并非所有候选项都生而平等：一种用于推荐系统预排序的异质性感知方法

Pengfei Tong 字节跳动 杭州，中国
tongpengfei@bytedance.com

Siyuan Chen 字节跳动 北京，中国
chensiyuan.279@bytedance.com

Chenwei Zhang 字节跳动 深圳，中国
zhangchenwei.929@bytedance.com

Bo Wang 字节跳动 北京，中国
wangbo.9830@bytedance.com

Qi Pi †字节跳动 杭州，中国
wk@bytedance.com

Pixun Li 字节跳动 北京，中国
lipixun@bytedance.com

Zuotao Liu 字节跳动 上海，中国
mic
hael.liu@bytedance.com

arXiv:2603.03770v1 [cs.IR] 4 Mar 2026

摘要

大多数大规模推荐系统采用多阶段级联架构，包括召回、预排序、排序和重排序。粗排阶段的一项关键挑战，源于从粗粒度召回结果、细粒度排序信号和曝光反馈中采样得到的训练样本存在异质性。我们的分析表明，当前的预排序方法不加区分地混合异质样本，因而面临梯度冲突：困难样本主导训练，而简单样本未得到充分利用，导致性能不佳。我们进一步发现，对所有样本统一扩展模型复杂度的常见做法效率低下，因为它在简单样本上耗费了过多计算，拖慢训练速度，却未带来相称的收益。为解决这些局限，本文提出**异**质性感知**自**适应**预**排序（**HAP**）框架。该统一框架通过冲突敏感采样与定制化损失设计相结合来缓解梯度冲突，同时在候选项之间自适应分配计算预算。具体而言，HAP 将简单样本与困难样本分离，并为每个子集采用专门的优化路径。在此基础上，HAP 首先对所有候选项使用轻量级模型，以高效实现覆盖；然后进一步对困难样本使用更强大的模型，在降低成本的同时保持准确率。这种方法不仅提升了预排序的有效性，也为工业推荐系统中的扩展策略提供了实用思路。HAP 已在今日头条生产系统中部署 9 个月，使用户应用使用时长最高提升**0.4%**，活跃天数提升**0.05%**，且未增加计算成本。我们还发布了一个大规模工业级混合样本数据集，以支持对预排序中来源驱动的候选项异质性开展系统研究。

CCS概念

• **信息系统** →**召回模型与排序**。

关键词

预排序、双阶段架构、候选项异质性、梯度协调、性能与效率的权衡

†通讯作者。

1 引言

现代推荐系统旨在从数亿候选项中识别出符合用户兴趣的内容。为平衡有效性与效率，多阶段级联已成为主流范式 [3, 5, 13, 21], , 包括召回 [8, 31, 33],、预排序 [16, 22, 24], 、排序 [19, 20, 32],以及重排序 [12, 14,23] 等阶段。预排序位于召回与排序之间，需在数毫秒内将数千个召回项筛选至数百个，以供后续排序。此处的核心挑战在于候选项的多样性：许多项目显然无关且易于过滤，而另一些则是近似正样本，与真实标签高度相似，难以区分。这种简单样本与困难样本的混合引入了固有异质性，导致相互竞争的学习信号，从而给有效优化带来挑战。

以往研究主要通过两种截然不同的范式来应对这种异质性，而我们认为这两种范式都不够理想。**1）不加区分地扩充数据及其优化陷阱**。应对分布多样性的一种常见策略是扩充负样本训练集。早期研究 [22, 27]将这一问题表述为样本选择偏差（SSB），并采用困难负样本挖掘来提升模型在分布偏移下的泛化能力。近期研究 [28] 进一步拓展了这一思路，将来自上游召回的未曝光负样本，以及来自预排序和下游排序的低排序物品纳入其中。尽管这些策略无疑扩大了数据覆盖率，却也引发了严重的优化冲突。由此形成的易负样本与困难负样本混合数据会使困难样本主导训练，将模型引向次优局部极小值，使其过拟合噪声并破坏优化过程（图1）。**2）盲目模型扩展及其不对称收益**。除了扩充数据，另一种常见的工业实践是提升模型容量，以期捕捉复杂的用户-物品交互模式。研究表明，与排序阶段相匹配的大型模型能够提高预排序准确率 [24, 25, 29, 30]。然而，收益大多局限于难以排序的候选项，对大多数较容易排序的候选项仅带来微乎其微的提升。因此，在不解决异质候选项造成的梯度冲突的情况下，一味追求模型表达能力，是一种次优且代价高昂的策略。

基于这些洞见，我们提出**异**质性感知**自**适应**预**排序（**HAP**），这是一个统一框架，旨在

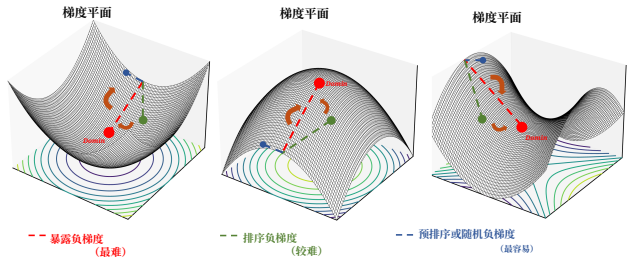


图 1：不同难度负样本之间的梯度冲突，展示了候选项固有的异质性。

应对异质候选项带来的两个关键挑战：优化冲突和计算效率低下。**1）梯度协调对比学习（GHCL）：**我们对二元交叉熵（BCE）与InfoNCE的理论分析表明，困难负样本会诱发不成比例的大梯度，从而破坏训练的稳定性。GHCL在今日头条[1]平台中构建了四个难度感知负样本集，遵循先前策略 [28], ，并引入一种定制的对比损失，以协调各负样本集之间的梯度贡献。该方法平衡简单样本与困难样本，确保全面的分布覆盖，并防止困难负样本占据主导，从而实现稳定收敛和更强的泛化能力。**2）难度感知模型路由（DAMR）：**并非所有候选项都需要相同的建模复杂度。DAMR采用渐进式路由策略。轻量级模型首先处理所有候选项以获得广泛覆盖率，而困难样本则被选择性地转发到表达能力更强的模型，从而在固定预算内高效分配计算。

通过大量定性和定量实验，我们证明 **HAP**能够有效应对异质候选项分布带来的挑战，并在无需额外工程成本的情况下显著提升预排序性能。我们的贡献总结如下：

- **具有理论基础的异质性感知预排序框架**。我们提出了一个统一框架，用于处理预排序中的海量异质候选项。基于对样本构建和梯度冲突的理论与实证研究，**HAP** 实现了推荐效果与计算效率之间的有效平衡。

- **开放式多阶段数据集**。我们发布了一个大规模、完整标注的多阶段推荐数据集，涵盖今日头条[1]，以促进可复现研究，并推动预排序和多阶段推荐领域的未来发展。

- **显著的实际收益**。我们已将该框架部署于工业级推荐系统今日头条1，用户活跃天数提升了 0.05%，应用使用时长提高了 0.4%；在这一规模下，这代表着显著的收益。

2 相关工作

预排序研究主要沿两个方向推进：样本选择优化和模型架构改进。

样本部分优化旨在通过拓宽训练分布来缓解曝光偏差，例如纳入未曝光或低排序物品 [2, 6, 15]。近期研究采用对比学习和反事实建模 [21, 27] 来提升对未见候选项的泛化能力。IRRD [9]引入物品侧正则化以对齐排序行为，而 HCCP [28]则探索跨阶段协同，以提升对低排序物品的泛化能力。然而，这些方法将所有候选项一视同仁，忽略了固有的分布异质性，可能导致优化过程中的梯度干扰。

模型架构改进侧重于提升表征能力。深度神经网络（DNN）已成为点积匹配的标准方法，可分别对用户和物品进行嵌入，以计算相关性分数 [8, 18]。后续研究 [10, 11, 16, 24] 探索先进架构和特征建模技术，以增强模型的能力。COLD[24] 采用特征选择来捕捉显著信号，FSCD [16]通过变分丢弃进行自适应特征剪枝，而 COPR [30] 则使预排序与精排阶段输出保持一致。尽管这些方法增强了对困难样本的建模能力，但对简单样本带来的收益有限，导致回报不对称且计算成本增加。此外，扩大模型规模并不能解决由样本难度异质性引发的根本性梯度冲突，因此这些策略既效率低下，也不足以解决问题。

总之，现有工作已通过更强的表达能力、更广的样本覆盖以及阶段协同推进了预排序，但它们忽视了异质样本优化与高效模型分配这一双重挑战。为克服这些局限，我们提出HAP，一种异质性感知的预排序框架，它联合协调梯度学习，并根据候选项的难度自适应地路由候选项，从而同时实现有效优化与高效计算。

3 方法

3.1 预备知识

我们首先详细说明预排序推荐场景中的数据构成。对于每个用户请求 q，系统通过多阶段流水线处理数十亿候选项，并在以用户信息 u为条件的情况下，返回一组排名靠前的项目 $E^q = \{e_1, e_2, \ldots, e_n\}$以供最终曝光，其中 e_i表示一项曝光内容， n是每个会话的最大项目数。在阶段k，候选集表示为$C_k^q = \{c_1, c_2, \ldots, c_l\}$，其中 c_l是该阶段的第 l个候选项。源自单个请求的训练样本被分为五种类型：

- **曝光正样本（EP）：**用户实际点击的项目，定义为 SEP $= \{(u, x_i) | x_i \in E^q,$ 点击 $= 1\}$。这些样本反映了经验证的用户兴趣。

- **曝光负样本（EN）：**已展示但未被点击的项目，定义为 SEN $= \{(u, x_i) | x_i \in E^q,$ 点击 $= 0\}$。这些样本提供明确的负反馈信号。

- **排序负样本（RN）：**在后续排序阶段中排序低于阈值 RankL 的候选项，定义为 SRN $= \{(u, x_i) | x_i \in C^q$ 排序$i >$ RankL$\}$。这些样本有助于使预排序与下游排序目标保持一致。

- **预排序负样本（PRN）：**由当前预排序阶段评分低于阈值 PrerankL 的候选项，定义为 SPRN $= \{(u, x_i) | x_i \in C^q$ pre-ranking$i >$ PrerankL$\}$。这些样本通过拉大排名靠前与靠后的候选项之间的分数差距，增强模型的排序区分能力。

[1]中国最大的新闻推荐平台之一

表1：二元交叉熵和InfoNCE损失函数下不同难度负样本的梯度特征比较。样本类型：已曝光（EN）、排序（RN）、预排序（PRN）和全局随机样本（GN）。

Loss	样本	难度	梯度范数	Logits	余弦相似度
BCE	EN	最难	1.101	0.322	–
	RN	较难	0.254	0.232	0.816
	PRN	较容易	0.140	0.191	0.496
	GN	最容易	0.060	0.173	0.237
InfoNCE	EN	最难	1.750	0.326	–
	RN	较难	0.401	0.235	0.820
	PRN	较易	0.352	0.192	0.607
	GN	最易	0.162	0.176	0.576

• **全局随机负样本（GN）：** 从召回结果中随机采样的候选项，定义为 SGN $= \{x_i \sim \text{Uniform}(\text{C}^q \text{ retrieval})\}_1^N$，其中 N为样本量。这些样本可以缓解样本选择偏差（SSB），并提高分布覆盖。

尽管多样化样本能够增强模型泛化，但它们也会在数据分布和学习难度上引入显著的异质性。这种异质性对统一优化构成了关键挑战，从而促使我们提出异质性感知设计。整体框架如图2所示。

3.2 异质性导致的梯度主导

为了解样本异质性如何影响优化，我们分析了常用预排序损失函数下不同样本类型的梯度表现。我们根据样本的学习难度对其进行分类，并考察其梯度范数（梯度范数）、logit 值（*Logits*）和梯度余弦相似度（余弦相似度），如表1所示。我们的分析发现，logit 值通常会随样本难度的增加而增大。具体而言，对于用户 u，困难负样本和易负样本的 logit 如下：

$$z_u^{\text{hn}} = f(x_u^i, \theta), \quad x_u^i \in D_u^{\text{hn}} \tag{1}$$

$$z_u^{\text{en}} = f(x_u^i, \theta), \quad x_u^i \in D_u^{\text{en}} \tag{2}$$

从经验上看，我们观察到 $z_u^{\text{hn}} > z_u^{\text{en}}$，这符合直觉，因为困难负样本在语义上更接近正样本，因此会获得更高的模型评分。我们还发现，困难负样本会产生更大的梯度范数，且与较易样本的余弦相似度更低，表明其梯度方向存在冲突。为解释这些观察结果，我们分析了二元交叉熵（BCE）和 InfoNCE 损失函数的梯度公式——这两种目标函数是粗排模型中具有代表性的损失函数。

*3.2.1 BCE*中的梯度主导效应。 BCE损失广泛用于建模为二元分类的推荐任务。对于用户请求 q，训练数据定义为$\text{D}_q^{\text{BCE}} = \{(u_q, x_i, y_i)\}_{i=1}^{N_q}$，其中 N_q是样本数量，每个元组表示一个带有二元标签 y_i的用户-物品对。点击的物品标记为 $y_i = 1$，未点击的物品标记为 $y_i = 0$。令$z_i = f(u_q, x_i; \theta) \in \mathbb{R}$表示样本 i的Logits输出，$\hat{y}_i = \sigma(z_i)$表示预测概率，其中 σ是Sigmoid函数。BCE损失及其相对于 z_i的梯度为：

$$\mathcal{L}_{\text{BCE}} = \sum_q \sum_{i=1}^{N_q} [-y_i \log \sigma(z_i) - (1 - y_i) \log(1 - \sigma(z_i))] \tag{3}$$

对于负样本（$y_i = 0$），梯度可简化为 $\frac{\partial \mathcal{L}_{\text{BCE}}}{\partial z_i} = \sigma(z_i)$。

由于Sigmoid函数是单调递增的，Logits较大的负样本（这表示其难度更高）往往会获得更强的梯度。如表1所示，这会形成明显的层次结构（$z_{\text{EN}} > z_{\text{RN}} > z_{\text{PRN}} > z_{\text{GN}}$），其中最难的负样本主导梯度更新。这种主导效应可能掩盖较容易的样本，阻碍平衡优化，并导致不稳定收敛。

3.2.2 InfoNCE 中的梯度主导效应。预排序阶段的目标是选出 Top-K 候选项，因此对比式 InfoNCE 损失是自然之选。对于用户请求 q，数据表示为 $\text{D}_q^{\text{InfoNCE}} = \{(u_q, x_i^+, x_1^- \cdots x_M^-)\}_{i=1}^{N_p}$，其中每个元组包含一个正样本项 x^+和不同类型的负样本项 $x_1^- \cdots x_M^-$，M 表示负样本的数量，N_p 表示正样本的数量。模型 f_θ 计算用户 u 与一个物品 $s^+ = f(x_i^+, u_q; \theta)$（针对正样本）和 $s_j^- = f(x_j^-, u_q; \theta)$（针对负样本）之间的相似度。为方便起见，下面的分析将聚焦于单个样本对的 InfoNCE。特定负样本s_k^- 对应的 InfoNCE 损失和梯度为：

$$\mathcal{L}_{\text{InfoNCE}} = -\log \frac{e^{s^+}}{e^{s^+} + \sum_j e^{s_j^-}}, \tag{4}$$

$$\frac{\partial L}{\partial s_k^-} = \frac{e^{s_k^-}}{e^{s^+} + \sum_j e^{s_j^-}}. \tag{5}$$

这一表述揭示了由负样本难度的分布偏移所引起的梯度耦合效应。每个负样本的梯度对应于其各自的 softmax 概率 $P(k^-)$；因此，极难的困难负样本（具有较高的s_k^-）会主导分母，从而抑制所有其他样本的梯度贡献。这种耦合会导致不稳定且有偏的优化，因为模型更新会不成比例地受到每个批次中最难样本的影响。因此，样本异质性会放大梯度幅度不平衡并引入样本间干扰，从而阻碍收敛稳定性。

3.3 梯度协调对比学习

我们先前的分析揭示了预排序中的一个根本挑战：不同难度负样本之间存在梯度冲突和耦合。逐点损失（如二元交叉熵）可以通过简单的样本重加权缓解这一问题，而列表式损失（如 InfoNCE）则更为敏感，因为 softmax 归一化会引发梯度相互依赖。为此，我们提出梯度协调对比损失（GHCL），这是对原始 InfoNCE 目标的一种简单而有效的改进，可显式解耦不同难度级别负样本之间的梯度。具体而言，GHCL 定义如下：

$$\mathcal{L}_{\text{GHCL}} = \underbrace{\mathcal{L}_{\text{RN}} + \mathcal{L}_{\text{EN}}}_{\mathcal{L}_{\text{hard}}} + \underbrace{\mathcal{L}_{\text{GN}} + \mathcal{L}_{\text{PRN}}}_{\mathcal{L}_{\text{easy}}} \tag{6}$$

$$\mathcal{L}_{\text{hard}} = -\log \frac{e^{s_p}}{e^{s_p} + \sum_{j \in \mathcal{N}_{\text{hard}}} e^{s_j^-}}, \tag{7}$$

$$\mathcal{L}_{\text{easy}} = -\log \frac{e^{s_p}}{e^{s_p} + \sum_{j \in \mathcal{N}_{\text{easy}}} e^{s_j^-}}. \tag{8}$$

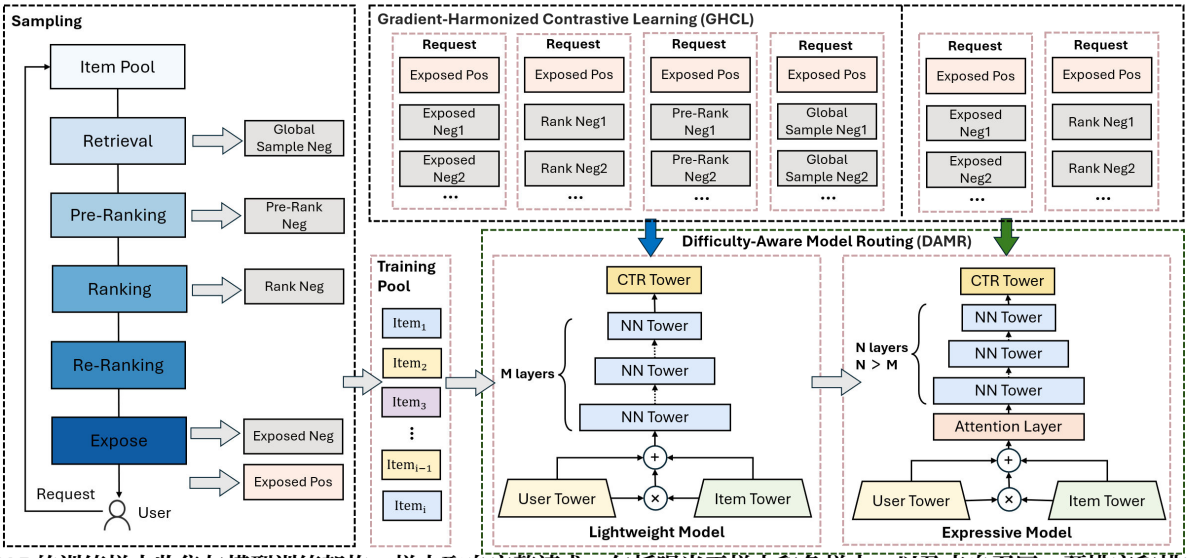


图2：HAP的训练样本收集与模型训练架构。样本取自完整请求，包括曝光正样本和负样本，以及来自召回、预排序和排序阶段的负样本。HAP采用DAMR架构：轻量级模型通过GHCL从所有样本类型中学习，对检索候选项打分并过滤简单样本，而表达能力强的模型则使用更深的基于注意力的网络，仅在困难样本上训练。

这里，s_p 表示正样本的分数，s_j^- 分别表示来自困难集 $N_{hard} \in \{N_{EN}, N_{RN}\}$ 和简单集 $N_{easy} \in \{N_{PRN}, N_{GN}\}$ 的负样本分数。通过拆分负样本，我们有效地将梯度计算限制在各个子组内，从而减轻跨组耦合。设 s_{hk}^- 和 s_{ek}^- 分别为一个困难负样本和一个简单负样本的分数。在原始 InfoNCE（L_{ORG}）下，它们的梯度为：

$$G_{ORG}^{hk} = \frac{\partial \mathcal{L}_{ORG}}{\partial s_{hk}^-} = \frac{e^{s_{hk}^-}}{e^{s_p} + \sum_j e^{s_j^-}}, \tag{9}$$

$$G_{ORG}^{ek} = \frac{\partial \mathcal{L}_{ORG}}{\partial s_{ek}^-} = \frac{e^{s_{ek}^-}}{e^{s_p} + \sum_j e^{s_j^-}}. \tag{10}$$

该比率 与它们的梯度幅值之比因此可表 示为：

$$R_{ORG} = \frac{G_{ORG}^{hk}}{G_{ORG}^{ek}} = e^{s_{hk}^- - s_{ek}^-}. \tag{11}$$

由于困难负样本通常得分较高（$s_{hk}^- > s_{ek}^-$），该比率可能呈指数级增长，表明困难负样本主导了梯度，并抑制了模型从较简单样本中学习。

在 GHCL 中，梯度是在每个子组内计算的：

$$G_{GHCL}^{hk} = \frac{e^{s_{hk}^-}}{e^{s_p} + \sum_{j \in \mathcal{N}_{hard}} e^{s_j^-}}, \tag{12}$$

$$G_{GHCL}^{ek} = \frac{e^{s_{ek}^-}}{e^{s_p} + \sum_{j \in \mathcal{N}_{easy}} e^{s_j^-}}. \tag{13}$$

所得梯度比为：

$$R_{GHCL} = \frac{G_{GHCL}^{hk}}{G_{GHCL}^{ek}} = e^{s_{hk}^- - s_{ek}^-} \cdot \left(\frac{e^{s_p} + \sum_{j \in \mathcal{N}_{easy}} e^{s_j^-}}{e^{s_p} + \sum_{j \in \mathcal{N}_{hard}} e^{s_j^-}} \right). \tag{14}$$

括号中的项充当校正因子 ：

$$C = \frac{e^{s_p} + \sum_{j \in \mathcal{N}_{easy}} e^{s_j^-}}{e^{s_p} + \sum_{j \in \mathcal{N}_{hard}} e^{s_j^-}} < 1. \tag{15}$$

因此，整体比率满足：

$$R_{GHCL} = R_{ORG} \cdot C < R_{ORG}. \tag{16}$$

GHCL 弥合了困难负样本与简单负样本之间的梯度幅值差距，有效协调了它们各自的贡献。通过平衡不同难度层级上的优化动态，GHCL 稳定了训练并增强了表征学习，尤其适用于负样本难度各异的预排序场景。

3.4 难度感知模型路由

尽管 GHCL 协调了梯度信号，但并未解决处理不同样本难度时的架构效率低下问题。我们的实证分析表明，模型容量应与样本难度相匹配。如图 3 所示，轻量级模型在简单样本上的表现与复杂模型相当，但在困难样本上的性能则急剧下降。这一观察促使我们采用难度自适应的模型层级结构：由简单模型处理简单样本，并由高表达模型专门处理困难样本。

3.4.1 使用轻量级模型进行粗粒度过滤。轻量级模型 f_l 能够高效处理召回阶段生成的完整候选集，以较低的计算成本过滤掉简单负样本。为确保 f_l 模型在整个难度谱上具有良好的泛化能力，我们使用多种负样本的混合数据对其进行训练：召回级（GN）、预排序（PRN）、排序（RN）和曝光负样本（EN）。遵循 GHCL，我们针对每种负样本类型分别计算独立的 InfoNCE 损失函数$t \in \{EN, RN, PRN, GN\}$：

$$\mathcal{L}_t = -\log \frac{e^{f(\mathbf{x}^+, \mathbf{u})}}{e^{f(\mathbf{x}^+, \mathbf{u})} + \sum_{j=1}^{n_t} e^{f(\mathbf{x}_{t,j}^-, \mathbf{u})}}, \tag{17}$$

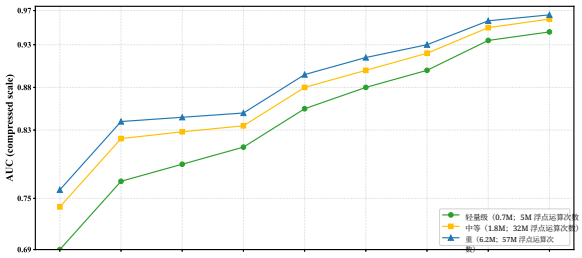


图3：模型曲线下面积与样本难度的关系，其中"已曝光"表示曝光负样本，"排序-N"/"预排序-N"表示排在位置 N之后的负样本； N 越小意味着难度越高。

其中 n_t 是类型为 t 的负样本数量。此阶段的总体损失函数为：

$$\mathcal{L}_{\text{light}} = \lambda_{\text{RN}}\mathcal{L}_{\text{RN}} + \lambda_{\text{EN}}\mathcal{L}_{\text{EN}} + \lambda_{\text{GN}}\mathcal{L}_{\text{GN}} + \lambda_{\text{PRN}}\mathcal{L}_{\text{PRN}} \tag{18}$$

该分层目标函数可防止对任何单一难度级别过拟合，并促进用于早期筛选的可泛化特征。

3.4.2 使用复杂模型进行细粒度排序。高容量模型 f_c 在上一阶段缩减后的候选集上执行细粒度排序。由于这些候选项主要是困难样本， f_c 通过仅使用最具挑战性的负样本进行训练来实现专门化：已曝光样本（EN）和排序样本（RN）。损失函数遵循与公式 17 相同的 InfoNCE 形式：

$$\mathcal{L}_{\text{complex}} = \lambda_{\text{RN}}\mathcal{L}_{\text{RN}} + \lambda_{\text{EN}}\mathcal{L}_{\text{EN}} \tag{19}$$

这一两阶段设计确保了资源的高效分配：简单样本由 f_l以较低成本处理，而 f_c的高昂计算容量则留给需要它的困难样本。这直接解决了开篇所指出的架构效率低下问题，在准确率与系统性能之间实现了最佳平衡。

3.4.3 最终损失函数。 InfoNCE至少需要每个请求包含一个正样本，因此没有点击的会话会被丢弃，从而造成显著的数据损失。为解决这一问题，我们引入了一个利用批次内所有样本的辅助二元交叉熵损失：

$$\mathcal{L}_{\text{global}} = \frac{1}{|\mathcal{B}|}\sum_{(x,y)\in\mathcal{B}} [-y\log\sigma(s(x)) - (1-y)\log(1-\sigma(s(x)))] \tag{20}$$

这里，B 是当前批次中所有样本的集合。 $y \in \{0,\ 1\}$ 是标签， $\sigma(\cdot)$ 是 Sigmoid 函数。我们联合优化上述三个损失函数。最终的训练目标为：

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{light}} + \mathcal{L}_{\text{complex}} + \alpha \cdot \mathcal{L}_{\text{global}} \tag{21}$$

这种统一优化将分层难度感知建模与全面监督相结合，在效率、鲁棒性和覆盖率之间取得平衡。

3.5 在线部署与样本工程

在本节中，我们介绍将 HAP框架部署到今日头条推荐系统中的实践经验。为使 HAP适应大规模在线环境，我们将预排序模块重构为级联架构，并构建实时列表式样本流水线，以实现高效的模型训练和

效果归因。整体系统设计如图 4 所示。

3.5.1 两阶段预排序架构。 今日头条传统的预排序模块采用单阶段设计，对召回的数千个候选项进行排序，并将排名靠前的几百个候选项送入精排序阶段。然而，这种单体式结构与 HAP 不兼容，因为 HAP 会针对不同难度的候选项采用不同模型（轻量级模型与表达能力强的模型）。为支持 HAP 并实现更优的效率—有效性权衡，我们将系统重新设计为两阶段级联。第一阶段采用特征维度较少、层结构简化的轻量级模型，过滤掉明显不相关的候选项，将候选集规模缩减至 G。门控阈值 G 通过经验确定，以平衡计算成本与模型性能，详见第 4.5.2 节。该阶段优先考虑推理速度，因为过高的复杂度相较于其计算成本只能带来微乎其微的收益。第二阶段采用表达能力强的模型，对剩余候选项进行精细筛选。由于该阶段自然聚焦于更难处理的样本，我们仅使用困难负样本训练模型。这种专门化增强了模型的判别能力，使其能够更好地区分细微的正负样本差异。总体而言，这种 HAP 两阶段实现显著提升了排序精度，且没有引入明显的延迟开销。

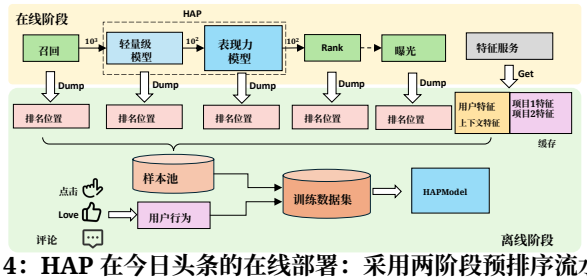


图 4：HAP 在今日头条的在线部署：采用两阶段预排序流水线，轻量级模型处理简单候选项，表达能力强的模型处理困难候选项；样本分组并共享通用特征以提升效率。

3.5.2 高效的数据收集与样本生成。 为进一步降低系统负载，我们通过全连接、列表式样本转储机制优化数据收集和采样流程。对于每个用户请求，系统会记录所有用户、物品和上下文特征，以及完整的候选列表。为尽量减少冗余，我们采用共享特征引用策略，每个请求的用户级特征仅存储一次，使每个请求的存储量减少 60% 以上。此外，我们定义了 40 分钟的反馈窗口，以捕获大多数用户交互。随后异步附加延迟行为，确保标签完整性和长尾覆盖。这一设计不仅提升了数据效率，还通过优化数据存储、减少冗余，确保模型训练更加稳定。

4 实验

4.1 ToutiaoRec数据集

在推荐系统中，数据对于评估模型至关重要。高性能算法的开发在很大程度上依赖高质量数据集的可用性。大多数开源数据集 [4, 26] 聚焦于推荐流水线的特定阶段，无法对单个用户请求提供全流程可观测性。

表 2：ToutiaoRec 与现有数据集的比较。

数据集	多阶段覆盖率	完整排序信息	多类型反馈	工业级-规模
MovieLens[7]	✗	✗	✗	✗
亚马逊[17]	✗	✗	✗	✗
Yelp[1]	✗	✗	✗	✗
淘宝[33]	✗	✗	✓	✓
Tenrec[26]	✗	✗	✓	✓
KuaiRec[4]	✗	✗	✓	✓
ToutiaoRec	✓	✓	✓	✓

为弥补这一不足，我们推出了 **ToutiaoRec**，这是一个大规模、公开可用的工业数据集，旨在支持推荐系统中的端到端优化。表 2 总结了 ToutiaoRec 与现有数据集的异同。与现有数据集不同，ToutiaoRec 提供全阶段覆盖和排序级标注，并严格遵守隐私保障要求。

ToutiaoRec 通过对今日头条信息流场景中 3 天内的 7000 万次用户请求进行采样构建而成，最终得到 3.13 亿次请求。该数据集包含真实用户反馈，如点击、点赞、评论等。对于每次请求，我们保留全部曝光物品，以及所有流水线阶段的排序信息。为处理大量未曝光数据，我们按如下方式采样负样本：1）全局随机负样本（GN）：从整个候选池中随机采样，与曝光正样本的比例为 1:2。2）排序负样本（RN）：排序阶段中排名低于第 50 位的 10 个未曝光候选项。3）预排序负样本（PRN）：粗排阶段中排名低于第 200 位的 10 个未曝光候选项。所有类型的负样本均保留其在相应阶段的排序位置。此外，该数据集包含 29 类匿名特征，例如用户行为和物品属性。鉴于今日头条的内容形式多样（文章、微博、短视频等），每个候选项均标注了其内容类型。为确保用户隐私与法律合规，我们遵循通用数据保护条例[2] 标准，采用严格的两步匿名化流程。ToutiaoRec 现已公开[3],，更多细节见其中说明。

4.2 实验设置

4.2.1 样本构建。由于现有公开数据集缺乏对完整流程的覆盖，因此不足以直接验证我们的方法。为此，我们在公开发布的 ToutiaoRec 数据集上开展实验。数据按时间顺序划分：按时间戳排列的前 80% 请求用于训练，其余 20% 用于评估，以反映真实的生产环境。训练期间，我们利用全部四种负样本：已曝光负样本（EN）、排序负样本（RN）、预排序负样本（PRN）和全局随机负样本（GN）。在评估中，我们模拟所提出的难度感知模型路由（DAMR）。负样本按难度分类并据此路由：困难负样本（EN 和 RN）由表达能力强的模型打分，而简单负样本（PRN 和 GN）则由轻量级模型处理。正样本始终定义为点击物品（EP），并默认纳入所有测试集。为严格评估模型在不同负样本类型上的性能，我们构建了仅包含负样本的分解测试集：TEN（EN）、TRN（RN）、TPRN（PRN）和 TGN（GN）。为进一步评估不同难度级别下的性能，我们定义了聚合集合：THard（EN + RN）表示困难负样本，TEasy（PRN + GN）表示简单负样本。这两种分组策略分别对应互补的分析目的：按负样本类型和按难度级别分析性能。后续所有实验均遵循此评估协议。

4.2.2 实现细节。我们使用 TensorFlow 实现 HAP，按请求组织样本，并将序列长度固定为 45，全局批大小设为 1350。模型基于 DeepFFM：轻量级模型有 3 个浅层，而表达能力强的模型则采用更深的架构，并通过注意力和按类别划分的塔式结构来捕获复杂模式。两个模型都应用梯度协调对比学习（GHCL），针对每种负样本类型分别计算 InfoNCE 损失，以缓解梯度冲突；此外还使用全局 BCE 损失，以处理没有正样本的请求。我们使用 RMSPropV2 优化器，初始学习率为 0.01。训练在 3000 个 CPU 核心上进行，共执行 1000 万步，耗时约 12 小时。为确保可复现性，代码见[4]。

4.2.3 评估指标。

- **曲线下面积** 量化模型在所有用户中区分正样本与负样本的整体能力。
- **点击率** 反映点击物品与曝光物品的比例，体现用户参与度以及整体推荐吸引力。
- **时长** 衡量用户在应用中停留的平均时间，反映模型促进长期参与的能力。
- **活跃天数** 表示用户保持活跃的平均天数，反映留存与持续兴趣。
- **延迟** 记录从接收请求到返回响应之间的时间，反映系统响应能力。
- **模型参数与浮点运算次数** 从可训练参数和浮点运算的角度体现模型复杂度。
- **成本比**报告离线训练和在线服务的总成本相较于此前最先进方法的相对变化。

4.2.4 对比方法。我们将 HAP 与多种具有代表性且最先进的预排序模型进行基准比较。为确保比较公平，所有基线模型的参数规模均设置为与我们的 HAP 网络相当（≈ 6.2M）。

- **DSSM**[8]：将用户和物品映射到低维语义空间，通过预计算嵌入实现高效召回。
- **COLD**[24] 由一个灵活的深度模型构成，支持跨特征交互和实时系统优化。
- **COPR**[30] 采用顺序对齐模块和分块采样，以提升跨阶段排序一致性。
- **HCCP**[28] 融合上游和下游信号，以减轻样本选择偏差并提升一致性。
- **HAP** 是我们的异质性感知自适应预排序模型，由用于简单候选项的轻量级模型（1.8M）和用于复杂候选项的表达能力强的模型（6.2M）组成。
- **HAP轻量版** 是 HAP 的一种高成本效益变体，在减少计算量的同时保持了竞争力强的效果，且模型配置更小（1.8M + 4.4M 个参数）。

[2]欧盟《通用数据保护条例》 [3] https://github.com/Toutiao-Rec/ToutiaoRec

[4]https://github.com/Toutiao-Rec/HAP

表3：各测试集上最先进预排序模型的离线 AUC 对比。最佳结果以粗体显示，次佳结果以下划线标出。

模型	TEN	TRN	TPRN	TGN	THard	TEasy
DSSM	0.6901	0.8247	0.9008	0.9785	0.7527	0.9257
COLD	0.6936	0.8303	0.9038	0.9803	0.7601	0.9261
COPR	0.7221	0.8533	0.9172	0.9871	0.7824	0.9402
HCCP	0.7244	0.8653	0.9218	0.9874	0.7917	0.9441
HAP	**0.7377**	**0.8693**	**0.9236**	**0.9878**	**0.8023**	**0.9468**
HAP_{lite}	0.7331	0.8666	0.9225	0.9872	0.7998	0.9452

模型规模：DSSM (6.2M), COLD (6.2M), COPR (6.2M), HCCP (6.2M), HAP (1.8M 6 2 1 8 + 4 4 . M),HAP_{lite} (. M . M).

4.3 离线性能对比

如表3所示，在不同负样本类型和难度层级下，HAP 的表现始终优于所有最先进的预排序模型。在细分集合（TEN、TRN、TPRN、TGN）上，提升在困难负样本（曝光样本、排序样本）上最为明显，而在较容易的负样本（PRN、GN）上则有适度提升。对于聚合难度集合，性能随样本难度增加而提升，其中 THard 的提升幅度大于 TEasy。尽管 HAP 两阶段模型的参数总量超过单阶段基线，但表达能力强的模型仅处理一小部分困难候选项，因此整体浮点运算次数更低，服务延迟也相当。我们还报告了 **HAP_{lite}** 的结果，这是一种高成本效益变体，在较小预算下仍能保持鲁棒性和出色性能。采用其他指标进行的补充评估也得出了相同结论（见附录中的表 7）。

4.4 消融研究

*4.4.1 GHCL*的消融研究。梯度协调对比学习（GHCL）模块协调各负例类型之间的梯度贡献，从而提升训练稳定性与表示学习效果。我们首先从HAP中移除GHCL，并观察到所有测试集上的曲线下面积均出现明显下降，如图5所示，这凸显了其重要性。其次，我们研究GHCL如何应对混合多种负例类型所带来的挑战。我们在表4中比较了三种范式下训练的模型：在SingleNeg中使用各个负例类型的模型、在MultiNeg中采用朴素混合策略（使用下标 Mix）的模型，以及我们的GHCL。我们观察到三个关键规律：1）在单一负样本类型上训练时，模型在对应测试集上取得最佳性能，但会引发跷跷板效应，其他负例类型上的性能显著下降。2）不加区分地混合所有负样本，可在各类型上获得更均衡的结果，但仍不及GHCL。3）GHCL始终优于SingleNeg和MultiNeg，不仅提升了各个测试集的表现，也提升了简单负样本与困难负样本上的聚合集合表现。总体而言，这些结果验证了GHCL对性能至关重要，尤其是在多负样本场景中。

*4.4.2 DAMR*消融研究。我们进一步研究 DAMR 在平衡模型有效性与效率方面的作用。如附录中的表6所示，模型性能始终随容量增加而提升，揭示了准确率与计算成本之间明显的权衡。统一大模型虽然能显著提升困难样本上的表现，却会在简单样本上产生冗余计算。相比之下，HAP 借助 DAMR 仅将困难候选项路由至表达能力强的模型，在保持准确率的同时

表4：负样本采样策略比较。对于每种负样本类型（GN、PRN、RN、EN），SingleNeg评估BCE损失与组合目标 InfoNCE+BCE。MultiNeg则组合全部四种类型。最佳结果以粗体和下划线标示。

负样本类型	模型	TEN	TRN	TPRN	TGN	THard	TEasy
SingleNeg	BCE_{GN}	0.5930	0.6859	0.7717	**0.9992**	0.6391	0.8680
	二元交叉熵$_{PRN}$	0.6186	0.7732	**0.9616**	0.9406	0.6955	**0.9502**
	BCE_{RN}	0.6238	**0.9017**	0.8595	0.8861	0.7623	0.8728
	二元交叉熵$_{EN}$	**0.7591**	0.6815	0.7121	0.7815	0.7207	0.7484
	$InfoNCE_{GN}$	0.5941	0.6834	0.7677	0.9991	0.6378	0.8634
	$InfoNCE_{PRN}$	0.6148	0.7564	0.9531	0.9348	0.6845	0.9487
	$InfoNCE_{RN}$	0.6417	0.8973	0.8556	0.8749	**0.7677**	0.8618
	$InfoNCE_{EN}$	0.7577	0.6851	0.7167	0.7788	0.7213	0.7357
MultiNeg	二元交叉熵$_{Mix}$	0.7232	0.8626	0.9173	0.9826	0.7875	0.9371
	$InfoNCE_{Mix}$	0.7234	0.8634	0.9233	0.9872	0.7882	0.9430
	GHCL	**0.7261**	**0.8662**	**0.9295**	**0.9898**	**0.7938**	**0.9488**

提升效率。为单独考察 DAMR 的作用，我们将 HAP 与统一大模型进行比较。该模型与 HAP 的表达能力强的组件采用相同架构，但以单阶段方式运行。如表5所示，尽管 HAP 的参数略多，其服务成本却降低了约 6%，延迟相当甚至更低。值得注意的是，在 THard数据集上，HAP 的表现甚至优于统一大模型。这是因为 DAMR 具有专业化能力：表达能力强的模型仅针对……进行训练和服务

在困难负样本上，缓解优化冲突，并实现专家级判别能力。这些优势也在在线环境中得到了进一步验证：与统一大模型相比，HAP 的点击率更高、延迟更低。在离线和在线环境中均保持一致的改进，凸显了 DAMR 在大规模推荐系统中实现性能与可部署性更优平衡方面的有效性。

4.5 HAP 组件的深入分析

4.5.1 梯度协调分析。我们从梯度层面进行分析，以实证验证 GHCL 在缓解不同难度样本间梯度冲突方面的有效性。具体而言，我们跟踪以下两项指标：1）不同负样本集的梯度之间的余弦相似度；2）每个集合内的梯度范数。如图 6 所示，GHCL 保持了较高的集合间余弦相似度，表明梯度方向更加一致，并能稳定简单样本和困难样本的梯度幅值。相比之下，随着训练的进行，BCE_{Mix} 和 $InfoNCE_{Mix}$ 的角度差异不断增大，梯度强度也愈发失衡。这些发现证实，GHCL 能有效协调梯度优化，避免破坏性干扰，并促进异构负样本上的稳定收敛。

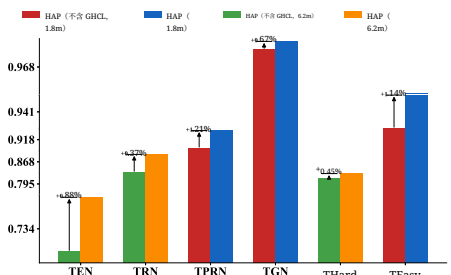
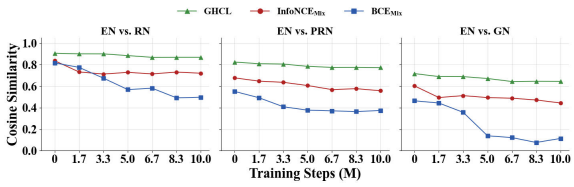


图5：HAP中GHCL的消融实验：移除GHCL会导致所有测试集上的曲线下面积显著下降。

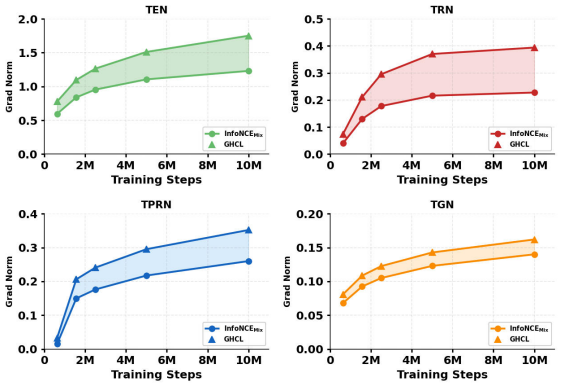
表5：HAP中DAMR的消融实验：在今日头条上基于大规模日志训练的在线A/B测试结果。

模型	THard[a]	TEasy[a]	CTR[c]	计算成本[b]	浮点运算次数（10^6）	延迟（*ms*）	成本比[c]
统一大模型	0.8158	0.9586	+0.55%	38.1	70	225	+15%
HAP	0.8197	0.9579	+3.22%	39.2	89	203	-6%

a 全规模工业数据评估；b 训练时长 × CPU 千核（h·千核）；c 相对于此前最先进方法的相对增益。



（a）集合间梯度余弦相似度。



（b）集合内梯度范数。

图 6：异构负样本的梯度统计。与现有方法相比，GHCL 稳定了梯度的方向和幅度，从而缓解了冲突。

4.5.2 路由策略分析。我们进一步分析路由策略，该策略决定在两阶段服务系统中有多少候选项从轻量级模型传递给表达能力强的模型。离线评估基于静态测试集进行，而在线部署的动态性更强，对延迟约束也更为敏感。路由候选项的数量会显著影响服务成本（延迟）和模型有效性。为量化这一权衡，我们通过改变门控阈值 G进行在线 A/B 测试，该阈值决定由表达能力强的模型评估的候选项数量。如图 7 所示，点击率起初会随着路由比例的提高而上升，但当系统负载超出承载能力后便会下降，这主要是由于请求失败，以及表达能力强的模型仅使用困难样本进行训练。这条权衡曲线为大规模工业系统在给定服务约束下选择最优门控阈值提供了有据可依的方法。

4.6 在线部署与性能

为进一步评估 HAP 在真实场景中的适用性，我们将其部署在今日头条的生产环境中。在线评估使用了大规模推荐日志，其中包含用户交互以及覆盖所有阶段的完整请求跟踪记录，包括召回、预排序、排序和重排序，这些数据收集自 60 天的时间跨度。我们训练了两个模型：一个轻量级模型（3.1M）和一个表达能力强的

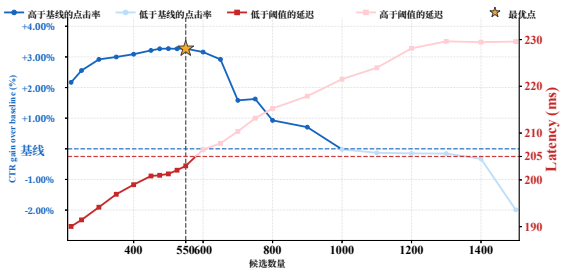


图 7：HAP 在不同门控阈值下的在线性能–效率权衡。

模型 基于我们的路由分析，门控阈值被设为550，这意味着仅轻量级模型中排名前30%的候选项会由表达能力更强的模型进行重排序。我们针对先前的生产环境最先进模型开展了在线A/B测试。结果表明，HAP在用户活跃天数上取得了 **+0.05%** 的提升，在应用使用时长上取得了 **+0.4%** 的提升，在点击率上取得了 **+3.0%** 的提升。考虑到今日头条服务数亿用户，即便用户活跃度提升0.01%也具有实际意义，因此所观察到的0.05%改进尤为值得关注。此外，HAP在保持与最先进模型相当的服务延迟的同时，将CPU 使用率降低了6%，证明其能够在大规模生产环境中以更低的计算成本实现更高的有效性。在取得这些积极成果之后，HAP已在过去9个月中全面部署于今日头条生产系统。

5 结论

在本文中，我们提出了HAP，这是一种用于应对预排序中候选项异质性的新颖框架。与受梯度冲突和计算低效制约的传统方法不同，HAP集成了用于稳定优化的梯度协调对比学习（GHCL）以及用于自适应资源分配的难度感知模型路由（DAMR）。在我们新发布的公开数据集上的大量实验，以及在今日头条的大规模在线部署表明，HAP在降低服务成本的同时持续提升推荐性能。这些结果确立了HAP作为一种实用且可扩展的解决方案，可用于构建更有效、更高效的工业推荐系统。

参考文献

[1] Nabiha Asghar。2016。Yelp 数据集挑战赛：评论评分预测。*arXiv* 预印本 *arXiv:1605.05362*（2016）。[2] Jiawei Chen、Hande Dong、Xiang Wang、Fuli Feng、Meng Wang 和 Xiangnan He。2023。推荐系统中的偏差与去偏：综述与未来方向。*ACM* 信息系统汇刊 41, 3, 67 2023 39 10 1145 3564284 文章 [3] 2016（2月），页码。doi：。/ Paul Covington、Jay Adams 和 Emre Sargin。深度神经网络用于 YouTube 推荐。载于第*10*届 *ACM* 推荐系统会议论文集。191–198。[4] Chongming Gao、Shijun Li、Wenqiang Lei、Jiawei Chen、Biao Li、Peng Jiang、Xiangnan He、Jiaxin Mao 和 Tat-Seng Chua。2022。KuaiRec：用于评估推荐系统的全观测数据集与见解。载于第*31*届 *ACM* 信息与知识管理国际会议（*CIKM '22*）论文集。ACM，540–550。doi:10.1145/3511808.3557220

[5] Chen Gao、Yu Zheng、Wenjie Wang、Fuli Feng、Xiangnan He 和 Yong Li。2024。推荐系统中的因果推断：综述与未来方向。*ACM* 信息系统汇刊 42，4（2024），1–32。[6] Feng Gao、Xin Zhou、Yinning Shao、Yue Wu、Jiahua Gao、Yujian Ren、Fengyang Qi、Ruochen Deng 和 Jie Liu。2025。兼顾供给与精度：大规模预排序系统的样本去偏与排序一致性联合学习。*AAAI*人工智能会议论文集 39，11（2025年4月），11672–11680。doi:10.1609/aaai.v39i11.33270[7] F Maxwell Harper 和 Joseph A Konstan。2015。MovieLens数据集：历史与背景。*ACM* 交互智能系统汇刊（*TIIS*）5，4（2015），1–19。[8] Po-Sen Huang、Xiaodong He、Jianfeng Gao、Li Deng、Alex Acero 和 Larry Heck。2013。利用点击数据学习用于网页搜索的深度结构化语义模型。载于第*22*届 *ACM* 信息与知识管理国际会议论文集。2333–2338。[9] SeongKu Kang、Junyoung Hwang、Wonbin Kweon 和 Hwanjo Yu。2021。用于推荐系统的物品侧排序正则化蒸馏。信息科学 580，C（2021年11月），15–34。doi:10.1016/j.ins.2021.08.060[10] Xiangyang Li、Bo Chen、HuiFeng Guo、Jingjie Li、Chenxu Zhu、Xiang Long、Sujian Li、Yichao Wang、Wei Guo、Longxia Mao 等。2022。IntTower：面向预排序系统的下一代双塔模型。载于第*31*届 *ACM* 信息与知识管理国际会议论文集。3292–3301。[11] Xiang Li、Xiaojiang Zhou、Yao Xiao、Peihao Huang、Dayao Chen、Sheng Chen 和 Yunsen Xian。2022。AutoFAS：面向预排序系统的自动特征与架构选择。载于第*28*届 *ACM SIGKDD*知识发现与数据挖掘会议论文集。3241–3249。[12] Xiao Lin、Xiaokai Chen、Chenyang Wang、Hantao Shu、Linfeng Song、Biao Li 和 Peng Jiang。2024。用于推荐重排序的离散条件扩散。载于*2024*年 *ACM*万维网会议附属论文集。161–169。[13] Shichen Liu、Fei Xiao、Wenwu Ou 和 Luo Si。2017。面向运营型电商搜索的级联排序。载于第*23*届 *ACM SIGKDD*知识发现与数据挖掘国际会议论文集。1557–1565。[14] Weiwen Liu、Yunjia Xi、Jiarui Qin、Fei Sun、Bo Chen、Weinan Zhang、Rui Zhang 和 Ruiming Tang。2022。多阶段推荐系统中的神经重排序：综述。*arXiv* 预印本 *arXiv:2202.06602*（2022）。[15] Jiaqi Ma、Zhe Zhao、Xinyang Yi、Ji Yang、Minmin Chen、Jiaxi Tang、Lichan Hong 和 Ed H Chi。2020。两阶段推荐系统中的离策略学习。载于*2020*年万维网会议论文集。463–473。[16] Xu Ma、Pengjie Wang、Hui Zhao、Shaoguo Liu、Chuhan Zhao、Wei Lin、Kuang-Chih Lee、Jian Xu 和 Bo Zheng。2021。探索预排序中有效性与效率之间的更佳权衡：一种基于可学习特征选择的方法。载于第*44*届国际*ACM SIGIR*信息检索研究与发展会议论文集。2036–2040。[17] Jianmo Ni、Jiacheng Li 和 Julian McAuley。2019。利用远程标注评论和细粒度方面解释推荐。载于*2019*年自然语言处理实证方法会议与第*9*届自然语言处理国际联合会议（*EMNLP-IJCNLP*）论文集。188–197。[18] Hamid Palangi、Li Deng、Yelong Shen、Jianfeng Gao、Xiaodong He、Jianshu Chen、Xinying Song 和 R Ward。2014。用于信息检索的长短期记忆语义建模。*arXiv* 预印本 *arXiv:1412.6629*（2014）。[19] Qi Pi、Weijie Bian、Guorui Zhou、Xiaoqiang Zhu 和 Kun Gai。2019。点击率预测中的长序列用户行为建模实践。载于第*25*届 *ACM SIGKDD*知识发现国际会议论文集

知识发现与数据挖掘. 2671–2679.[20] Pi Qi, Xiaoqiang Zhu, Guorui Zhou, Yujing Zhang, Zhe Wang, Lejian Ren, Ying Fan, and Kun Gai. 2020. 基于搜索的用户兴趣建模：利用终身序列行为数据进行点击率预测。arXiv:2006.05639 [cs.IR] https://arxiv.org/abs/2006.05639[21] Jiarui Qin, Jiachen Zhu, Bo Chen, Zhirong Liu, Weiwen Liu, Ruiming Tang, Rui Zhang, Yong Yu, and Weinan Zhang. 2022. Rankflow：将多阶段级联排序系统作为流进行联合优化。载于国际*ACM SIGIR*信息检索研究与发展会议论文集（第*45*届）. 814–824.[22] Jinbo Song, Ruoran Huang, Xinyang Wang, Wei Huang, Qian Yu, Mingming Chen, Yafei Yao, Chaosheng Fan, Changping Peng, Zhangang Lin, et al. 2022. 重新思考大规模预排序系统：全链路跨域模型。载于第*31*届 *ACM* 信息与知识管理国际会议论文集. 4495–4499.[23] Fan Wang, Xiaomin Fang, Lihang Liu, Yaxue Chen, Jiucheng Tao, Zhiming Peng, Cihang Jin, and Hao Tian. 2019. 组合推荐系统的序列评估与生成框架。*arXiv* 预印本 *arXiv:1902.00245* (2019年)。 [24] Zhe Wang, Liqin Zhao, Biye Jiang, Guorui Zhou, Xiaoqiang Zhu, and Kun Gai. 2020. COLD：迈向下一代预排序系统。*arXiv* 预印本 *arXiv:2007.16122* (2020年)。[25] Chen Xu, Quan Li, Junfeng Ge, Jinyang Gao, Xiaoyong Yang, Changhua Pei, Fei Sun, Jian Wu, Hanxiao Sun, and Wenwu Ou. 2020. 淘宝推荐中的特权特征蒸馏。载于第*26*届 *ACM SIGKDD*知识发现与数据挖掘国际会议论文集. 2590–2598.[26] Guanghu Yuan, Fajie Yuan, Yudong Li, Beibei Kong, Shujie Li, Lei Chen, Min Yang, Chenyun Yu, Bo Hu, Zang Li, et al. 2022. Tenrec：面向推荐系统的大规模多用途基准数据集。神经信息处理系统进展35 (2022年), 11480–11493.[27] Zhixuan Zhang, Yuheng Huang, Dan Ou, Sen Li, Longbin Li, Qingwen Liu, and Xiaoyi Zeng. 2023. 重新思考预排序在大规模电子商务搜索系统中的作用。*arXiv* 预印本 *arXiv:2305.13647* (2023年)。[28] Binglei Zhao, Houying Qi, Guang Xu, Mian Ma, Xiwei Zhao, Feng Mei, Sulong Xu, and Jinghe Hu. 2025. 面向在线推荐系统的混合式跨阶段协同预排序模型。载于*ACM Web Conference* 伴随论文集 *2025* *25* 621 630 10 1145 3701716 3715208 *(WWW '*)。[29] ACM，–。doi:。/。Qihang Zhao, Rui-jie Zhu, Liu Yang, He Yongming, Bo Zhou, and Luo Cheng. 2023. 效率与有效性兼得！搜索系统中的大规模预排序框架。*arXiv* 预印本 *arXiv:2304.02434* (2023年)。 [30] Zhishan Zhao, Jingyue Gao, Yu Zhang, Shuguang Han, Siyuan Lou, Xiang-Rong Sheng, Zhe Wang, Han Zhu, Yuning Jiang, Jian Xu, et al. 2023. COPR：面向在线广告的一致性导向预排序。载于第*32*届 *ACM* 信息与知识管理国际会议论文集. 4974–4980. [31] Chang Zhou, Jianxin Ma, Jianwei Zhang, Jingren Zhou, and Hongxia Yang. 2021. 大规模推荐系统中用于去偏候选生成的对比学习。载于第*27*届 *ACM SIGKDD*知识发现与数据挖掘会议论文集. 3985–3995.[32] Guorui Zhou, Weijie Bian, Kailun Wu, Lejian Ren, Qi Pi, Yujing Zhang, Can Xiao, Xiang-Rong Sheng, Na Mou, Xinchen Luo, et al. 2020. CAN：重新审视点击率预测中的特征交互。*arXiv* 预印本 *arXiv:2011.05625* (2020年)。 [33] Han Zhu, Xiang Li, Pengye Zhang, Guozheng Li, Jie He, Han Li, and Kun Gai. 2018. 面向推荐系统的基于树的深度模型学习。载于第*24*届 *ACM SIGKDD*知识发现与数据挖掘国际会议论文集. 1079–1088.

A 附录

A.1 模型复杂度与有效性权衡分析

如表6所示，随着模型复杂度提高，模型性能也随之提升，这体现了有效性与效率之间的权衡。不过，实验还表明，随着复杂度增加，统一大模型在困难样本上的提升幅度比在简单样本上更显著。相比之下，HAP 使用 DAMR 将计算精准分配给“困难候选项”，避免在“简单候选项”上浪费计算资源。因此，在不牺牲性能的情况下实现了更高的效率。

表6：不同复杂度模型的性能与效率权衡

（a）性能比较

指标	模型	TEN	TRN	TPRN	TGN	THard	TEasy
AUC	轻量级	0.7192	0.8617	0.9267	0.9886	0.7836	0.9471
AUC	中等	0.7219	0.8625	0.9270	0.9887	0.7867	0.9475
AUC	重	0.7261	0.8662	0.9295	0.9898	0.7938	0.9488

(b) 效率比较

模型	计算成本[a]	参数量	浮点运算次数 (10^6)
轻量级	28.5	1.8m	32
中等	33.9	4.4m	47
重	36.1	6.2m	57

a 训练时长 × CPU 千核（小时·千核）

A.2 多种离线指标评估

为了进行更全面的比较，我们使用曲线下面积以外的其他指标评估了模型。具体而言，我们引入了组曲线下面积（分组曲线下面积）和对数损失。结果如表 7 所示。可以看出，分组曲线下面积和对数损失上的性能趋势与曲线下面积的结果完全一致。这表明，我们的 HAP 模型带来的提升稳健且显著。

表 7：HAP 与工业推荐基线的对比评估。最佳结果以粗体和下划线标示。带有 ↑ 的指标表示数值越高越好，而带有 ↓的指标表示数值越低越好。

	模型	TEN	TRN	TPRN	TGN	THard	TEasy
AUC↑	DSSM	0.6901	0.8247	0.9008	0.9785	0.7527	0.9257
	COLD	0.6936	0.8303	0.9038	0.9803	0.7601	0.9261
	COPR	0.7221	0.8533	0.9172	0.9871	0.7824	0.9402
	HCCP	0.7244	0.8653	0.9218	0.9874	0.7917	0.9441
	HAP	**0.7377**	**0.8693**	**0.9236**	**0.9878**	**0.8023**	**0.9468**
	HAP_{lite}	0.7331	0.8666	0.9225	0.9872	0.7998	0.9452
分组曲线下面积↑[a]	DSSM	0.6483	0.8229	0.9002	0.9794	0.7377	0.9277
	COLD	0.6499	0.8245	0.9021	0.9809	0.7393	0.9337
	COPR	0.6617	0.8389	0.9112	0.9875	0.7514	0.9404
	HCCP	0.6659	0.8545	0.9184	0.9891	0.7602	0.9449
	HAP	**0.6783**	**0.8627**	**0.9203**	**0.9897**	**0.7691**	**0.9473**
	HAP_{lite}	0.6736	0.8565	0.9194	0.9887	0.7664	0.9457
对数损失↓[b]	DSSM	0.0489	0.0422	0.0381	0.0325	0.0453	0.0353
	COLD	0.0482	0.0420	0.0378	0.0323	0.0451	0.0351
	COPR	0.0473	0.0409	0.0372	0.0322	0.0447	0.0341
	HCCP	0.0471	0.0403	0.0366	0.0319	0.0438	0.0337
	HAP	**0.0459**	**0.0385**	**0.0359**	**0.0315**	**0.0434**	**0.0332**
	HAP_{lite}	0.0461	0.0398	0.0362	0.0321	0.0435	0.0335

a 分组曲线下面积在用户层面计算曲线下面积，反映每位用户情境下的个性化排序效果。b 对数损失量化了预测概率与真实值之间的差距。数值越低，准确率越高。