品的曝光度 [Shi 等，2023；Shi 等， 2024]。这些方法都通过将长尾物品直接融入推荐结果来增加其曝光。这种突兀的公平性强制策略会破坏推荐物品与用户偏好之间的一致性，导致用户参与度和满意度下降。图1(a)展示了传统的公平性感知推荐场景：用户与热门物品交互后，推荐模型会立即优先提升长尾物品的曝光。尽管这种方法改善了长尾物品曝光的分布，但这种突兀的

推荐转变可能会令用户错愕，打消他们参与的意愿，并最终降低用户满意度。

近期研究提出了主动推荐的概念，指出推荐系统应随着时间推移，逐步引导用户偏好的演变 [Tang 等，2019；Wu等，2019；Zhu等， 2023]。受这一观察启发，我们提出一种主动公平引导策略，旨在提升交互式推荐中的物品侧公平性，同时保持用户满意度。与现有方法不同，后者不考虑用户偏好与推荐物品之间的匹配，强行增加长尾物品的曝光；我们的策略则在整个交互过程中迭代生成个性化且可执行的推荐。通过这种方式，我们的方法逐步调整用户偏好，使其适应长尾物品的特点，从而在不损害用户参与度的情况下实现物品侧公平性。图 1(b) 展示了采用我们主动公平引导策略的交互式推荐系统，该策略根据用户当前偏好逐步降低热门物品的曝光，引导用户随着时间推移欣赏长尾物品。这使用户偏好能够从长尾分布逐步转向更加均衡的分布，同时不影响长期满意度。

实施主动公平引导策略面临若干挑战：1）用户在与推荐系统动态交互的过程中，用户偏好会不断变化。静态或基于规则的方法无法适应这种变化，因而会带来负面用户体验。2）为实现物品侧公平性而随机选择长尾物品，可能导致推荐低质量物品，从而削弱系统的整体效果。3）如何在引导用户偏向长尾物品与确保用户满意度之间取得平衡，仍是交互式推荐研究中一个复杂且尚未解决的挑战。

本文提出了主动公平引导的分层强化学习（HRL4PFG），这是一种新颖的交互式推荐框架，利用强化学习应对上述挑战。HRL4PFG采用分层学习方法，将引导过程分为两个阶段：宏观学习和微观学习。在宏观学习阶段，高层智能体根据多步反馈制定公平性引导目标，同时兼顾物品侧公平性和用户满意度。这些目标将引导微观学习阶段的过程。在微观学习阶段，低层智能体将这些高层目标转化为适合个体用户的可执行动作。低层智能体在实现既定目标与维持用户满意度之间权衡，逐步引导用户偏好朝期望目标发展。通过这种渐进式、自适应的引导策略，HRL4PFG在提升推荐系统物品侧公平性的同时，最大限度地减少对用户满意度的影响。总之，本文的贡献如下：

- 我们强调，有必要采用主动公平引导策略，逐步影响用户对长尾物品的偏好，并利用该策略提升交互式推荐中的物品侧公平性。
- 我们提出HRL4PFG，一种分层强化学习框架，以提升交互式推荐系统中的物品侧公平性。HRL4PFG采用分层学习方法，将引导过程分为两项任务：高层智能体生成公平引导目标，低层智能体基于这些目标选择推荐物品。
- 我们开展了广泛的交互式实验，从最大交互长度和累积交互奖励两个方面验证了 HRL4PFG 的有效性和优越性。

2 相关工作

推荐中的公平性。
公平性在推荐研究中日益受到关注，旨在减少流行度偏差和不平等现象 [Geyik 等， 2019]。公平性通常从群体层面或个体层面进行考量。群体公平性确保不同人口统计群体之间受到公平对待，而个体公平性确保相似的用户和物品受到同等对待 [Abdollahpouri 等，2019；Singh 和 Joachims，2019]。

已有多种方法致力于解决推荐系统中的公平性问题 [Wei 等，2021；Beutel 等，2019；Zhu 等，2021]。例如，Abdollahpouri 等人 [Abdollahpouri 等，2017] 提出了一个正则化项，以减少流行度偏差；Zhu 等人 [Zhu等， 2018] 引入了一个结合潜在因子矩阵与公平性正则化的公平性增强框架。上述方法均关注静态单次推荐场景中的公平性问题，并在单个时间点优化公平性。因此，它们忽视了用户—系统交互的动态性，不适用于在交互式推荐过程中实现物品侧公平性。

主动推荐。
主动推荐偏离了传统推荐范式：从被动适应用户偏好，转向主动引导用户偏好朝期望方向发展 [Bi等， 2024]。这种主动方法有多项优势，例如缓解信息茧房问题并实现有效的广告投放。主动推荐研究主要聚焦于两个方向：1）在推荐系统中对偏好转变进行建模；2）主动引导用户偏好。

偏好转变模型研究用户兴趣如何随着与推荐系统的交互而演变 [Curmei 等，2022；Dean 和 Morgenstern，2022]。相比之下，偏好引导方法旨在让用户平滑地转向目标物品/主题，同时保持满意度 [Tang等，2019；Wu 等，2019；Zhu 等， 2023]。然而，大多数既有研究并未明确说明如何选择引导目标。我们通过生成公平性导向目标，并相应地应用平滑引导策略，来解决交互推荐中的物品侧公平性问题。

用于推荐的强化学习。
近年来，强化学习（RL）在应对交互式推荐场景的动态性方面日益受到关注 [Bharadhwaj，2019]。传统的

推荐算法通常优化静态的一次性目标，例如最大化点击率，却无法考虑用户不断变化的偏好和行为 [Zhang 等, 2019; Zhang 等 , 2025]。相比之下，强化学习将推荐过程建模为动态环境中的一系列决策，从而能够持续适应用户偏好的变化 [Zheng等, 2018]。近年来，分层强化学习（HRL）已成为一种强大的方法，应用于包括推荐系统在内的多个领域 [Zhao等, 2020; Zhao 等, 2023]。在这项工作中，我们采用分层强化学习来提升交互式推荐系统中的物品侧公平性。

一些研究通过修改奖励函数或添加约束，探索了利用强化学习解决交互式推荐系统中物品侧公平性问题的方法 [Shi 等, 2023; Shi 等, 2024]。然而，向偏好热门内容的用户直接推荐长尾物品，可能会严重破坏推荐物品与用户偏好之间的一致性，导致用户参与度和满意度下降。鉴于用户偏好可以受到推荐系统的逐步影响 [Zhu等, 2023],我们的方法旨在逐步引导用户接触长尾物品。这种渐进式方法旨在增加长尾物品的曝光，同时保持用户满意度。

3 方法

3.1 任务定义

令 $u \in \mathcal{U}$ 和 $i \in \mathcal{I}$ 分别表示一个用户和一个物品。在交互式推荐中，系统会在多个轨迹中与用户交互。在步骤 t，系统会观察一个用户状态 s_t^u，该状态由交互历史概括而成，并推荐一个物品 i_t^u，进而接收反馈 r_t^u。用户偏好会随着交互而变化，导致状态随时间呈现非平稳性。

物品侧公平性。我们关注物品侧公平性，其旨在系统层面更公平地在各物品之间分配曝光机会 [Wang 等, 2023]。具体而言，它强调减轻长尾物品在曝光方面的不利处境。令 $\mathcal{I}_{\text{tail}} \subset \mathcal{I}$ 为长尾集合。给定用户集合 $\mathcal{U}$ 上的一组交互轨迹，物品 i 的总曝光量定义为：

$$\text{Exp}(i) \;=\; \sum_{u \in \mathcal{U}} \sum_{t=1}^{T_u} \mathbb{I}[i_t^u = i]\,.$$

相应地，长尾曝光率计算如下：

$$\rho \;=\; \frac{\sum_{i \in \mathcal{I}_{\text{tail}}} \text{Exp}(i)}{\sum_{i \in \mathcal{I}} \text{Exp}(i)},$$

该指标衡量了在所有交互步骤中分配给长尾物品的总曝光量所占的比例。

主动公平性引导推荐。 为在不降低用户满意度的情况下改善ρ ，系统采用主动引导策略。具体而言，系统生成一个公平性引导目标 j （即一个虚拟长尾物品），并在 M 个交互推荐步骤中引导用户偏好转向该目标。目标是学习一种策略 π，在提升系统层面长尾曝光的同时，最大化长期用户满意度。

3.2 框架概述

基于上述设置，我们提出了 HRL4PFG，这是一种新颖的交互式推荐框架，利用分层强化学习提升物品侧公平性。HRL4PFG 将引导过程分为两个阶段（如图 2 所示）：

1）宏观学习：高层智能体（HRA）由一个高层行动者和一位评论家组成。行动者负责生成公平性引导目标，而评论家则根据多步用户反馈以及推荐结果的物品侧公平性对这些目标进行评估。基于评估结果，评论家迭代优化行动者的决策过程。

2）微观学习：低层智能体（LRA）由一个低层行动者和一位评论家组成。行动者同时考虑用户当前偏好和公平性引导目标来生成物品推荐，旨在逐步使用户偏好与公平目标保持一致。评论家评估这些推荐在引导用户偏好方面的效果，同时评估其与用户当前偏好的契合程度。基于这些评估，评论家优化行动者的决策过程。

在以下小节中，我们将进一步详细介绍 HRL4PFG 的各个组成部分。

3.3 宏观学习

宏观学习过程包括由高层智能体（HRA）生成公平性引导目标，以提升公平的物品曝光。该公平性引导目标将作为参考，在接下来的 M 步中引导低层智能体采取行动。

高层状态跟踪器。
高层状态包含有关最近 ^hN 次用户交互 $\{h_{t-N+1}, \ldots, h_t\}$的信息，反映了用户的偏好。为对这些交互之间的依赖关系进行建模，我们引入了注意力机制。具体而言，首先通过嵌入层将用户交互编码为向量表示 $\{e_{t-N+1}, \ldots, e_t\}$。随后，将所有历史嵌入向量映射到三个矩阵：查询矩阵 Q^h、键矩阵 K^h 和值矩阵 V^h。最后，高层状态表示可按如下方式计算：

$$s_t^h = Attention(Q^h, K^h, V^h) = softmax\left(\frac{Q^h (K^h)^T}{\sqrt{d^h}}\right) V^h, \tag{1}$$

其中 d^h 是缩放因子。

高层智能体模型。
高层行动者旨在生成动态的公平性引导目标，引导用户偏好朝向物品的更公平曝光。为促进全面探索并防止过早收敛，我们采用高斯分布。具体而言，高层状态s_t^h被输入多层感知机，以生成该分布的均值u_t^h和方差σ_t^2。

随后，从由均值 u_t^h 和

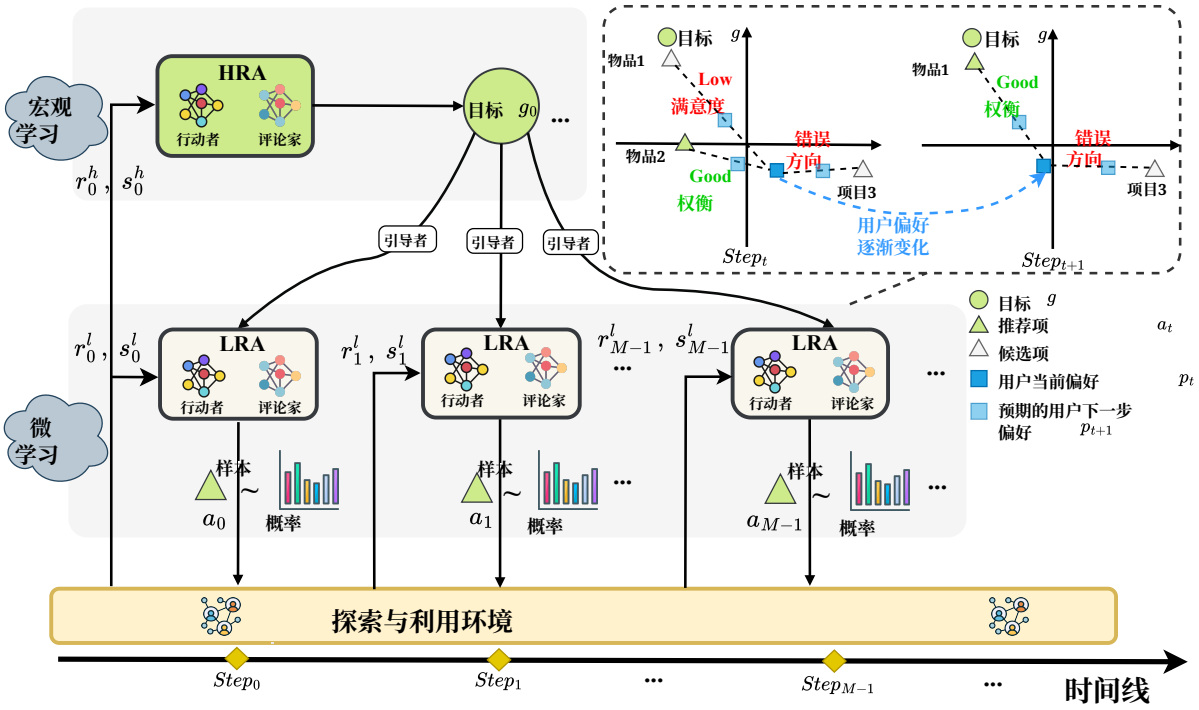

图 2：HRL4PFG 的整体架构。

方差 σ_t^2:

$$g_t \sim \mathcal{N}(u_t^h, \sigma_t^2). \tag{2}$$

高层智能体采用演员-评论家算法[Konda 和 Tsitsiklis, 1999]实现。演员-评论家算法的优势在于其稳定的学习过程、较低的方差，以及通过结合策略梯度和基于价值的方法实现高效的策略优化。评论家接收 s_t^h 并估计状态回报 $V_{\varphi^h}(s_t^h)$。高层评论家网络通过最小化以下损失函数进行训练:

$$\mathcal{L}_v^h = \mathbb{E}_t[(V_{\varphi^h}(s_t^h) - R_t^h)^2], \quad R_t^h = \gamma V_{\varphi^h}(s_{t+1}^h) + r_t^h, \tag{3}$$

其中，φ^h 和 φ_-^h 是高层评论家中在线网络和目标网络的参数。折扣因子 $\gamma \in [0,1]$用于平衡即时奖励与未来奖励的重要性。行动者的主网络通过最小化以下损失函数进行训练:

$$\mathcal{L}_a^h = -\mathbb{E}\left[\log \pi_{\theta^l}(g_t|s_t^h) \cdot A^h(s_t^h, g_t)\right], \tag{4}$$

其中，A^h 是优势函数，可按如下方式计算:

$$A^h(s_t^h, g_t) = \gamma V_{\varphi^h}(s_{t+1}^h) + r_t^h - V_{\varphi^h}(s_t^h), \tag{5}$$

高层奖励机制。
公平性引导目标 g_t 用于引导 LRA 执行接下来的 M 步。然而，并非每个生成的 g_t 都有意义，因为它可能会严重偏离真实的物品空间。为解决这一问题，我们在奖励函数中引入以下约束:

$$\|g_t - v_{center}\|_2 \leq \max_{i \in \mathcal{I}} \left(\|v_i - v_{center}\|_2\right), \tag{6}$$

其中，$\mathcal{I}$ 表示物品集合，v_i 是物品 i的编码向量，v_{center} 表示所有物品编码的中心点。函数 $\|\cdot\|_2$ 表示 $L2$ 范数，即欧几里得距离。此外，高层智能体的奖励函数应同时考虑接下来 M 步的推荐准确性和物品侧公平性。奖励函数公式如下:

$$r_t^h = \begin{cases} \sum_{j=t}^{t+M} \left(r_j^a + \lambda_f \cdot r_j^f\right), & \text{if Equation(6) is ture} \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

其中，r_j^a 表示在时间 j推荐的物品所获得的准确性奖励，该奖励可根据用户反馈（例如评分或点击次数）推导得到。r_j^f 表示物品侧公平性奖励，其定义如下:

$$r_j^f = -\log(pop_j), \quad pop_j = \frac{N_j}{|U|}, \tag{8}$$

其中，N_j 是在时间 j对所推荐物品提供正面反馈的用户数量，$|U|$ 是用户总数。参数 λ_f 是一个加权因子，用于平衡推荐准确率和公平性。

3.4 微学习

低层智能体（LRA）根据 HRA 生成的公平性引导目标和用户当前偏好选择推荐物品。如图 2 所示，在每个推荐步骤中，LRA 都旨在引导用户偏好趋向公平性引导目标 g_t，同时确保用户满意度持续保持。因此，在 $Step_t$ 时，物品 2 是最佳选择；而在 $Step_{t+1}$ 时，物品 1 则成为最佳推荐。

低层状态跟踪器。
低层状态不仅包含用户近期交互的信息，反映用户偏好，还包含一个公平性引导目标，引导系统促进公平的物品曝光。与 HRA 类似，它首先将最近的 N 用户交互 $\{h_{t-N+1}, \ldots, h_t\}$ 通过嵌入层映射，然后将嵌入向量投影到查询矩阵 Q^l、键矩阵 K^l和值矩阵 V^l。最后，使用注意力层提取用户偏好：

$$p_t = Attention(Q^l, K^l, V^l) = softmax\left(\frac{Q^l (K^l)^T}{\sqrt{d^l}}\right) V^l, \tag{9}$$

其中， d^l 是一个缩放因子。将偏好嵌入 p_t 与公平性引导目标 g_t 拼接，得到低层状态： $s_t^l = p_t \oplus g_t$。

低层智能体模型。
低层行动者负责生成推荐物品，使其既符合用户当前偏好，又能引导用户偏好朝公平性引导目标g_t靠拢。然而，庞大的物品空间带来了挑战，因为许多物品偏离了公平性引导目标所指示的预期方向。为实现 LRA 的高效学习，我们对候选物品应用**过滤机制**。首先，我们计算每个物品的编码与公平性引导目标 g_t之间的 $L2$ 距离，然后选出与 g_t 距离最近的前 L 个物品作为候选物品。每个物品被推荐的概率计算如下：

$$P(\mathcal{I} \mid s_t^l) = \phi\left(MLP(s_t^l) \cdot Mask\right), \tag{10}$$

其中， MLP表示多层感知机， $\mathcal{I}$ 表示整个物品空间。 $Mask$ 是一个掩码向量，其中已选候选物品对应的位置设为 1，其余位置均设为 0。 $\phi(\cdot)$ 是一种归一化操作。

在得到分布 $P(\mathcal{I} \mid s_t^l)$ 后，对物品进行采样，形成一个既符合用户偏好又能促进物品侧公平性的列表。低层智能体采用行动者-评论家方法，其损失函数与公式 (3) 和 (4) 类似。

低层奖励机制。
无效的公平性引导目标 g_t 与真实物品空间存在显著偏差，会削弱低层智能体学习过程的有效性。因此，我们也将公式6中的约束纳入低层奖励函数。此外，低层智能体的推荐既应准确符合用户当前偏好，也应引导用户偏好逐步朝期望目标 g_t 演变。为实现这些目标，低层奖励函数设计如下：

$$r_t^l = \begin{cases} r_t^a + \lambda_g \cdot r_t^g, & \text{if Equation}(6) \text{ is true} \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

$$r_t^g = Distance(p_t, g_t) - Distance(p_{t+1}, g_t), \tag{12}$$

其中，r_t^a 表示准确率奖励，可根据评分或点击次数等用户反馈得出，而 λ_g 是一个超参数，用于权衡用户满意度与用户偏好的引导。引入引导奖励分量 r_t^g ，以引导用户偏好朝公平性引导目标 g_t指定的方向发展。函数 $Distance(\cdot)$使用 $L2$ 范数。

4 实验

本节将开展一系列实验，以回答以下与 HRL4PFG 相关的研究问题：

- **RQ1**：在交互式推荐场景中，HRL4PFG 与其他最先进的方法相比表现如何?
- **RQ2**：在不同的公平性约束下，HRL4PFG 在各种环境中的表现如何?
- **RQ3**：HRL4PFG 中的不同模块在提升推荐性能方面有多大作用?
- **RQ4**：采用不同的超参数设置时，HRL4PFG 的表现如何?

4.1 实验设置

本节介绍实验设置，包括环境模拟器、评估指标、基线方法和参数设置的详细信息。

环境模拟器。
由于在线实验成本高昂，我们基于 EasyRL4Rec 库 [Yu 等， 2024] 开发了一个交互式仿真环境，用于研究在交互式推荐场景中提升物品侧公平性并最大化用户长期满意度的方法。我们使用近期提出的 KuaiRec [Gao 等，2022b] 和 KuaiRand [Gao等，2022a] 数据集构建了两个仿真环境，这些数据集提供了高质量的用户交互日志。我们按照表 1 所示划分数据集，以构建独立的训练和测试环境。在训练推荐模型期间，测试环境始终不可见。

引入最大交互长度（最大长度），用以建模真实的会话级推荐场景，即用户在有限时间段内与系统进行交互。我们考虑最大长度为 = 30 和50，以模拟中期与长期交互。这意味着用户在每个会话中最多可产生30次或50次交互。与 [Shi 等人，2024；Yu 等人， 2024], 类似，我们引入流行度感知退出机制，使仿真环境能够评估模型推荐的物品侧公平性：若推荐模型连续推荐 W 个按流行度排名前20%的物品，则交互终止。 W 的取值越小，意味着

数据集	用途	用户	物品	交互
KuaiRec	训练集	7,176	10,728	12530.8k
	Test	1,411	3,327	4676.5k
KuaiRand	训练集	26,285	7,551	1436.6k
	Test	27,285	7,583	1186.1k

表 1：数据集统计

Envs	方法	最大长度 30				最大长度 50			
		R_{cum} ↑	R_{single} ↑	长度 ↑	基尼指数（%）↓	R_{cum} ↑	R_{single} ↑	长度 ↑	基尼指数（%）↓
KuaiRec	SQN	13.46 ± 1.96	1.00 ± 0.03	13.61 ± 2.35	99.40 ± 0.10	13.23 ± 4.32	0.99 ± 0.04	13.65 ± 5.13	99.30 ± 0.10
	PG	14.06 ± 1.23	0.79 ± 0.08	18.32 ± 2.95	99.10 ± 0.20	17.76 ± 1.60	0.71 ± 0.10	26.02 ± 5.19	98.70 ± 0.20
	DDPG	8.23 ± 3.99	0.80 ± 0.22	13.03 ± 8.59	99.50 ± 0.20	12.27 ± 4.41	0.42 ± 0.17	32.92 ± 10.29	99.40 ± 0.30
	TD3	8.27 ± 5.68	0.44 ± 0.28	23.13 ± 8.29	99.60 ± 0.30	9.34 ± 6.69	0.50 ± 0.34	31.39 ± 19.43	99.30 ± 0.60
	C51	13.90 ± 3.38	0.77 ± 0.09	18.84 ± 6.03	99.00 ± 0.20	19.12 ± 5.50	0.76 ± 0.06	25.61 ± 8.47	98.40 ± 0.30
	SAC4IR	4.86 ± 1.57	0.25 ± 0.12	23.39 ± 4.13	99.70 ± 0.20	9.60 ± 3.04	0.26 ± 0.08	38.21 ± 4.19	99.60 ± 0.30
	DNAIR	17.40 ± 3.09	0.84 ± 0.05	21.04 ± 4.42	99.20 ± 0.20	21.85 ± 5.58	0.82 ± 0.05	26.80 ± 6.96	98.90 ± 0.30
	HRL4PFG	**29.99 ± 2.98**	**1.07 ± 0.04**	**27.82 ± 2.34**	**98.20 ± 1.30**	**43.37 ± 4.00**	**1.08 ± 0.03**	**40.25 ± 3.63**	**97.20 ± 1.10**
KuaiRand	SQN	1.74 ± 0.08	**0.42 ± 0.02**	4.11 ± 0.07	99.90 ± 0.00	1.76 ± 0.08	**0.43 ± 0.02**	4.07 ± 0.05	99.90 ± 0.00
	PG	3.32 ± 0.51	0.33 ± 0.02	10.24 ± 1.63	99.60 ± 0.10	6.46 ± 0.70	0.28 ± 0.01	22.70 ± 2.35	98.40 ± 0.70
	DDPG	5.41 ± 1.76	0.26 ± 0.03	21.23 ± 7.26	99.70 ± 0.10	6.96 ± 3.02	0.25 ± 0.04	29.20 ± 13.81	99.60 ± 0.20
	TD3	4.35 ± 1.52	0.26 ± 0.04	18.50 ± 7.22	99.60 ± 0.10	6.09 ± 2.97	0.27 ± 0.05	25.61 ± 14.12	99.60 ± 0.30
	C51	2.68 ± 0.62	0.32 ± 0.03	8.79 ± 2.77	98.70 ± 0.20	2.88 ± 0.99	0.34 ± 0.02	8.98 ± 4.08	98.30 ± 0.30
	SAC4IR	6.21 ± 2.17	0.28 ± 0.03	23.56 ± 8.12	99.70 ± 0.10	5.85 ± 3.34	0.28 ± 0.05	24.00 ± 16.43	99.70 ± 0.20
	DNAIR	5.97 ± 0.87	0.26 ± 0.02	22.78 ± 3.16	98.10 ± 1.00	8.00 ± 1.36	0.27 ± 0.02	30.54 ± 5.88	98.20 ± 0.80
	HRL4PFG	**10.03 ± 1.85**	0.36 ± 0.04	**27.90 ± 2.25**	**95.50 ± 5.30**	**15.41 ± 2.78**	0.35 ± 0.03	**43.40 ± 4.83**	**94.70 ± 6.60**

表 2：不同环境下所有方法的平均结果。

对模型推荐公平性的要求越高。

评估指标。
我们在交互推荐场景中从用户满意度和物品侧公平性两个维度评估模型性能。

用户满意度指标。 遵循 [Yu 等人， 2024],我们采用**单轮奖励**（R_{single}）和**累积奖励** （R_{cum}）分别衡量即时与长期用户满意度。R_{single} 反映每个交互步骤的推荐质量，而R_{cum} 刻画模型在整个会话中维持高质量交互的能力。我们进一步报告**交互长度** （长度），用于衡量用户在退出系统前完成的交互次数。

物品侧公平性指标。为明确量化物品之间的曝光公平性，我们采用**基尼指数**[Ge 等人， 2021], 该指标衡量整个推荐周期内物品曝光分布的不平等程度。较低的基尼指数反映更好的物品侧公平性。

基线方法。
为公平、全面地评估HRL4PFG的有效性，我们将其与基于强化学习的方法以及公平性感知强化学习方法进行比较。基线方法包括：SQN [Xin等人， 2020], 一种结合交叉熵损失与强化学习进行推荐的双头模型。策略梯度 [Williams，1992], 一种通过估计奖励梯度直接优化策略的策略梯度方法。深度确定性策略梯度 [Lillicrap，2015], 一种使用行动者-评论家网络进行优化的确定性策略梯度算法。TD3 [Fujimoto 等人， 2018], 一种通过使用两个评论家网络及最小Q值来降低过估计偏差、从而实现稳定学习的强化学习算法。C51 [Bellemare等人， 2017], 一种对完整奖励分布而非期望值进行建模的分布式强化学习方法。SAC4IR [Shi等人， 2024], 一种利用最大熵约束来增强长期公平性的公平性感知强化学习方法。DNAIR [Shi 等人， 2023], 一种融合物品流行度与质量以缓解流行度偏差的基于强化学习的方法。

SQN 使用离线日志进行训练，而其他算法则采用预训练用户模型（DeepFM [Guo 等人， 2017]）以引导在线智能体的学习。所有方法在各个环境中均使用退出参数 $W = 3$ 。我们为所有方法调优了关键参数。详细的参数设置和代码见附录。

4.2 总体性能比较（RQ1）

我们在不同的交互环境中对每种方法进行了 50 个训练轮次。在每个轮次学习之后，我们用 100 个回合（即交互轨迹）评估所有方法。结果如表 2 所示，

物品侧公平性分析。 HRL4PFG 在 KuaiRec 和 KuaiRand 的不同交互时域下均持续取得最低的基尼指数。这表明 HRL4PFG 有效降低了热门物品与长尾物品之间的曝光不平等。与公平性感知基线相比，HRL4PFG 进一步降低了基尼指数，证明了主动的、与偏好对齐的公平性引导相对于基于反应式正则化的方法的优势。

公平性约束下的用户满意度。 重要的是，曝光公平性的提升并非以牺牲用户满意度为代价。相反，HRL4PFG 在所有设置下同时取得了最高的累积奖励和最长的交互长度。这是因为 HRL4PFG 采用渐进式引导机制，平滑地将用户偏好转向长尾物品，从而实现物品侧公平性目标与用户偏好之间的有效对齐。SAC4IR 和 DNAIR 等公平性感知基线在一定程度上缓解了曝光失衡，但在维持长期参与度和累积奖励方面仍不及 HRL4PFG。

4.3 不同环境下的结果（RQ2）

为验证 HRL4PFG 在不同环境设置下的鲁棒性，我们将退出机制参数W 分别设为 2、3、4、5 和 6，同时将最大交互长度固定为 30。较大的 W 表示系统对物品侧公平性的要求较低。实验结果如图 3 所示。在不同的物品侧公平性要求下，HRL4PFG 始终表现最佳，证明了其在多种环境中的鲁棒性。

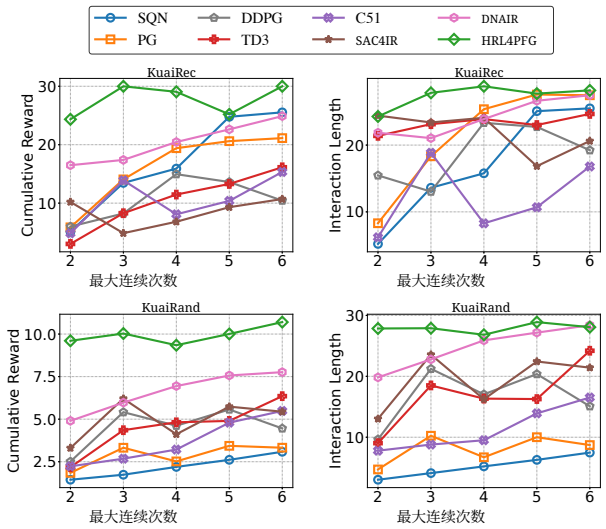

图 3：不同退出条件下的结果。

4.4 消融研究（研究问题3）

为评估所提出框架中各组件的有效性，我们通过定义以下模型变体开展消融研究：1）**HRL4PFG-wo-hie**：该变体移除了层次结构，即去掉高层智能体，仅保留低层智能体进行推荐；2）**HRL4PFG-wo-tc**：该变体移除了对高层智能体生成的公平性引导目标的约束，如公式6所述；3）**HRL4PFG-wo-fm**：该变体移除了公式10所述的过滤机制 $Mask$；4）**HRL4PFG**：所提出框架的完整版本。

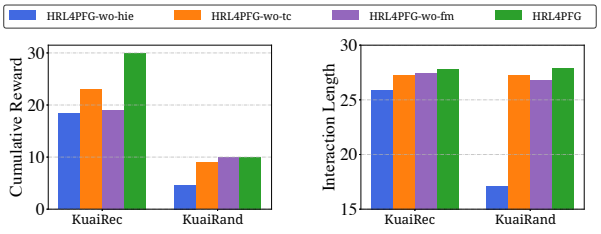

图4：HRL4PFG 的消融研究结果。

实验在仿真环境中进行，将最大用户交互长度设为30，并将退出机制参数 W 设为3。结果如图4所示。**HRL4PFG-wo-hie** 在物品侧公平性和准确率方面的表现都最差，说明分层学习对于公平性感知推荐是必要的。**HRL4PFG-wo-tc** 的表现优于**HRL4PFG-wo-hie** ，但仍不及完整模型，这表明对高层智能体生成的公平性引导目标施加约束可以提升模型性能。**HRL4PFG-wo-fm** 的表现不如**HRL4PFG**，说明根据公平性引导目标过滤物品有助于低层智能体的学习过程。

4.5 超参数敏感性分析（研究问题4）

在本小节中，我们考察公式11中的 λ_g 和公式7中的 M 对模型性能的影响。其中，λ_g 控制施加于低层智能体的引导强度， M 则决定高层智能体公平性引导目标的更新间隔。

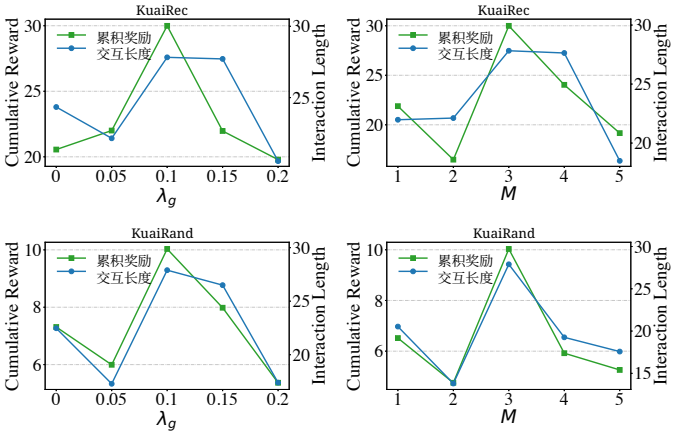

图5：两个数据集上的超参数研究。

在将最大交互长度设为 30、退出机制参数 W 设为 3 的条件下，我们在两个数据集上进行了实验。不同参数设置下的结果如图5所示。随着 λ_g 增加，推荐准确率和公平性均得到提升，直到达到 $\lambda_g = 0.1$，之后性能开始下降。这表明，高层智能体生成的公平性引导目标不应在单步内对低层智能体的决策过程进行过度引导。同样，增大 M 会提升性能，直到达到 $M = 3$，超过该值后性能会下降，这可能是因为更新间隔过短会缺乏前瞻性，而间隔过长又会使目标过于容易达成，从而削弱引导效果。

5 结论

本文提出了 HRL4PFG，这是一种用于提升交互式推荐系统中物品侧公平性的主动引导策略。HRL4PFG 采用分层方法，逐步引导用户偏好转向长尾物品，在提升物品侧公平性的同时维持用户满意度。实验结果表明，与现有方法相比，HRL4PFG 同时提升了物品侧公平性和用户满意度。

未来的研究方向包括将 HRL4PFG 应用于人口统计公平性等其他公平性目标，整合更丰富的用户行为以提升适应能力，并开展现实世界研究，评估其对公平性和满意度的长期影响。

参考文献

[Abdollahpouri 等, 2017] Himan Abdollahpouri、Robin Burke 和 Bamshad Mobasher。《学习排序推荐中的流行度偏差控制》。载于论文集

第十一届 *ACM* 推荐系统会议，第 42–46 页，2017 年。

[Abdollahpouri 等， 2019] Himan Abdollahpouri、Masoud Mansoury、Robin Burke 和 Bamshad Mobasher。《推荐中的流行度偏差之不公平性》。*arXiv* 预印本 *arXiv:1907.13286*，2019 年。

[Bellemare等， 2017] Marc G Bellemare、Will Dabney 和 R´emi Munos。《强化学习的分布视角》。载于国际机器学习会议，第 449–458 页。PMLR，2017 年。

[Beutel等， 2019] Alex Beutel、Jilin Chen、Tulsee Doshi、Hai Qian、Li Wei、Yi Wu、Lukasz Heldt、Zhe Zhao、Lichan Hong、Ed H Chi 等。《通过成对比较实现推荐排序中的公平性》。载于第 *25* 届 *ACM SIGKDD* 国际知识发现与数据挖掘会议论文集，第 2212–2220 页，2019 年。

[Bharadhwaj， 2019] Homanga Bharadhwaj。面向用户冷启动推荐的元学习。载于 *2019*年国际神经网络联合会议（*IJCNN*），第1–8页。IEEE，2019年。

[Bi等， 2024] Shuxian Bi、Wenjie Wang、Hang Pan、Fuli Feng 和 Xiangnan He。通过迭代式偏好引导实现主动推荐。载于*2024*年*ACM Web*会议附属论文集，第871–874页，2024年。

[Curmei等， 2022] Mihaela Curmei、Andreas A Haupt、Benjamin Recht 和 Dylan Hadfield-Menell。迈向心理学基础的动态偏好模型。载于第*16*届*ACM*推荐系统会议论文集，第35–48页，2022年。

[Dean 和 Morgenstern， 2022] Sarah Dean 和 Jamie Mor- genstern。个性化推荐下的偏好动态。载于第*23*届*ACM*经济学与计算会议论文集，第795–816页，2022年。

[Fujimoto等, 2018] Scott Fujimoto、Herke Hoof 和 David Meger。《解决行动者-评论家方法中的函数近似误差》。载于国际机器学习会议，第 1587–1596 页。PMLR，2018 年。

[Gao等, 2022a] Chongming Gao、Shijun Li 等。《KuaiRand：包含随机曝光视频的无偏序列推荐数据集》。载于第 *31* 届 *ACM* 信息与知识管理国际会议论文集，第 3953–3957 页，2022 年。

[Gao等, 2022b] Chongming Gao、Shijun Li 等。《KuaiRec：用于评估推荐系统的全观测数据集及其启示》。载于第 *31* 届 *ACM* 信息与知识管理国际会议论文集，第 540–550 页，2022 年。

[Ge等人, 2021] Yingqiang Ge、Shuchang Liu、Ruoyuan Gao、Yikun Xian、Yunqi Li、Xiangyu Zhao、Changhua Pei、Fei Sun、Junfeng Ge、Wenwu Ou 等人。迈向推荐中的长期公平性。发表于第*14*届 *ACM* 网络搜索与数据挖掘国际会议论文集，第445–453页，2021年。

[Geyik等人, 2019] Sahin Cem Geyik、Stuart Ambler 和 Krishnaram Kenthapadi。将公平性感知排序应用于搜索和推荐系统：以 LinkedIn 人才搜索为例。发表于第*25*届 *ACM SIGKDD* 知识发现与数据挖掘国际会议论文集，第2221–2231页，2019年。[Guo等人, 2017] Huifeng Guo、Ruiming Tang、Yunming Ye、Zhenguo Li 和 Xiuqiang He。DeepFM：一种基于因子分解机的神经网络，用于点击率预测。*arXiv* 预印本 *arXiv:1703.04247*，2017年。

[He 等，2016] Chen He、Denis Parra 和 Katrien Verbert。《交互式推荐系统：现状综述及未来研究挑战与机遇》。《专家系统与应用》，56：9–27，2016 年。

[Konda 和 Tsitsiklis， 1999] Vijay Konda 和 John Tsitsiklis。演员-评论家算法。神经信息处理系统进展，第 12 卷，1999 年。

[Lillicrap， 2015] TP Lillicrap。基于深度强化学习的连续控制。*arXiv* 预印本 *arXiv:1509.02971*，2015 年。

[Quan等， 2023] Yuhan Quan、Jingtao Ding、Chen Gao、Nian Li、Lingling Yi、Depeng Jin 和 Yong Li。缓解微视频推荐中的视频长度效应。*ACM* 信息系统汇刊，42(2)：1–24，2023 年。

[Shi等， 2023] Xiaoyu Shi、Quanliang Liu、Hong Xie、Di Wu、Bo Peng、MingSheng Shang 和 Defu Lian。缓解交互式推荐中的流行度偏差：一种多样性与新颖性感知强化学习方法。*ACM* 信息系统汇刊，42(2)：1–30，2023 年。

[Shi 等, 2024] Xiaoyu Shi、Quanliang Liu、Hong Xie、Yanan Bai 和 Mingsheng Shang。交互式推荐系统中长期公平性的最大熵策略。*IEEE* 服务计算汇刊，2024。

[Singh 和 Joachims， 2019] Ashudeep Singh 和 Thorsten Joachims。排序公平性的策略学习。神经信息处理系统进展，32，2019。

[Tang 等, 2019] Jianheng Tang、Tiancheng Zhao、Chenyan Xiong、Xiaodan Liang、Eric P Xing 和 Zhiting Hu。目标引导的开放域对话。*arXiv* 预印本 *arXiv:1905.11553*，2019。

[Wang 等, 2023] Yifan Wang、Weizhi Ma、Min Zhang、Yiqun Liu 和 Shaoping Ma。推荐系统公平性综述。*ACM* 信息系统汇刊，41(3):1–43，2023。

[Wei 等, 2021] Tianxin Wei、Fuli Feng、Jiawei Chen、Zi- wei Wu、Jinfeng Yi 和 Xiangnan He. 一种用于消除推荐系统中流行度偏差的模型无关的反事实推理. 载于第 *27* 届 *ACM SIGKDD* 知识发现与数据挖掘会议论文集，第 1791–1800 页，2021 年。

[Williams, 1992] Ronald J Williams. 用于连接主义强化学习的简单统计梯度跟随算法. 机器学习，8:229–256，1992 年。

[Wu等人， 2019] Wenquan Wu、Zhen Guo、Xiangyang Zhou、Hua Wu、Xiyuan Zhang、Rongzhong Lian和 Haifeng Wang。具有显式对话目标的主动式人机对话。*arXiv*预印本 *arXiv:1906.05572*，2019。

[Xin 等， 2020] Xin Xin、Alexandros Karatzoglou、Ioannis Arapakis 和 Joemon M Jose。用于推荐系统的自监督强化学习。载于 第*43*届国际 *ACM SIGIR* 信息检索研究与发展会议论文集，第931–940页，2020年。[Yu 等， 2024] Yuanqing Yu、Chongming Gao、Jiawei Chen、Heng Tang、Yuefeng Sun、Qian Chen、Weizhi Ma 和 Min Zhang。Easyrl4rec：一个易于使用的基于强化学习的推荐系统库。载于第*47*届国际 *ACM SIGIR* 信息检索研究与发展会议论文集，第977–987页，2024年。[Zhang等， 2019] Shuai Zhang、Lina Yao、Aixin Sun 和 Yi Tay。基于深度学习的推荐系统：综述与新视角。*ACM* 计算调查（*CSUR*），52(1)：1–38，2019年。[Zhang等， 2025] Tong Zhang、Nitin Bisht、Zihao Li、Guandong Xu 和 Xianzhi Wang。Sarrec：统计保证的增强检索推荐。载于第*34*届 *ACM* 信息与知识管理国际会议论文集，第4181–4190页，2025年。[Zhao等， 2020] Dongyang Zhao、Liang Zhang、Bo Zhang、Lizhou Zheng、Yongjun Bao 和 Weipeng Yan。Mahrl：基于多目标抽象的深度分层强化学习推荐。载于 第*44*届 *ACM SIGIR* 信息检索研究与发展国际会议论文集，第871–880页，2020年。[Zhao等， 2023] Sen Zhao、Wei Wei、Yifan Liu、Ziyang Wang、Wendi Li、Xian-Ling Mao、Shuai Zhu、Minghui Yang 和 Zujie Wen。面向基于超图强化学习的对话式推荐分层策略学习。载于第三十二届国际人工智能联合会议论文集，第2459–2467页，2023年。[Zheng等， 2018] Guanjie Zheng、Fuzheng Zhang、Zihan Zheng、Yang Xiang、Nicholas Jing Yuan、Xing Xie 和 Zhenhui Li。Drn：用于新闻推荐的深度强化学习框架。载于*2018* 年万维网大会论文集，第167–176页，2018年。[Zhu等， 2018] Ziwei Zhu、Xia Hu 和 James Caverlee。公平性感知的张量推荐。载于第*27*届 *ACM* 信息与知识管理国际会议论文集，第1153–1162页，2018年。

[Zhu 等， 2021] Ziwei Zhu、Jingu Kim、Trung Nguyen、Aish Fenton 和 James Caverlee。冷启动推荐系统中新物品之间的公平性。载于 第*44*届 *ACM SIGIR* 信息检索研究与发展国际会议论文集，第767–776页，2021年。[Zhu 等， 2023] Haoren Zhu、Hao Ge、Xiaodong Gu、Pengfei Zhao 和 Dik Lun Lee。有影响力的推荐系统。载于*2023*年 *IEEE* 第*39*届国际数据工程会议（*ICDE*）论文集，第1406–1419页。IEEE，2023年。