arXiv:2603.00980v1 [cs.IR] 1 Mar 2026

超越扁平序列：分层且偏好感知的生成式推荐

Zerui Chen*
zrchen@ir.hit.edu.cn
哈尔滨工业大学
技术
中国哈尔滨

Heng Chang*
changh.heng@gmail.com
华为技术有限公司，
Ltd.
中国北京

Tianying Liu
tianying_liu@outlook.com 华为技术有限公司 中国上海

Chuantian Zhou
alita@bupt.edu.cn
北京邮电大学
与电信
中国北京

Yi Cao
caoyi23@huawei.com
华为技术有限公司，
Ltd.
中国上海

Jiandong Ding
dingjiandong2@huawei.com
华为技术有限公司，
Ltd.
中国上海

刘铭†
mliu@ir.hit.edu.cn
哈尔滨工业大学
技术
中国哈尔滨

Bing Qin
qinb@ir.hit.edu.cn
哈尔滨工业大学
技术
中国哈尔滨

摘要

以分层序列转导单元（HSTU）为代表的生成式推荐器（生成式推荐系统）已成为建模长用户交互序列的强大范式。然而，我们发现，其“扁平序列”假设忽视了用户行为丰富且内在的结构。这带来了两个关键局限：一是无法捕捉基于会话的参与行为中的时间层次结构；二是计算效率低下，因为密集注意力会引入大量噪声，掩盖语义稀疏历史中的真实偏好信号，从而降低学习得到的表示的质量。为此，我们提出了一种名为 **HPGR**（**H**分层**P**偏好感知**G**生成式**R**推荐器）的新框架。该框架建立在两阶段范式之上，将这些关键结构先验注入模型，以克服上述缺陷。具体而言，HPGR 包含两个协同阶段。首先，结构感知预训练阶段采用基于会话的掩码项目建模（MIM）目标，学习具有层次信息且语义丰富的物品表示空间。其次，偏好感知微调阶段利用这些强大的表示，实现偏好引导稀疏注意力机制，动态地将计算限制在最相关的历史物品上，从而同时提升效率和信噪比。在来自华为应用市场的大规模专有工业数据集上进行的实证实验以及在线 A/B 测试验证了，HPGR 在多个强大的基线模型（包括 HSTU 和 MTGR）中取得了最先进的性能。

关键词

注意力机制，生成式推荐

ACM 参考文献格式：

Zerui Chen、Heng Chang、Tianying Liu、Chuantian Zhou、Yi Cao、Jiandong Ding、MingLiu 和 Bing Qin。2026 年。超越扁平序列：分层且偏好感知的生成式推荐。载于*2026* 年 *ACM Web* 会议论文集（*WWW '26*），*2026*年*4*月*13*日至*17*日，迪拜，阿拉伯联合酋长国。ACM，纽约，纽约州，美国，9 页。https://doi.org/10.1145/ 3774904.3792790

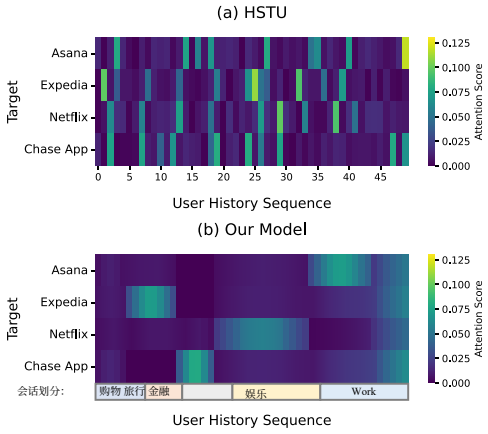

图 1：注意力机制对比，展示不同模型如何解读用户历史。(a) 基线 HSTU 模型将历史视为“扁平序列”，其分散的项目级注意力无法捕捉用户意图的内在结构。(b) 相比之下，HPGR 学会理解用户行为的层次化、基于会话的特性。

CCS 概念

• **信息系统** →**推荐系统**。

*两位作者对本研究的贡献相同。
†通讯作者。

本作品采用知识共享署名 4.0 国际许可协议授权。*WWW '26*，迪拜，阿拉伯联合酋长国© 2026 版权归所有者/作者所有。ACM ISBN 979-8-4007-2307-0/2026/04 https://doi.org/10.1145/3774904.3792790

1 引言

推荐系统是现代互联网平台的基石，每天为数十亿用户提供个性化体验。

近年来，该领域见证了从传统深度学习推荐模型（DLRMs）[5, 7, 12,

26, 29, 37, 38] 向生成式推荐系统（GRs） [35]转变的范式变革。作为这一变革的引领者，分层序列转导单元（HSTU） [35] 取得了巨大成功，将推荐模型扩展至数万亿参数，并展现出与大语言模型（LLMs） [1, 2, 22, 27]相似、前景广阔的缩放定律。

通过将推荐重新定义为针对用户长历史的序列转导任务，HSTU 有效克服了深度学习推荐模型的可扩展性瓶颈，并将异构特征的建模统一到单一且强大的架构中。

尽管 HSTU 在处理长序列方面能力出众，其成功却建立在一个根本但隐含的假设之上：它将用户历史记录建模为一连串扁平、未加区分的行为。尽管这种方法行之有效，“扁平序列”视角却忽略了人类行为内在丰富的结构，由此带来两项重大挑战。

首先，它忽视了用户参与行为的分层时间结构。用户活动并非均匀分布，而是自然地聚集为多个会话，每个会话都代表一种明确的短期意图。扁平模型难以区分连贯的会话内模式与更广泛的会话间兴趣变化 [33]。

其次，它忽视了用户偏好的稀疏性：用户的下一步行为取决于过往交互中紧凑且高度相关的子集，而非完整的历史记录。强迫模型对数千个历史物品计算密集注意力，不仅会造成计算冗余，还会引入显著噪声，进而损害学习得到的表示质量 [4]。

为了解决这些局限，我们提出 HPGR，这是一种将关键结构先验融入生成式推荐过程的新框架。我们的核心洞见是，通过显式建模用户行为的时间层次结构与偏好，我们能够构建效果更好、效率更高的推荐系统。

具体而言，我们提出一种两阶段范式。在第一阶段，即结构感知预训练阶段，我们引入会话增强模块（SEM），并采用基于会话的掩码项目建模（MIM）目标。SEM 通过处理会话内及会话间的交互，显式建模用户行为的时间层次结构。这一过程促使模型学习可泛化的物品语义和复杂的行为动态，从而形成强大的预训练嵌入空间。

在第二个阶段，即偏好感知微调阶段，我们针对下游预测任务调整预训练模型。这是通过我们提出的偏好引导稀疏注意力（PGSA）机制实现的，该机制旨在为给定的候选物品重构最显著的偏好上下文。PGSA 不会对全部历史记录应用密集注意力，而是将候选物品视为查询，从过往交互中检索一个规模小且高度相关的子集。随后，注意力仅作用于这个偏好丰富子集，从而大幅减少计算量，并通过增强真实的偏好信号、抑制历史噪声来提升准确率。

我们的工作主要有以下贡献：

- 我们提出了会话增强模块（SEM）、一种分层 Transformer 架构，以及与之相应的结构感知预训练策略。这种组合方法使模型能够显式捕捉会话内与会话间动态，克服以往生成式推荐器存在的关键“扁平序列”局限。
- 我们为微调阶段提出了偏好引导稀疏注意力（PGSA）机制。PGSA 以偏好驱动的稀疏性取代了以往的随机采样技术。该方法通过聚焦于最相关的信号来提高注意力的准确性，同时对于较长的用户序列，其计算效率也天然高于密集注意力。
- 我们在一个大规模工业级推荐系统上开展了广泛的离线实验和在线 A/B 测试。结果验证了 HPGR 显著优于最先进的基线模型，并在真实生产环境中实现了显著的 +1.99 % eCPM 提升，展现了其实用效果和商业价值。

2 相关工作

2.1 序列式推荐系统与生成式推荐器

传统的序列式推荐系统旨在从交互序列中捕捉用户动态，其发展历程从 GRU4Rec [15] 等基于 RNN 的模型，演进到 SASRec [17] 等基于注意力的架构。在大规模工业系统中，这些方法通常以更大型 DLRM 框架中的目标注意力模块的形式实现。

受大语言模型 [1, 2, 22, 27] 成功的启发，近年来出现了向生成式推荐器转变的范式。生成式推荐系统不再将序列视为众多特征来源之一，而是将整个推荐任务重新表述为基于统一用户行为流的序列到序列问题。HSTU [35] 的开创性工作表明，这种方法可以扩展至数万亿参数，并展现出良好的缩放定律。后续工作（如 MTGR [13] ）进一步调整了这一架构，以更好地融入 DLRM 中传统使用的丰富交叉特征。我们的工作直接建立在这一强大的生成式推荐范式之上。然而，现有生成式推荐系统将用户历史视为“扁平序列”，而我们认为，融入显式结构先验至关重要；这正是我们的 HPGR 框架所解决的关键区别。

2.2 用户序列中的结构建模

推荐领域的文献已充分确立这样一个观点：用户行为并非单一整体，而是由多个意图会话构成，尤其是在基于会话的推荐中 [14, 30]。这类研究通常侧重于预测短期、正在进行的会话中的下一步行为。然而，在生成式推荐范式倡导的超长用户历史建模背景下，显式建模多个相互关联会话的层次结构仍是一个探索较少的领域 [18, 20]。大多数长序列模型要么将序列视为扁平结构 [17, 25] ，要么使用位置偏置等隐式机制 [28] 来区分近期物品与较早的物品。为填补这一空白，我们引入会话增强模块（SEM），这是一种分层 Transformer 架构，旨在显式建模长用户序列中的会话内与会话间动态，从而为预测提供更丰富、具备结构感知能力的上下文。

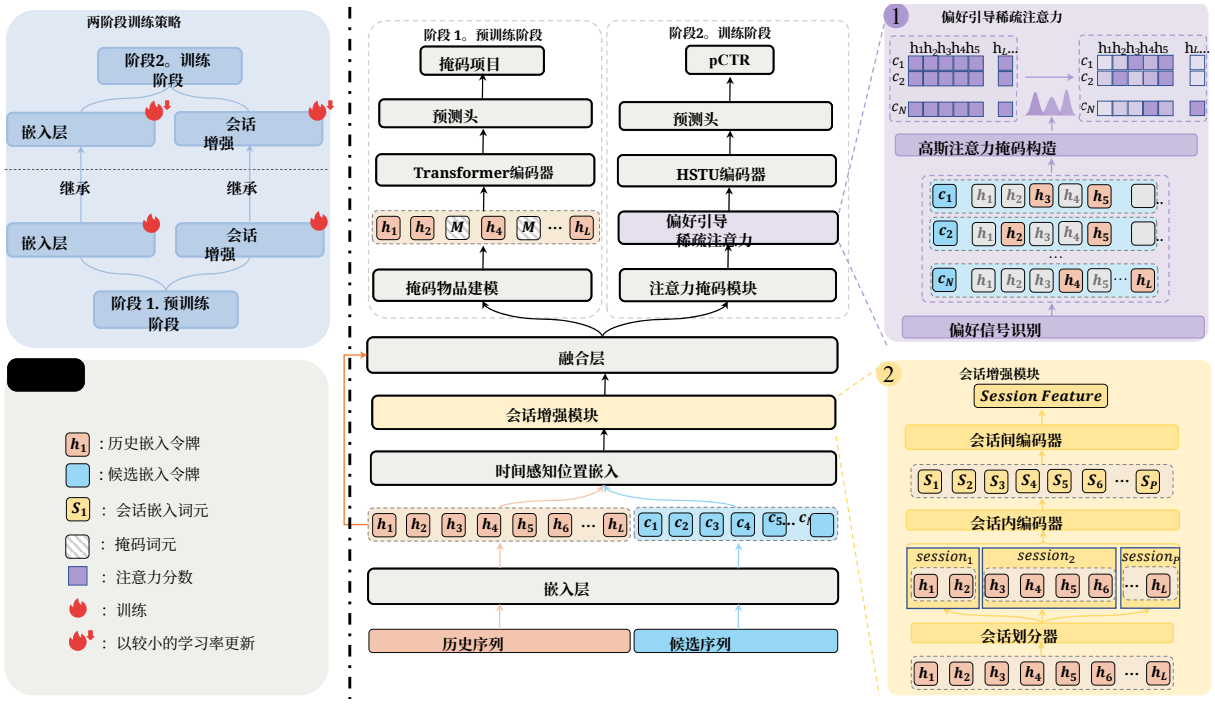

图 2：所提出的 HPGR 框架的整体架构。该模型采用两阶段训练范式。中央流程展示了共享基础架构，包括提供层次化上下文的会话增强模块（SEM）。随后，架构分为两个不同阶段：（1）阶段 1（预训练），使用 Transformer编码器和掩码物品建模目标，学习具备结构感知能力的物品表示；（2）阶段2（微调），采用 HSTU编码器执行下游点击率预测任务，并通过我们提出的偏好引导稀疏注意力（PGSA）机制加以增强。右侧面板详细展示了 SEM 的层次结构和 PGSA 的运行步骤。左上方面板突出显示了权重继承策略：在微调过程中，以较小的学习率更新预训练模块。

2.3 长序列的高效注意力

自注意力的二次复杂度仍然是长序列建模的一大瓶颈 [28, 40]。大量研究致力于通过低秩近似 [31] 等方法，或 FlashAttention [8, 9, 24]等面向硬件的实现，提升注意力效率。另一类工作则在注意力矩阵中引入稀疏性 [3, 6]。例如，原始 HSTU 模型提出了一种受随机深度启发的随机长度缩减方法 [16, 35]。

尽管这些方法行之有效，但它们通常与内容无关，或依赖于随机采样。相比之下，我们的PGSA提供了一种更有原则、具备内容感知能力的方法。其灵感源于如下观察：对于给定的候选物品，用户历史记录中只有一小部分真正相关。通过利用学习得到的表示，为每次预测动态识别这一偏好丰富子集，我们的方法引入了强大的归纳偏置，从而形成一种由任务本身引导的有信息引导的稀疏性，这与以往随机或固定的稀疏模式有着根本区别。

2.4 面向推荐的预训练

由 BERT[10] 和 GPT [1], 等模型普及的预训练、微调范式，也在推荐领域 [19, 23, 34]中日益受到关注。一种显著趋势是利用预训练大语言模型丰富的世界知识和上下文学习能力来提升推荐效果 [11, 32]。另一种趋势则侧重于根据用户的行为序列预训练用户表征，随后将其迁移至各种下游任务 [39]。

我们的工作与后一种趋势相一致，但提出了一种针对我们目标定制的新颖预训练目标。我们的结构感知预训练阶段并非学习通用的用户表示，而是利用我们新颖的SEM来显式建模行为层级。随后，该架构通过基于会话的掩码项目建模目标 [25]进行训练，以同时共同学习丰富的物品语义以及用户层级行为的动态。这构建了强大的基础，明确为下游模型做好准备，而该下游模型同时依赖于语义理解与层次化上下文。

3 方法

我们提出的框架 HPGR 通过显式建模用户行为的内在结构，增强了生成式推荐范式。该框架由两阶段训练框架组成：（1）结构感知预训练阶段，用于学习通用的、融合层次信息的物品表示，

以及（2）偏好感知微调阶段，通过高效的偏好引导稀疏注意力（PGSA）机制，使模型适应特定的下游任务。该机制利用预训练知识，聚焦于用户历史记录中最相关的偏好信号。图 2 展示了我们框架的整体架构。

3.1 阶段 1：结构感知预训练

预训练阶段的主要目标有两个：首先，学习一个能够捕捉丰富语义关联的高质量物品嵌入空间；其次，训练一个强大的会话增强模块（SEM），使其能够理解用户兴趣的层次结构。这两个组件均以自监督方式学习，不依赖于任何特定的下游任务，从而为微调奠定稳健的基础。

3.1.1 会话增强模块。为突破"平铺序列"的假设，我们引入一个以层次化方式建模用户历史行为的 SEM。给定用户的长期交互历史 $H = \{h_1, h_2, \ldots\ldots, h_N\}$，我们首先根据时间间隔不活跃情况，将其划分为一系列会话 $S = \{S_1, S_2, \ldots\ldots, S_M\}$。随后，该模块采用双层 Transformer [28] 架构：

- **会话内 Transformer**：对每个会话 S_i 分别应用共享的 Transformer编码器，以建模连贯交互上下文中物品之间的局部短期依赖关系。随后，将 [CLS] 标记的输出表示作为该会话的摘要向量 s_i。

- **会话间 Transformer**：将会话摘要向量序列 $[s_1, s_2, \ldots, s_M]$ 输入第二个更高层级的 Transformer编码器，以建模用户兴趣在不同会话间的长期演变与转移。

形式化地，对于每个会话 S_i，会话内 Transformer 会生成一个摘要向量 s_i：

$$s_i = \text{Intra-Transformer}(\text{Emb}(S_i))_{\texttt{[CLS]}} \tag{1}$$

其中，$\text{Emb}(S_i)$ 表示会话 S_i 的嵌入物品序列，$(\cdot)_{\texttt{[CLS]}}$ 表示选取 [CLS] 标记的输出表示。随后，会话间 Transformer 会处理所有会话向量组成的序列，从而得到最终的全局会话上下文向量 F_{sess}：

$$F_{\text{sess}} = \text{Inter-Transformer}([s_1, s_2, \ldots, s_M])_{\text{GlobalPool}} \tag{2}$$

其中，$(\cdot)_{\text{GlobalPool}}$ 可以是另一个 [CLS] 标记的输出，或是均值对所有会话表示执行池化操作。随后，通过将上下文向量 F_{sess} 以广播方式加到整个历史序列的初始词元嵌入中，将其整合到物品表示中：$H_{\text{enhanced_emb}} = H_{\text{emb}} + F_{\text{sess}}$。

3.1.2 预训练目标。我们采用掩码物品建模目标进行预训练 [25]。对于给定用户，设 H_m 为其会话增强历史记录 H_{enhanced}中的掩码物品索引集合。预训练目标是最小化对正确掩码物品进行预测的负对数似然，公式如下：

$$\mathcal{L}_{\text{MIM}} = -\sum_{i \in H_m} \log P(h_i | H_{\text{unmasked}}) \tag{3}$$

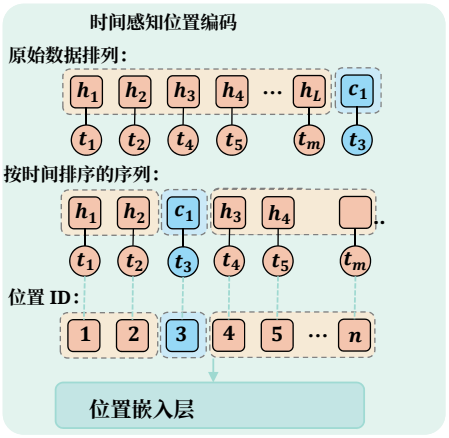

图 3：时间感知位置编码的工作流程。首先，将候选物品 c_1 插入用户历史序列中正确的时间顺序位置，形成完整的按时间排序的序列。随后，为该最终序列分配一组新的位置 ID，确保每个物品的位置都反映其精确的时间顺序，然后将其传递至嵌入层。

其中，$P(h_i | H_{\text{unmasked}})$ 表示在给定未掩码上下文的情况下预测原始物品 h_i 的概率，该概率通过对 Transformer 编码器输出应用 softmax 层计算得出。此任务促使模型学习物品与用户行为模式之间深层的语义关系，从而生成稳健且有意义的物品嵌入表。本阶段的输出为预训练物品嵌入 $\mathbf{E}^{\text{pre}}$ 和 SEM 的权重 Θ_{SEM}。

3.2 阶段 2：偏好感知微调

在偏好感知微调阶段，我们针对下游预测点击率任务调整模型。本阶段的核心思想是继承阶段 1 的预训练权重：物品嵌入层和 SEM 分别使用 $\mathbf{E}^{\text{pre}}$ 和 Θ_{SEM} 进行初始化。随后，通过添加以下组件来增强架构，以便进行微调。

3.2.1 时间感知位置编码。 为精确建模历史物品与候选物品C之间的时间关系，我们用一种时间感知策略取代标准位置编码，该策略会构建一组符合时间顺序的索引。如图 3 所示，该过程分为两步：首先，给定历史序列 $H = \{h_1, \ldots\ldots, h_N\}$ 和候选物品 C，我们根据其时间戳将 C 插入正确的位置，从而形成一个新的统一序列。其次，为这个按时间排序的序列分配一组连续的位置索引 $\{1, 2, \ldots\ldots, N+1\}$。候选物品 C 的具体位置索引正式定义如下：

$$\text{Index}(C) = 1 + \sum_{i=1}^{N} \mathbb{I}(t_i < t_c) \tag{4}$$

其中，t_i 是历史项目 h_i的时间戳，$\mathbb{I}(\cdot)$ 是指示函数：条件成立时取 1，否则取 0

。这样可以确保最终序列中的所有物品都有唯一且时间顺序正确的位置表示。

3.2.2 偏好引导稀疏注意力。 为克服在包含数千个物品的长历史记录上应用密集注意力时效率低下的问题，我们提出了偏好引导稀疏注意力（PGSA）。PGSA 为每个候选物品构建紧凑而强大的偏好上下文，并将建模能力集中在最显著的用户偏好信号上。具体而言，该方法分为以下三个步骤：

(i) 偏好信号识别：对于给定的候选物品 C，我们首先识别用户历史记录中的潜在偏好信号。具体来说，我们通过计算 C 与每个历史物品 h_i 的预训练嵌入点积，来计算它们之间的偏好对齐分数：$\text{score}_i = \text{emb}(C) \cdot \text{emb}(h_i)$。分数越高，表示偏好对齐程度越强。

(ii) 偏好子集选择：然后，我们选取得分最高的 Top-K 个历史物品。索引为 $\{p_1, p_2, \ldots, p_K\}$ 的选定物品构成一个稀疏子集，作为该候选物品的主要**用户偏好锚点**，超越单纯的语义相似度。

(iii) 高斯注意力掩码构造：我们不使用硬二元掩码，而是构造一种软注意力掩码，在每个偏好锚点周围施加高斯形状的影响。对于历史中的任意位置 j，最终掩码值 M_j 按公式（5）计算：

$$M_j = \max_{k \in \{1,\ldots,K\}} \left\{ \exp\left(-\frac{(j - p_k)^2}{2\sigma^2} \right) \right\} \tag{5}$$

其中，σ 是一个控制影响范围宽度的超参数。该掩码逐元素应用于 HSTU编码器中的注意力分数矩阵，使计算集中于偏好对齐程度较高的区域，同时保持梯度平滑，并有效滤除无关交互中的噪声。

会话增强词元嵌入及其时间感知位置嵌入和高斯掩码会一同输入主 HSTU编码器。随后，将得到的候选项表示传入多层感知机，以生成预测分数。

3.2.3 微调目标。在此阶段，整个 HPGR模型将进行端到端训练，以预测点击率。给定一个由用户—候选项对组成的训练批次，目标是最小化二元交叉熵损失，其定义如下：

$$\mathcal{L}_{\text{pCTR}} = -\frac{1}{B} \sum_{i=1}^{B} \left(y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \right) \tag{6}$$

其中，B是批次大小，$y_i \in \{0,1\}$是真实标签（点击或未点击），$\hat{y}_i$ 是由最终的多层感知机和 Sigmoid激活函数生成的第 i 个样本对的预测 CTR 分数。

3.3 训练策略

我们的两阶段方法兼顾通用知识获取与特定任务适配。首先进入结构感知预训练阶段，模型在大规模无标签用户交互数据语料上进行训练，以优化 L_{MIM} （式 3）。

随后，在偏好感知微调阶段，我们使用学得的权重初始化核心模块，并通过优化 L_{pCTR} （式 6），在带标签的下游任务上训练整个 HPGR模型。在第二阶段，与模型其余随机初始化的部分相比，我们对继承的权重（物品嵌入和 SEM）采用较小的学习率。这种差分学习率策略至关重要：它既能保留预训练期间学到的丰富通用语义空间，又能让模型有效适配下游任务的特定监督信号。

4 实验

4.1 实验设置

4.1.1 数据集。 为在真实世界的大规模工业环境中验证我们的方法，我们利用**华为应用市场** 应用分发平台生产环境中的推荐系统构建了一个专有数据集。该完整数据集涵盖连续一个月的用户交互日志。表1汇总了这一**完整的未采样数据集**的统计信息，彰显了其庞大的规模。

为便于开展包括超参数调优和消融研究在内的大量实验，我们的主要实验（见表2）基于一个具有代表性的子集进行。该子集通过**随机抽取 2.5% 的用户**及其完整交互历史构建。所有数据均经过严格匿名化，以保护用户隐私。

为确认从抽样数据中得出的结论能否推广至完整的数据分布，我们还在完整测试集上进行了最终验证，详情见第 4.3 节。

表格 1：华为应用市场数据集的统计信息。

指标	数值
用户数	~45.8 百万
物品数量	~200,000
交互（训练集）	~285 百万
交互（测试集）	~7.38 百万

4.1.2 基线模型。 为全面评估我们提出的模型 HPGR，我们将其与多种基线方法进行比较，并根据这些方法所基于的技术范式对其进行分类。

- **判别式推荐模型：**这类模型涵盖了多种将推荐建模为评分或分类任务的方法，范围从基础架构到当前高阶特征交互建模的最先进水平。

  – **深度神经网络 [7]**：一种基础的深度学习模型，使用多层感知机处理特征。– **MoE [21]**：深度神经网络的一种扩展，采用混合专家架构来应对不同的用户群体。– **GRU4Rec [15]**：一种经典模型，使用门控循环单元捕捉用户行为序列中的时间模式。

– **SASRec [17]**：一项开创性工作，首次引入自注意力，以有效捕捉用户序列中的长程依赖关系。– **BERT4Rec [25]**：一种受 BERT 启发的模型，采用深层双向自注意力网络进行序列推荐。– **悟空 [36]**：一种达到最先进水平的模型，通过"密集缩放"方法，利用堆叠的因子分解机捕捉高阶特征交互，并为推荐确立了缩放定律。

- **生成式推荐模型：**此类别代表了本工作所依托的新兴范式。这些模型将推荐重新表述为序列生成任务，在建模用户长历史方面展现出巨大潜力。

– **HSTU [35]**：生成式推荐领域的开创性工作，也是本研究的直接基础。– **MTGR [13]**：生成式推荐范式中的强大后继模型，进一步改进了架构，使其能够处理工业系统中的长序列。

对于我们提出的模型 **HPGR**，我们尝试了若干不同规模的配置，以验证其可扩展性，详见后续分析。

4.2 主要结果

为评估我们提出的框架 HPGR 的有效性，我们在 APPGallery 数据集上与所有基线模型进行了全面比较。表2按底层范式分组，展示了排序任务的主要结果。从表中可以得出以下几点关键观察：

(1) **判别式范式中的表现：** 代表判别式范式的第一组模型显示出明显趋势：显式建模用户序列是有益的。SASRec 和 GRU4Rec 等序列模型显著优于 DNN 等非序列模型。值得注意的是，采用自注意力机制的 SASRec 是整个范式中表现最强的模型，为该数据集上的判别式方法设定了性能上限。

(2) **生成式范式的优势：** 一项重要发现是，生成式范式展现出根本性的性能优势。即使是基线生成模型 HSTU 和 MTGR，也优于表现最佳的判别式模型 SASRec。这验证了将推荐重新表述为序列生成任务，是在大规模工业场景中建模冗长且复杂的用户历史的一种更有力的方法。

(3) **HPGR树立了新的最先进水平：** 我们提出的模型 **HPGR**不仅证明了生成式方法的优势，也显著拓展了其边界。完整的HPGR模型整体表现最佳，曲线下面积达到**0.8377**。与最强基线MTGR相比，这一成绩大幅提升，曲线下面积绝对值提高了 +0.0124，相对提升了 **+1.50%**。这一结果有力地表明，显式注入结构先验是对生成式推荐框架的一项关键且极为有效的增强。

4.3 完整数据集验证

为确认在抽样子集上观察到的优异性能具有稳健性和普适性，我们在完整的未抽样测试数据集上进行了验证实验。如图4所示，排名靠前的模型之间相对性能排名保持一致。在完整数据集上，HPGR 的曲线下面积达到令人瞩目的**0.89288**，比最强基线 MTGR 领先 **+0.01139** 个绝对曲线下面积点。这一结果有力地表明，我们的架构创新所带来的优势是根本性的，并且能够有效迁移至完整的真实世界数据分布，从而证实了主要发现的普适性。

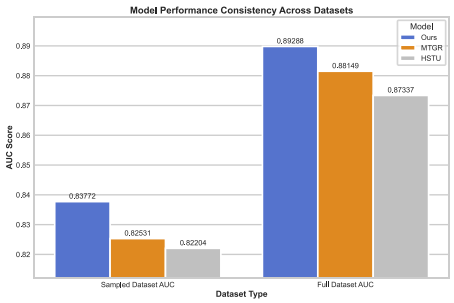

图4：排名靠前的模型在抽样测试数据集和完整测试数据集上的性能一致性。

4.4 消融研究

为剖析关键创新各自的贡献——结构感知预训练、PGSA机制和 SEM架构——我们开展了详细的消融研究。结果见表2下半部分，从基础配置开始逐步添加各个组件。

- **预训练的影响：HPGR（基础版，仅预训练）**模型采用标准的 HSTU 架构，但使用结构感知预训练阶段（采用 MIM目标）得到的物品嵌入进行初始化。其曲线下面积达到 0.8327。这一结果已优于表现最强的基线 MTGR，有力地验证了仅凭我们的预训练目标，就能生成从根本上更优、泛化能力更强的物品嵌入空间，相较于面向特定任务的端到端训练更胜一筹。

- **PGSA的影响：**在基础模型中加入 PGSA机制（**+ PGSA模块**）后，曲线下面积从 0.8327 提升至 0.8348。这一提升表明，动态地将注意力集中于历史物品中稀疏且偏好丰富的子集，能够有效降低噪声并增强相关信号，尤其适用于长序列。

- **SEM的影响：** 相比之下，仅在基础模型中加入SEM（**+ SEM模块**），曲线下面积便大幅提升至0.8374。这一显著改进凸显了明确建模用户行为时间层次结构的重要性。

表2：在我们的工业数据集上的性能比较。最佳结果以粗体显示，次佳结果以下划线标出。Impr.% 表示最佳 HPGR 模型相较于最强基线模型（以下划线标出）的相对提升。

类别	模型	曲线下面积 (↑)
判别式模型	深度神经网络（Covington 等人，2016）	0.7923
	MoE（Ma 等人，2018）	0.8024
	GRU4Rec（Hidasi 等人，2015）	0.8146
	BERT4Rec（Sun 等人，2019）	0.8145
	SASRec（Kang 和 McAuley，2018）	0.8194
	悟空（Zhang 等人 2024）	0.8167
生成模型	HSTU（Zhai等人 2024）	0.8220
	MTGR（Han等人 2025）	0.8253
我们的模型	HPGR（基础版，仅预训练）	0.8327
	+ PGSA模块	0.8348
	+ SEM模块	0.8374
	HPGR（完整模型）	**0.8377**
改进。		**+1.50%**

基础架构的扁平序列假设忽略了这一点。

- **所有组件的协同作用：**最后，同时整合了SEM和PGSA的**HPGR（完整模型）**取得了最佳的整体效果。这证明了我们核心创新的协同作用，凸显了在生成式范式下，捕捉用户行为的层次结构并将注意力聚焦于相关历史物品的价值，从而实现最先进的性能。

4.5 预训练嵌入空间的定性分析

为了验证我们的结构感知预训练，我们分析了 MIM 阶段前后的物品嵌入表。图 5 比较了 Xavier初始化（“随机初始化”）与我们的预训练嵌入（“训练”）。这一比较揭示了三种关键的结构性变化：

- **学习到的幅值变化**：L2范数分布（左上）从均匀分布转变为高度偏斜的分布。大多数物品可能代表长尾物品，它们集中在低范数区域。相反，一部分具有重要语义的物品则表现出更高的范数，说明模型学会压缩不太突出的物品，同时突出重要物品。
- **各向异性的显现**：奇异值衰减（右上）表明，结构从各向同性转变为各向异性。这证实模型利用特定维度来捕捉语义方差的主轴，而不是均匀地覆盖整个空间。

- **结构化语义景观**：UMAP 可视化（底排）展示了从无结构云团到明显子结构的转变。预训练嵌入形成了高密度区域，表明语义相似的物品已成功聚集在一起。

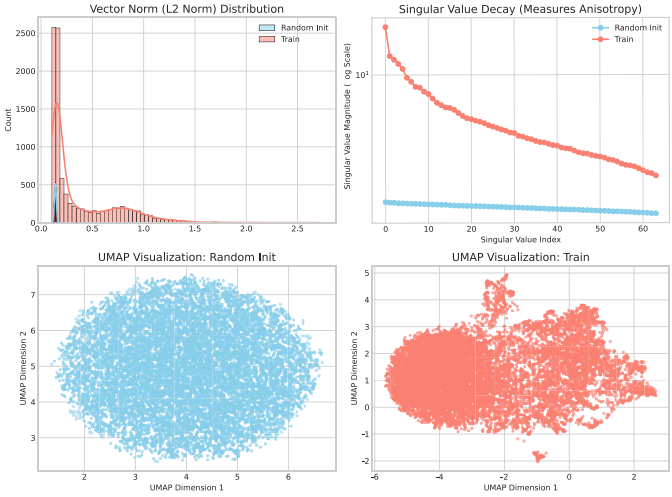

图 5：随机初始化与预训练的嵌入属性对比。左上：L2范数分布转变为高密度峰。底部一行：UMAP 可视化显示，嵌入形成了有意义的语义模式，与初始的随机点云形成对比。

4.6 超参数研究

我们研究了 PGSA 机制中 Top-K 值 K 的敏感性。如图 6 所示，性能在 $K = 50$ 时达到峰值，这表明该值能够捕获足够的上下文，同时避免引入过多噪声。其他所有实验中，我们均使用 $K = 50$ 。

4.7 可扩展性分析

我们分析了 HPGR 的扩展特性。如图 7 所示，随着用户历史记录变长，以及模型深度和宽度增加，性能持续提升。这表明

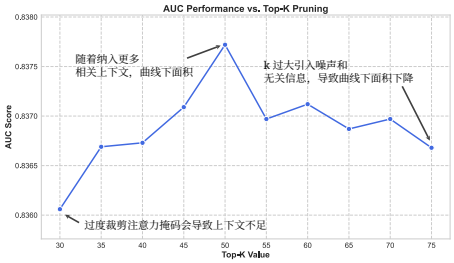

图 6：AUC 随 Top-K 值变化的超参数分析。

HPGR 是一种稳健且可扩展的架构，有望通过更多数据和计算资源进一步提升性能。

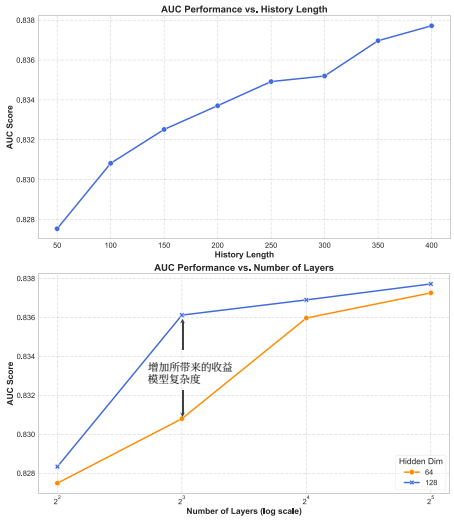

图 7：HPGR 随历史长度（上）和模型规模（下）的扩展性。

4.8 效率分析

我们还比较了 HPGR 与 MTGR 的计算效率。图 8 显示，尽管我们的模型使训练时间和推理时间有所增加，但这一投入带来了 +0.0124 的绝对曲线下面积提升。这种极具优势的权衡充分体现了我们在架构上的改进所具有的实用价值。

5 在线评估

为了最终验证我们提出的 HPGR 框架在现实场景中的影响和实际价值，我们在**华为应用市场**应用推荐平台上部署了模型，并进行了在线 A/B 测试。实验持续了 7 天周期，流量在我们提出的 HPGR模型与现有模型之间平均分配；现有模型是经过高度优化的生产环境DLRM，作为在线基线。

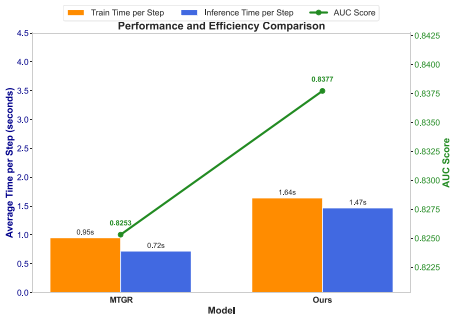

图 8：MTGR 与我们的 HPGR 模型在训练时间、推理时间和曲线下面积得分方面的比较。

我们的评估重点关注主要业务指标**eCPM（千次展示有效成本）**，该指标直接衡量平台收入，也是工业级推荐系统的一项关键指标。汇总结果以相对于在线基线的提升幅度表示，如表 3 所示。

表 3：在线 A/B 测试结果，离线曲线下面积在完整数据集上评估。该比较表明，HPGR 的离线优势转化为显著的在线 eCPM 提升。

模型	离线AUC	在线eCPM
生产环境DLRM（基线）	0.8880	基线
HPGR（我们的方法）	**0.8929**	**+1.99%**

在线 A/B 测试结果有力地证实了我们的模型在在线生产环境中的有效性。如表格所示，与经过高度优化的基线相比，HPGR 的 eCPM 提升了具有统计显著性的 +1.99%。

6 结论

本文介绍了 HPGR，这是一种创新框架，旨在解决现代生成式推荐器中的"扁平序列"局限。HPGR 通过两阶段范式注入关键结构先验：首先，在结构感知预训练阶段，借助会话增强模块（SEM）学习层次化表示；随后，在偏好感知微调阶段，使用偏好引导稀疏注意力（PGSA）聚焦相关历史记录。这种双重方法使模型能够捕捉复杂的、基于会话的用户行为。大量离线实验和在线 A/B 测试验证了 HPGR 的有效性，表明其达到了最先进的性能，并在生产环境中实现了显著的 +1.99% eCPM 提升。

致谢

本文研究得到国家自然科学基金（U22B2059、62276083）、黑龙江省重点研发计划（2024ZX01A05）和5G应用创新联合研究院项目（A003）的资助。

参考文献

[1] Josh Achiam、Steven Adler、Sandhini Agarwal、Lama Ahmad、Ilge Akkaya、Florencia Leoni Aleman、Diogo Almeida、Janko Altenschmidt、Sam Altman、Shyamal Anadkat等。2023。GPT-4技术报告。*arXiv*预印本 *arXiv:2303.08774*（2023）。[2] Jinze Bai、Shuai Bai、Yunfei Chu、Zeyu Cui、Kai Dang、Xiaodong Deng、Yang Fan、Wenbin Ge、Yu Han、Fei Huang等。2023。通义千问技术报告。*arXiv*预印本 *arXiv:2309.16609*（2023）。[3] Iz Beltagy、Matthew E Peters和Arman Cohan。2020。Longformer：长文档Transformer。*arXiv*预印本 *arXiv:2004.05150*（2020）。[4] Sneha Chaudhari、Varun Mithal、Gungor Polatkan和Rohan Ramanath。2021。注意力模型综述。*ACM*智能系统与技术汇刊（*TIST*）12卷5期（2021），1–32。[5] Heng-Tze Cheng、Levent Koc、Jeremiah Harmsen、Tal Shaked、Tushar Chandra、Hrishi Aradhye、Glen Anderson、Greg Corrado、Wei Chai、Mustafa Ispir等。2016。面向推荐系统的宽深度学习。载于第*1*届推荐系统深度学习研讨会论文集。7–10。[6] Rewon Child、Scott Gray、Alec Radford和Ilya Sutskever。2019。使用稀疏Transformer生成长序列。*arXiv*预印本 *arXiv:1904.10509*（2019）。[7] Paul Covington、Jay Adams和Emre Sargin。2016。用于YouTube推荐的深度神经网络。载于第*10*届*ACM*推荐系统大会论文集。191–198。[8] Tri Dao。2023。FlashAttention-2：通过更优的并行性和工作划分实现更快的注意力计算。*arXiv*预印本 *arXiv:2307.08691*（2023）。[9] Tri Dao、Dan Fu、Stefano Ermon、Atri Rudra和Christopher Ré。2022。FlashAttention：兼具快速、内存高效和I/O感知特性的精确注意力。神经信息处理系统进展35（2022），16344–16359。[10] Jacob Devlin、Ming-Wei Chang、Kenton Lee和Kristina Toutanova。2019。BERT：用于语言理解的深度双向Transformer预训练。载于*2019*年计算语言学协会北美分会会议论文集：人类语言技术，第*1*卷（长篇与短篇论文）。4171–4186。[11] Shijie Geng、Shuchang Liu、Zuohui Fu、Yingqiang Ge和Yongfeng Zhang。2022。作为语言处理的推荐（RLP）：统一的预训练、个性化提示与预测范式（P5）。载于第*16*届*ACM*推荐系统大会论文集。299–315。[12] Huifeng Guo、Ruiming Tang、Yunming Ye、Zhenguo Li和Xiuqiang He。2017。DeepFM：一种基于因子分解机的点击率预测神经网络。*arXiv*预印本 *arXiv:1703.04247*（2017）。[13] Ruidong Han、Bin Yin、Shangyu Chen、He Jiang、Fei Jiang、Xiang Li、Chi Ma、Mincong Huang、Xiaoguang Li、Chunzhen Jing等。2025。MTGR：美团工业级生成式推荐框架。*arXiv*预印本 *arXiv:2505.18654*（2025）。[14] Balázs Hidasi和Alexandros Karatzoglou。2018。基于会话的推荐中具有Top-k增益的循环神经网络。载于第*27*届*ACM*信息与知识管理国际会议论文集。843–852。[15] Balázs Hidasi、Alexandros Karatzoglou、Linas Baltrunas和Domonkos Tikk。2015。基于循环神经网络的基于会话的推荐。*arXiv*预印本 *arXiv:1511.06939*（2015）。[16] Gao Huang、Yu Sun、Zhuang Liu、Daniel Sedra和Kilian Q Weinberger。2016。具有随机深度的深度网络。载于欧洲计算机视觉会议。施普林格，646–661。[17] Wang-Cheng Kang和Julian McAuley。2018。自注意力序列推荐。载于*2018*年*IEEE*国际数据挖掘会议（*ICDM*）。IEEE，197–206。[18] Chao Li、Zhiyuan Liu、Mengmeng Wu、Yuchi Xu、Huan Zhao、Pipei Huang、Guoliang Kang、Qiwei Chen、Wei Li和Dik Lun Lee。2019。通过动态路由实现天猫推荐的多兴趣网络。载于第*28*届*ACM*信息与知识管理国际会议论文集。2615–2623。[19] Peng Liu、Lemei Zhang和Jon Atle Gulla。2023。预训练、提示与推荐：推荐系统中语言建模范式适配的全面综述。计算语言学协会汇刊11（2023），1553–1571。[20] Chen Ma、Peng Kang和Xue Liu。2019。用于序列推荐的分层门控网络。载于第*25*届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集。825–833。

[21] Jiaqi Ma、Zhe Zhao、Xinyang Yi、Jilin Chen、Lichan Hong 和 Ed H Chi。2018。使用多门混合专家模型对多任务学习中的任务关系进行建模。载于第*24*届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集。1930–1939。[22] Shervin Minaee、Tomas Mikolov、Narjes Nikzad、Meysam Chenaghlu、Richard Socher、Xavier Amatriain 和 Jianfeng Gao。2024。大语言模型：综述。*arXiv*预印本 *arXiv:2402.06196* (2024)。[23] Zhaopeng Qiu、Xian Wu、Jingyue Gao 和 Wei Fan。2021。U-BERT：通过预训练用户表征改进推荐。载于*AAAI*人工智能大会论文集，第35卷。4320–4327。[24] Jay Shah、Ganesh Bikshandi、Ying Zhang、Vijay Thakkar、Pradeep Ramani 和 Tri Dao。2024。FlashAttention-3：利用异步处理和低精度实现快速、准确的注意力。神经信息处理系统进展 37 (2024)，68658–68685。[25] Fei Sun、Jun Liu、Jian Wu、Changhua Pei、Xiao Lin、Wenwu Ou 和 Peng Jiang。2019。BERT4Rec：采用双向 Transformer 编码器表征的序列推荐。载于第*28*届*ACM*信息与知识管理国际会议论文集。1441–1450。[26] Hongyan Tang、Junning Liu、Ming Zhao 和 Xudong Gong。2020。渐进式分层提取（PLE）：一种用于个性化推荐的新型多任务学习（MTL）模型。载于第*14*届*ACM*推荐系统大会论文集。269–278。[27] Hugo Touvron、Thibaut Lavril、Gautier Izacard、Xavier Martinet、Marie-Anne Lachaux、Timothée Lacroix、Baptiste Rozière、Naman Goyal、Eric Hambro、Faisal Azhar 等。2023。Llama：开放且高效的基础语言模型。*arXiv*预印本 *arXiv:2302.13971* (2023)。[28] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。2017。注意力机制就是你所需要的一切。神经信息处理系统进展 30 (2017)。[29] Ruoxi Wang、Rakesh Shivanna、Derek Cheng、Sagar Jain、Dong Lin、Lichan Hong 和 Ed Chi。2021。DCN v2：改进的深度与交叉网络，以及网络规模学习排序系统的实践经验。载于*2021*年万维网大会论文集。1785–1797。[30] Shoujin Wang、Longbing Cao、Yan Wang、Quan Z Sheng、Mehmet A Orgun 和 Defu Lian。2021。基于会话的推荐系统综述。*ACM*计算调查（*CSUR*）54，7 (2021)，1–38。[31] Sinong Wang、Belinda Z Li、Madian Khabsa、Han Fang 和 Hao Ma。2020。Linformer：具有线性复杂度的自注意力。*arXiv*预印本 *arXiv:2006.04768* (2020)。[32] Likang Wu、Zhi Zheng、Zhaopeng Qiu、Hao Wang、Hongchao Gu、Tingjia Shen、Chuan Qin、Chen Zhu、Hengshu Zhu、Qi Liu 等。2024。面向推荐的大语言模型综述。万维网27，5 (2024)，60。[33] Wooseong Yang、Chen Wang、Zihe Song、Weizhi Zhang 和 Philip S Yu。2024。面向基于会话的推荐的物品簇感知提示学习。*arXiv*预印本 *arXiv:2410.04756* (2024)。[34] Zheni Zeng、Chaojun Xiao、Yuan Yao、Ruobing Xie、Zhiyuan Liu、Fen Lin、Leyu Lin 和 Maosong Sun。2021。通过预训练实现推荐中的知识迁移：回顾与展望。大数据前沿4 (2021)，602071。[35] Jiaqi Zhai、Lucy Liao、Xing Liu、Yueming Wang、Rui Li、Xuan Cao、Leon Gao、Zhao-jie Gong、Fangda Gu、Michael He 等。2024。行动胜于言语：用于生成式推荐的万亿参数序列转换器。*arXiv*预印本 *arXiv:2402.17152* (2024)。[36] Buyun Zhang、Liang Luo、Yuxin Chen、Jade Nie、Xi Liu、Daifeng Guo、Yanli Zhao、Shen Li、Yuchen Hao、Yantao Yao 等。2024。悟空：探索大规模推荐的缩放定律。*arXiv*预印本 *arXiv:2403.02545* (2024)。[37] Shuai Zhang、Lina Yao、Aixin Sun 和 Yi Tay。2019。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 (2019)，1–38。[38] Guorui Zhou、Xiaoqiang Zhu、Chenru Song、Ying Fan、Han Zhu、Xiao Ma、Yanghui Yan、Junqi Jin、Han Li 和 Kun Gai。2018。用于点击率预测的深度兴趣网络。载于第*24*届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集。1059–1068。[39] Kun Zhou、Hui Wang、Wayne Xin Zhao、Yutao Zhu、Sirui Wang、Fuzheng Zhang、Zhongyuan Wang 和 Ji-Rong Wen。2020。S3-Rec：通过互信息最大化进行序列推荐的自监督学习。载于第*29*届*ACM*信息与知识管理国际会议论文集。1893–1902。[40] Bohan Zhuang、Jing Liu、Zizheng Pan、Haoyu He、Yuetian Weng 和 Chunhua Shen。2023。Transformer高效训练综述。*arXiv*预印本 *arXiv:2302.01107* (2023)。