

超越语义理解：在基于大语言模型的推荐系统中保留协作频率成分

Minhao Wang
East China Normal University
Shanghai, China
51275901104@stu.ecnu.edu.cn

Yunhang He
East China Normal University
Shanghai, China
yhhe2004@gmail.com

Cong Xu
East China Normal University
Shanghai, China
cong xueric@gmail.com

Zhangchi Zhu
East China Normal University
Shanghai, China
zczhu@stu.ecnu.edu.cn

Wei Zhang*
East China Normal University
& Shanghai Innovation Institute
Shanghai, China
zhangwei.thu2011@gmail.com

Abstract

结合大型语言模型（LLMs）的推荐系统为生成基于语义的信息推荐提供了有前景的途径。然而，基于 LLM 的推荐系统在用户交互历史中表现出过分强调语义关联的倾向。当以预训练的协同 ID 嵌入作为输入时，随着嵌入逐层传递通过 LLM 主干，基于 LLM 的推荐系统逐渐削弱了固有的协同信号，这与在传统的基于 Transformer 的序列模型中协同信号通常被保留或甚至增强以实现最先进性能的做法相反。为了解决这一限制，我们引入了 FreLLM4Rec，该方法旨在从频谱视角平衡语义和协同信息。首先使用全局图低通滤波器（G-LPF）净化结合了语义和协同信息的物品嵌入，初步去除无关的高频噪声。然后，通过时间频率调制（TFM）逐层主动保留协同信号。请注意，TFM 的协同保留能力通过建立最佳但难以实现的局部图傅里叶滤波器与次优但计算效率高的频域滤波器之间的关系，得到了理论上的保证。在四个基准数据集上的大量实验证明，FreLLM4Rec 成功减轻了协同信号的衰减，并实现了具有竞争力的性能，在 NDCG@10 方面相比最佳基线提高了高达 8.00%。我们的研究结果提供了对 LLM 如何处理协同信息的见解，并为改进基于 LLM 的推荐系统提供了一种原则性的方法。我们的代码可在 <https://anonymous.4open.science/r/FreLLM4Rec/> 处获取。

CCS Concepts

• Information systems → Recommender systems.

Keywords

Sequential Recommendation, Large Language Models, Spectral Analysis, Graph Signal Processing, Collaborative Filtering

ACM Reference Format:

Minhao Wang, Yunhang He, Cong Xu, Zhangchi Zhu, and Wei Zhang. 2018. 超越语义理解：在基于大语言模型的推荐系统中保留协作频率成分. In *Proceedings of Make sure to enter the correct conference title from your rights*

*Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Conference acronym 'XX, Woodstock, NY

© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-XXXX-X/2018/06

<https://doi.org/XXXXXXX.XXXXXXX>

confirmation email (Conference acronym 'XX). ACM, New York, NY, USA, 10 pages. <https://doi.org/XXXXXXX.XXXXXXX>

1 介绍

大型语言模型作为推荐系统使用时，与主要关注拟合用户-项目交互数据的传统模型相比，提供了一种新的视角 [8, 24, 45]。最近的进展表明，LLM 可以有效地利用其预训练知识来解决推荐系统中的长期挑战，包括冷启动问题、跨领域转移和可解释性 [39, 54, 57]。这种演变引发了广泛的研究，从零样本排序到复杂的嵌入融合策略，探索将 LLM 集成到推荐管道中的各种范式。

基于 LLM 的推荐探索经历了不同的发展阶段，每个阶段都解决了集成挑战的不同方面。最初的方法主要集中于通过纯文本方法利用 LLM 的语义理解，将项目视为文本描述，并将推荐框定为语言生成任务 [1, 9]。然而，最近的研究表明，虽然语义信息可以捕捉项目属性和内容特征，但仅靠这一点不足以完全捕捉有效推荐所需的协同模式 [22, 30, 42]。这一认识促成了新一代的混合方法，结合了语义理解和协同信号。先进的框架现在将预训练的 ID 嵌入与文本表示结合，展示了显著的性能提升 [4, 14, 23, 39, 52]。这些发展强调了一个重要的见解：最有效的基于 LLM 的推荐系统是那些成功地平衡了语义理解和协同信息保留的系统。

尽管取得了这些进展，现有方法主要将 LLMs 视为黑箱，关注输入输出关系而未探究模型内部机制中协作信息是如何演变的。这一忽视导致了我们在理解上的一个根本性空白：LLMs 的深度架构如何处理和转换对推荐至关重要的协作信号？本文通过从光谱角度对 LLM 架构内协作信号转换进行首次系统性研究，以解决这一关键空白。具体而言，图光谱理论的最新发展表明，协作信息主要存在于由局部交互图得出的低频成分中。如果序列嵌入主要特征为低频成分，则假定其保留了大多数协作信息；反之，高频成分的占优将暗示上述衰减（见 Figure 1a）。相反，正如 Figure 1b 所示，在传统的基于 Transformer 的方法中，这种衰减并未观察到，其中低频成分在传播过程中甚至得到了增强，这通过在不同 LLM 骨干和附加数据集集中一致的观察进一步证实，如图 8 所示。我们将这种现象称为层内光谱衰减，其产生是因为基于 LLM 的推荐系统过度依赖其内部知识和推理能力，同时隐性地抵制协作信息的整合。

一个直接且理论上最优的解决方案是基于对应的局部图 [28, 33] 应用低通滤波，以保留大部分协作信息。然而，出现了两个实际障碍：(i) 采样局部图复杂且难以并行化，使该方法成本过高 [6]，以及 (ii) 该方法忽略了时间动态，丢弃了捕捉用

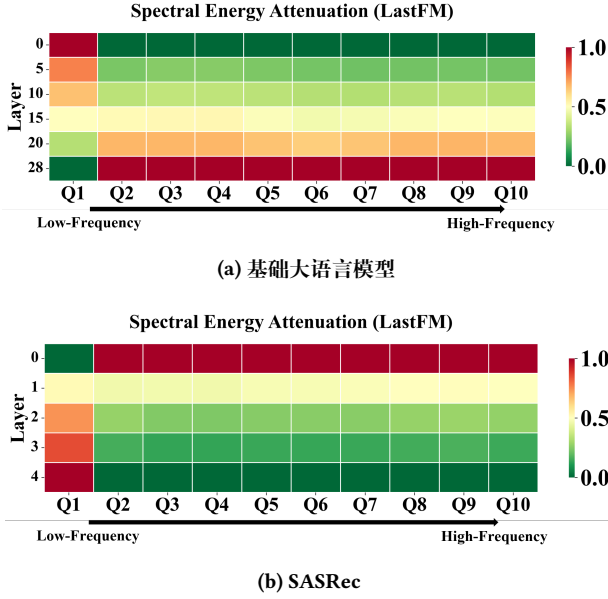


Figure 1: 发现跨多个数据集的层内光谱衰减现象。(a) 预训练的 LLM 显示随着层深度增加，低频协作信号能量系统地衰减。(b) 专为推荐设计的模型如 SASRec 保留了这些信号。这种差异解释了为什么直接应用 LLM 在推荐中表现不佳，尽管它们具有复杂的结构。

户偏好演变的序列模式。幸运的是，在某些温和的假设下，我们证明了频域滤波（特别是通过 (1) 离散傅立叶变换 (DFT) 转换信号和 (2) 衰减高频分量）可以保留等效的协作信息。虽然 DFT 之前已被用作 LLM 和序列推荐设置中的中间模块，但在我们的方法中，其应用是由一个完全不同的目标驱动的。例如，在推荐领域，FEARec 使用混合时频域注意力捕捉低/高频模式和周期性，而 BSARec 使用傅立叶变换与频率重新缩放来保留低高频模式，同时缓解过度平滑。在 LLM 领域，像 FourierFT 和 SFFT 这样的方法利用傅立叶变换进行参数高效微调，通过学习任务特定的频率分量。然而，这些方法将频率分析用作计算工具或架构组件，而我们的目标则是正交的——我们识别并抵消协作信号的系统性衰减，这是 LLM 推荐基于系统中独特且在之前研究中未被发现的现象。此外，实验证明，巴特沃斯滤波器比直接截断更适合作为自适应频域滤波的低通滤波器。自然地，当输入项目的嵌入仅来源于文本时，最初往往缺乏足够的协作信息。为了解决这一问题，FreLLM4Rec 建议项目嵌入结合语义特征（来自文本内容）和协作特征（来自预训练模型），随后应用全局低通滤波器以初步净化嵌入的协作信号。我们的工作做出了几个贡献。

- 发现层内谱衰减：我们在基于 LLM 的推荐系统中识别并正式描述这一现象，并为观察到的性能差距提供理论解释。
- 频率感知方法：我们提出了 FreLLM4Rec，它具有两个协同模块，能够在保持协作信号完整性的同时解决计算和架构上的挑战。
- 理论基础：我们提供严谨的分析，证明了时间频率调制与图谱特性保持之间的联系。
- 实证验证：在四个基准数据集上进行的广泛实验表明，NDCG@10 的性能提升了最多 8.00%，验证了协作信号的保留。

这项工作通过揭示谱衰减现象并提供一个有原则的解决方案，推动了对如何在推荐系统中有效应用大型语言模型的理解，为频率感知的神经架构研究开辟了新的途径。

2 初步

2.1 问题表述

设 $\mathcal{U}$ 表示用户集合， $\mathcal{V}$ 表示项目集合，其基数分别为 $|\mathcal{U}| = M$ 和 $|\mathcal{V}| = N$ 。每个用户 $u \in \mathcal{U}$ 有一个按时间顺序排列的交互序列 $S_u = (v_1^{(u)}, v_2^{(u)}, \dots, v_{|S_u|}^{(u)})$ ，其中 $v_t^{(u)} \in \mathcal{V}$ 表示用户在时间 t 交互的项目。序列推荐任务旨在预测使得用户效用最大化的下一个项目 $v_{|S_u|+1}^{(u)}$ ，其形式化为学习一个评分函数 $f: \mathcal{U} \times \mathcal{V} \rightarrow \mathbb{R}$ ，基于交互历史排列给定用户的所有候选项目。

2.2 图信号处理

图信号处理 (GSP) 将信号处理扩展到不规则数据，提供了一种在图上分析信息传播的方法 [28, 33]。在推荐系统中，GSP 揭示了协同信息如何通过神经架构演变。这个谱视角将协同信息分解为频率成分，携带着关于用户-物品关系的不同结构信息。

项目-项目共现图。我们构建了一个项目-项目图 $\mathcal{G} = (\mathcal{V}, W)$ ，其中节点代表项目，边量化协作信息。邻接矩阵 $W \in \mathbb{R}^{N \times N}$ 通过共现从用户项目交互矩阵 $R \in \mathbb{R}^{M \times N}$ 派生： $W = R^T R$ ，其中 W_{ij} 计算同时与项目 i 和 j 交互的用户数。这捕捉了协同过滤原则：在相同用户的交互历史中频繁出现的项目表现出相似的模式，并且很可能与相似的用户偏好相关 [41]。给定节点的图信号 $f \in \mathbb{R}^N$ 取自 W 。

对称归一化拉普拉斯矩阵。图的谱特性由以下公式表征： $L = I - D^{-1/2} W D^{-1/2}$ ，其中 D 是对角度矩阵，具有 $D_{ii} = \sum_j W_{ij}$ 。特征分解揭示了频率基： $L = U \Lambda U^T$ ，其中 $U = [u_1, u_2, \dots, u_N]$ 包含正交单位特征向量，和 $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_N)$ ，其中 $0 \leq \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_N \leq 2$ 。

图频率。特征值 λ_k 代表图频率。低频率 (λ_k 接近 0) 对应平滑信号，这些信号捕捉群体级别模式，对协同过滤至关重要。高频率 (λ_k 接近 2) 代表快速变化的信号，通常是对泛化不利的噪声 [27]。图傅里叶变换 (GFT) 对信号进行分解： $\hat{f} = U^T f$ ，其中 $\hat{f}_k = u_k^T f$ 是频率为 λ_k 的分量。逆变换： $f = U \hat{f} = \sum_{k=1}^N \hat{f}_k u_k$ 。

信号平滑度。平滑度通过拉普拉斯二次型来量化：

$$f^T L f = \sum_{i,j} w_{ij} (f_i - f_j)^2 = \sum_{k=1}^N \lambda_k |\hat{f}_k|^2. \quad (1)$$

，这惩罚了连接节点之间的差异。在频谱域中，平滑信号主要集中在低频率的能量。低频组件捕捉社区结构和项目相似性，这些对于协同过滤至关重要 [12, 44]。当这些组件在神经处理过程中减弱时，系统失去利用协同信息的能力。

3 基于 LLM 的推荐与频谱分析

将 LLMs 应用于推荐系统已经显示出很有前景的结果。然而，一个重要问题仍未被探索：预训练的语言模型如何经过优化以处理对语言任务的协同信号，这些信号对推荐至关重要呢？本节呈现了我们系统的调查，揭示了一种破坏基于 LLM 的推荐系统的现象。

尽管最近基于大语言模型 (LLM) 的推荐系统探索了多种范式，但嵌入作为令牌的方法已被视为特别有效的解决方案。这种范式能够高效地批量处理用户序列，自然而然地保留了预训练嵌入中编码的结构信息，并避免了基于文本方法固有的令牌词汇限制。此外，通过将项目嵌入视为输入令牌，这种方法允许 LLM 直接处理丰富的协作信号以及语义信息，在保持计算效率的同时实现竞争性的性能。嵌入输入的结构性也促进了我们系统性地研究合作信号如何在 LLM 架构中传播。

基于 LLM 的推荐中的一个关键挑战是在语言模型的输入空间内有效地表示物品。纯文本方法会遗漏协作模式，而仅使用协作嵌入则忽略了 LLM 擅长处理的语义信息 [13]。最近的研究表明，结合两种模式的混合方法优于单一模式的方法 [40, 53]。

为了解决这一挑战，我们的方法构建了整合两种信息来源的物品表示。对于每个物品 $i \in \mathcal{V}$ ，我们基于以下嵌入创建一个融合表示：

- 一种协作 ID 嵌入 $e_{id}(i) \in \mathbb{R}^{d_{id}}$ ，使用已建立的序列模型如 SASRec [19] 进行预训练，编码了物品的流行度和共同出现模式。
- 一种语义文本嵌入 $e_{text}(i) \in \mathbb{R}^{d_{text}}$ ，它源自 LLM 对项目元数据的表示，捕捉语言和上下文信息。

这些组件通过一个可学习的融合机制集成在一起，如下所示：

$$x_i = \text{MLP}([e_{id}(i); e_{text}(i)]). \quad (2)$$

这里，MLP 是一个带有非线性激活的两层网络，它将连接的嵌入投射到 LLM 的隐藏维度 d_{llm} 。

对于一个用户序列 S_u ，LLM 处理相应的嵌入序列：

$$(h_1^{(u)}, \dots, h_{|S_u|}^{(u)}) = \text{LLM}(x_{v_1^{(u)}}, \dots, x_{v_{|S_u|}^{(u)}}). \quad (3)$$

最终的隐藏状态 $h_{|S_u|}^{(u)}$ 作为用户表示 h_u ，并通过内积计算相关性得分： $s(u, j) = h_u^T x_j$ 。我们冻结 LLM 参数，只微调融合 MLP，这使我们能够在没有语言模型适应干扰因素的情况下，单独研究预训练架构如何影响协作信号。

3.1 揭开层内光谱衰减的面纱

借助分析方法，我们现在解决核心问题：协作信号在通过 LLM 的深层架构传播时如何演变？传统的评估只关注最终输出，未能提供产生这些输出的内部转变的洞察。我们开发了一种逐层谱分析方法，以研究信号在每个阶段的演变。

我们的方法认识到，Transformer 的自注意力机制会创建动态的、与上下文相关的关系。因此，与其分析单一的全局图，我们构建针对每个用户序列的本地图，确保我们的分析能够捕捉到与每次预测相关的实际协作结构。

DEFINITION 1 (LOCAL SPECTRAL ANALYSIS APPROACH). 对于每个用户序列 $S_u = (v_1^{(u)}, \dots, v_{|S_u|}^{(u)})$ ，我们定义：

- 局部图构建：我们将子序列 $(v_2^{(u)}, v_3^{(u)}, \dots, v_{|S_u|}^{(u)})$ 视为模型旨在预测的目标项的有序列表。从全局共现矩阵 W 中提取了邻接矩阵 $A_u \in \mathbb{R}^{T \times T}$ ，以捕捉特定于该序列的协作关系。
- 逐层信号提取：在每一层 l 中，我们提取隐藏状态 $H_u^{(l)} \in \mathbb{R}^{(T-1) \times d_{llm}}$ ，其中第 t 行包含用于预测项 $v_{t+1}^{(u)}$ 的表示。

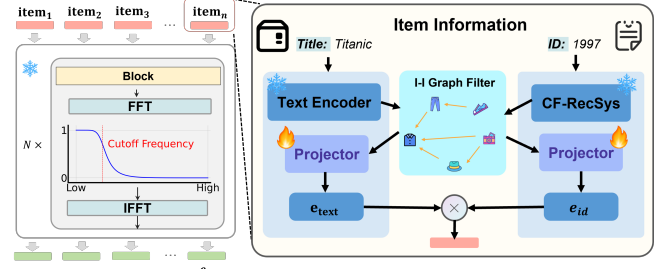


Figure 2: FreLLM4Rec 概述: 我们的频率感知方法将频谱衰减转换为频谱保留，G-LPF 模块净化输入协作信号，TFM 模块在每个 Transformer 层之后主动抵消衰减。

- 谱能量分析：局部广义傅立叶变换 (GFT) $\hat{H}_u^{(l)} = U_u^T H_u^{(l)}$ 将信号分解为频率分量。频率 k 处的能量如下：

$$\mathcal{E}_u(l, k) = \left\| (\hat{H}_u^{(l)})_k \right\|_F^2. \quad (4)$$

为了量化 LLM 架构中协同信号的演变，我们聚合跨用户的能量测量，并将频率划分为基于分位数的频带。通过这种系统的光谱特征，我们识别出一个在图 1a 中展示的一致现象：低频协同信号的能量在表示传播到更深层次时单调递减。这种对最平滑、最重要的协同模式的系统衰减，代表了 LLM 架构用于推荐的一种失效模式。

作为比较，我们还对 SASRec [19] 进行相同的分析，一种专门为序列推荐设计的 Transformer。如图 1b 所示，SASRec 在其各层中保持低频能量，表明频谱衰减不是 Transformer 架构固有的，而是语言导向的预训练中归纳偏差的结果。虽然 LLM 的自注意力机制擅长于语义处理，但它系统性地侵蚀了由低频图信号代表的重要、平滑的协同模式。

这一发现具有重要的意义。LLM 系统性衰减的低频成分恰好是那些编码社区偏好、项目相似性模式和协同过滤信号的成分——这些是有效推荐的基础。这一现象解释了为何许多基于 LLM 的推荐系统尽管具有强大的语义理解能力，却未能充分利用协同信息。这一现象促使我们在下一节提出频率感知的校正机制。

4 FreLLM4Rec 方法：频率感知推荐

受光谱衰减现象的启发，我们提出了 FreLLM4Rec（频率感知大语言模型的推荐），旨在在大语言模型处理流水线中保持协作信号的完整性。我们的方法引入了两个协同的频率感知模块：输入级别信号净化和层内信号保持，它们在基于大语言模型的推荐框架内的互补阶段进行操作。图 2 展示了完整的架构。

4.1 通过全局图低通滤波进行输入信号净化

虽然预训练的协同嵌入编码了有价值的互动模式，但它们也可能包含不太稳定的高频成分，这可能加剧了大型语言模型 (LLMs) 中的衰减问题。我们的 G-LPF 在全局项目-项目共现图 $\mathcal{G}$ （定义在第 2.2 节）上运行，该图捕获了整个项目目录中的协同关系。为了研究频率组成对推荐性能的影响，我们应用了一个理想的低通滤波器，逐步保留 ID 嵌入的低频成分的较大部分。图 3 展示的结果揭示了一个一致的模式：在包含更多低频成分后，性能最初有所改善，但在包含最高频成分后达到峰

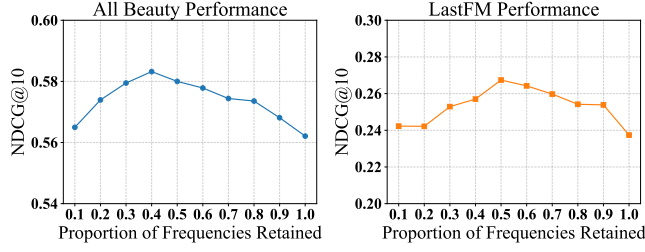


Figure 3: 保留低频成分百分比对 NDCG@10 的影响。

值并随后下降。这为最高频成分对推荐效果贡献较少提供了实证证据，激发了在输入层面进行纯化模块的必要性。

这一观察促使我们进行了第一次干预：即在大语言模型 (LLMs) 接收协作信号之前使用 G-LPF 进行净化。低频信号代表平滑变化，类似的项目具有相似的表示——这正是协作过滤的本质 [12]。高频成分通常编码流行偏差、测量噪声或阻碍泛化的虚假相关性 [5]。通过选择性地衰减这些可能有害的高频成分，我们为 LLMs 提供更纯净的协作信号。

理想的谱滤波操作将是：

$$E'_{item} = U \text{diag}(h(\lambda_1), \dots, h(\lambda_N)) U^T E_{item}, \quad (5)$$

，其中 $h(\lambda)$ 是一个低通频率响应函数。然而，这需要 $\mathcal{O}(N^3)$ 特征分解，对于推荐系统中大型项目目录来说，计算上是不可行的。

我们利用谱图理论中的一个结果：任何谱过滤器都可以通过拉普拉斯的多项式 [6] 来逼近：

$$h(\lambda) = \sum_{k=0}^K \theta_k \lambda^k \implies \mathcal{H}(L) = \sum_{k=0}^K \theta_k L^k. \quad (6)$$

我们的 G-LPF 应用了这个算子来纯化 ID 嵌入：

$$E'_{item} = \mathcal{H}(L) E_{item}. \quad (7)$$

现实世界的协作图的稀疏结构确保了计算的效率。在我们的实验中，我们采用了一个简化的一阶滤波器 ($K = 1$)，其响应为 $h(\lambda) = 1 - \alpha\lambda$ ，其中 $\alpha \in [0, 1]$ 控制滤波强度， $\alpha = 0$ 对应于无滤波，而 $\alpha = 1$ 提供最大平滑。

4.2 信号保留的层内时间频率调制

虽然输入净化有帮助，但它不能防止在 LLM 层内的系统衰减。为了解决这个问题，我们采用时间频率调制 (TFM)，它在每个 Transformer 层之后运行，以恢复退化的低频成分。然而，开发有效的层内校正存在挑战。首先，直接的局部图域操作需要每层进行昂贵的特征分解，使其计算上无法处理。其次，简单的频谱过滤无法利用用户序列中的时间依赖性。这一问题因需要保持与预训练的 LLM 兼容，同时确保校正不干扰语义处理而更加复杂。

实现 TFM 的关键洞见在于：在某些条件下，时间域和图频域可以互相关联：环图上的图傅立叶变换等同于离散傅立叶变换（见附录 B）。虽然用户序列并不构成完美的环图，但时间滤波可以通过以下方式影响图的谱特性：

ASSUMPTION 1 (SPATIO-TEMPORAL LOCALITY). 在用户序列中，时间上相近的项比时间上相隔较远的项表现出更高的协作相似性（更大的边权重）。

Table 1: 预处理数据集的统计信息。Avg. Len 表示用户的平均序列长度。

Dataset	# Users	# Items	# Interactions	Avg. Len
All Beauty	2,151	2,150	11,422	5.31
Luxury Beauty	11,490	6,534	71,898	6.26
Movies and TV	309,555	86,678	1,697,533	5.48
LastFM	1,090	3,646	52,551	48.21

这一假设在推荐领域中得到了验证 [11, 25, 31]，反映了平滑的偏好演变，即用户连续地与相似的物品进行交互。基于这一假设，我们可以建立时间域和图频域之间的联系：

THEOREM 1 (INFORMAL). 在假设 1 的情况下，对物品表示序列进行时间低通滤波可减少它们的图 Laplacian 二次型，并将能量集中在低图频率中。

定理 ??（在附录 C 中证明）提供了 TFM 的基础：时间上平滑隐藏状态可以恢复低频图组件，同时避免直接图操作的计算开销。

为了保留协作信号并避免可能扭曲序列模式的振铃伪影，我们使用 Butterworth 滤波器实现 TFM，该滤波器提供最大平坦的通带：

$$H^{(l)} = \text{TFM}(H^{(l)}) = \mathcal{F}^{-1} \left(\mathcal{B}(\omega) \odot \mathcal{F}(H^{(l)}) \right), \quad (8)$$

其中 $\mathcal{F}$ 表示沿着序列维度的 1D FFT，Butterworth 滤波器如下：

$$|\mathcal{B}(\omega)|^2 = \frac{1}{1 + (\omega/\omega_c)^{2n}}. \quad (9)$$

这里 ω_c 和 n 是截止频率和阶数超参数。

TFM 在每个 Transformer 层之后应用，创建持续的修正以对抗逐渐的衰减。这确保了协同信号在网络深度中保持强劲，使得 LLM 能够利用语义理解和保留的协同模式。

FreLLM4Rec 被设计为在大规模实际部署中使用。G-LPF 预处理的复杂度为 $\mathcal{O}(K \cdot |\mathcal{E}_{item}| \cdot d)$ ，其中 K 是多项式阶数， $|\mathcal{E}_{item}|$ 是稀疏项目图中的边数， d 是嵌入维度。这个预处理是一次性成本，并在整个训练过程中进行摊销。

TFM 模块在每层增加了 $\mathcal{O}(B \cdot d \cdot T \log T)$ 的计算成本，其中 B 是批处理大小， d 是隐藏维度， T 是序列长度。这个复杂度从渐近上看比 Transformer 的 $\mathcal{O}(B \cdot T^2 \cdot d)$ 自注意力复杂度要小。因此，FreLLM4Rec 保持了与标准 LLM 架构相同的可扩展性特征，同时提供了频率感知的校正。

5 实验

我们进行实验以验证 FreLLM4Rec 的有效性，并验证我们关于频谱衰减的理论见解。我们的评估解决了三个关键的研究问题：

- RQ1: FreLLM4Rec 的频率感知方法是否优于现有的序列推荐模型？
- RQ2: G-LPF 和 TFM 模块如何有助于缓解频谱衰减，FreLLM4Rec 在实践中是否能经验性地保留协作信号？
- RQ3: 光谱衰减现象和我们的解决方案在不同的 LLM 架构和协作信号源中是否具有鲁棒性？

Table 2: 在四个数据集上的总体性能比较。最佳结果以粗体显示，最好的基准线进行了下划线标记。“Improv.”表示 FreLLM4Rec 相对于最佳基准的相对提升。

Model	All Beauty		Luxury Beauty		Movies and TV		LastFM	
	NDCG@10	Recall@10	NDCG@10	Recall@10	NDCG@10	Recall@10	NDCG@10	Recall@10
GRU4Rec	0.5379	0.5904	0.4352	0.5797	0.5350	0.7531	0.1969	0.3440
Caser	0.5322	0.5895	0.3739	0.5104	0.5073	0.7013	0.1042	0.2073
SASRec	0.5774	0.6127	0.5069	0.6275	0.3566	0.5118	0.1996	0.3502
Bert4Rec	0.5370	0.5811	0.3961	0.4951	0.5486	0.7674	0.1782	0.3189
MoRec	0.5551	0.5809	0.4780	0.6209	0.4015	0.6260	0.2000	0.3710
FMLPRec	0.5662	0.6110	0.5065	0.6385	0.4440	0.6384	0.2185	0.3800
BSARec	0.5634	0.6013	0.4902	0.6081	0.3722	0.5760	0.2214	0.3862
SR-GNN	0.5502	0.5986	0.4274	0.5604	0.3909	0.5648	0.1563	0.2902
MAERec	0.5772	0.6145	0.4869	0.5982	0.4109	0.5996	0.1702	0.2911
LLaMA-3	0.1372	0.2527	0.0824	0.2033	0.1643	0.2824	0.0872	0.2147
LLARA	0.5418	0.5911	0.5072	0.6324	0.5346	0.7606	0.2810	0.4955
E4SRec	0.5415	0.5850	0.5120	0.6330	0.5550	0.7610	0.2305	0.3901
IDGenRec	<u>0.5821</u>	<u>0.6198</u>	0.5299	0.6412	0.3950	0.6100	0.2655	0.4705
LLM2Rec	0.5624	0.6071	<u>0.5354</u>	<u>0.6590</u>	<u>0.5976</u>	<u>0.7812</u>	<u>0.3097</u>	<u>0.5119</u>
FreLLM4Rec (Ours)	0.6287	0.6892	0.5618	0.7020	0.6311	0.8163	0.3327	0.5462
Improv.	8.00 %	11.20 %	4.93 %	6.52 %	5.61 %	4.49 %	7.43 %	6.70 %

5.1 实验设置

数据集。我们在四个基准数据集上进行评估，这些数据集涵盖电子商务和娱乐领域：来自亚马逊产品数据¹ [26] 的 All Beauty、Luxury Beauty 和 Movies and TV，以及用于音乐推荐的 LastFM² [3]。按照标准协议 [19, 59]，我们过滤出交互少于 5 次的用户和项目，并使用留一法策略进行训练/验证/测试集的划分。表 1 总结了数据集的统计信息。

基线。为了验证我们方法的有效性，我们将其与几条研究线进行比较：I) 传统的序列模型：GRU4Rec [15]，Caser [37]，SASRec [19]，BERT4Rec [34]，MoRec [53]；II) 频率域序列模型：FMLPRec [60]，BSARec [32]；III) 基于图的序列模型：SR-GNN [46]，MAERec [49]；IV) 基于大语言模型的推荐方法：LLaMA-3 [38]，LLARA [22]，E4SRec [16]，IDGenRec [55]，LLM2Rec [14]。更多细节见附录 ??。

实现细节。所有实验都使用在 NVIDIA A800 GPU 上的 PyTorch。对于 FreLLM4Rec 和 LLM 基线，我们使用 Qwen2.5-7B-Instruct [48] 作为默认主干。我们采用 AdamW 优化，批次大小为 32，学习率在 $\{1e-5, 5e-5, 1e-4, 5e-4\}$ 中搜索。G-LPF 参数 α 在 $[0, 1.0]$ 中调整，TFM 截止 ω_c 在 $[0, 1.0]$ 中调整。所有模型使用 50 维嵌入进行公平比较。

评估指标。我们报告 Recall@10 和 NDCG@10 作为我们的主要评估指标。对于每个序列，我们随机选择 100 个未交互的项目来构建候选集，确保包含正确的后续项目。这个 100 个项目的候选设置是为了能够与现有的大型语言模型推荐方法进行公平比较。对于传统模型，我们选择概率最高的候选项目作为预测，而基于 LLM 的模型通过各自的推理机制生成预测。

¹https://cseweb.ucsd.edu/~jmcauley/datasets/amazon_v2/

²<https://grouplens.org/datasets/hetrec-2011/>

Table 3: 在 LastFM、All Beauty 和 Luxury Beauty 数据集上的消融研究。

Model Variant	LastFM		All Beauty		Luxury Beauty	
	N@10	R@10	N@10	R@10	N@10	R@10
FreLLM4Rec (Full)	0.3327	0.5462	0.6287	0.6892	0.5618	0.7020
w/o G-LPF	0.2524	0.4523	0.5958	0.6671	0.5336	0.6575
w/o TFM	0.2507	0.4468	0.5840	0.6392	0.5181	0.6408
w/o G-LPF & TFM	0.2236	0.4037	0.5564	0.6090	0.5059	0.6257
w/o ID emb	0.0629	0.1404	0.5378	0.6134	0.4954	0.5946
w/o Text emb	0.2374	0.4257	0.5355	0.5909	0.4478	0.5967

5.2 整体性能比较 (RQ1)

表 2 展示了性能比较。我们的方法在所有数据集和指标上都取得了有竞争力的结果，尤其是在 LastFM 上有显著的提升 (NDCG@10 提升 7.43 %)，以及在 All Beauty 上提升 8.00 %。出现了几个关键的观察：

首先，FreLLM4Rec 优于所有基于 LLM 的基准方法，这表明解决光谱衰减对于有效的基于 LLM 的推荐非常重要。性能差异显著——尽管像 E4SRec 和 IDGenRec 这样的方法相比传统模型表现出适度的改进，FreLLM4Rec 实现了更大的性能提升。

其次，我们的方法超越了像 FMLPRec 和 BSARec 这样的专门频域模型，这验证了我们对 LLM 特定光谱衰减的有针对性的校正比通用的频率感知设计更为有效。

第三，在不同行业（美容产品、电影、音乐）中的持续改进，展示了光谱衰减现象及我们解决方案的普遍性。

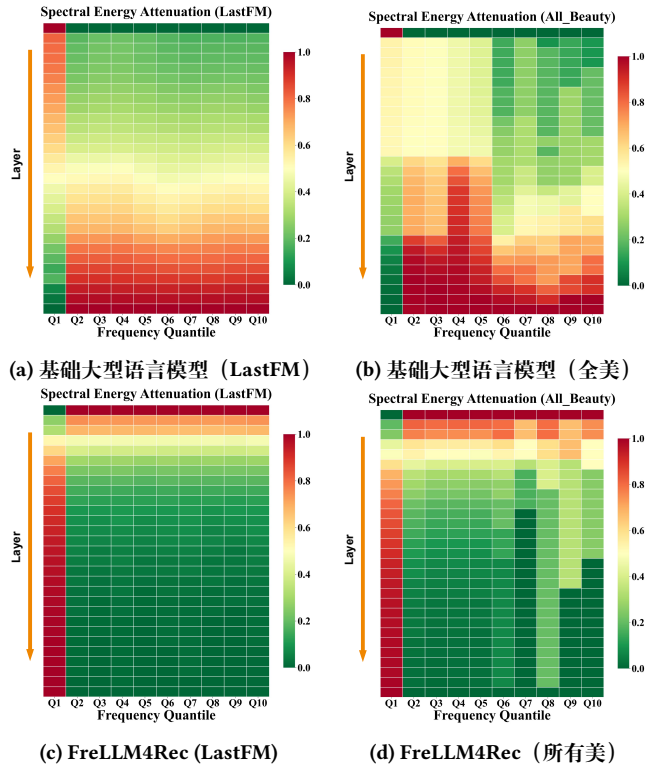


Figure 4: 谱衰减缓解的实证验证。FreLLM4Rec 在整个网络深度中成功保留了低频协同信号能量，这与传统 LLM 中的严重衰减形成鲜明对比。这种保留促进了我们方法的优越性能。

5.3 成分分析与频谱衰减缓解 (RQ2)

为了理解我们的频率感知模块如何缓解频谱衰减，我们进行了消融研究和经验频谱分析。

组件贡献。表 3 显示了各个模块的影响。去除 G-LPF 或 TFM 都会导致性能下降，其中 TFM 的影响更大（下降 6-8 %），这证实了层内校正对于对抗渐进式衰减是至关重要的。G-LPF 通过提供更清洁的输入信号贡献了 3-5 % 的提升。完整模型的表现优于单个组件改进之和，这表明输入净化和连续校正之间的协同作用。

嵌入分析显示，移除 ID 嵌入会导致性能下降高达 60 %，这确认了尽管大型语言模型具有语义能力，但协同信号仍然是不可或缺的。文本嵌入贡献了 5-10 % 的提升，表明语义理解是对协同过滤的补充，但不能替代它。

谱保真的实证验证。图 4 提供了视觉证据，表明 FreLLM4Rec 成功解决了谱衰减现象。利用我们的局部谱分析方法，我们通过网络层追踪低频协作信号能量。

在传统的大型语言模型（顶行）中，我们观察到低频能量呈指数衰减——到最后几层，协同信号已经损失了超过 70 % 的初始强度。这种显著的衰减解释了为什么直接应用大型语言模型在推荐任务上表现困难。相比之下，FreLLM4Rec（底行）通过 TFM 模块成功地在整个网络中保持了低频能量。这种协同信号的保留直接转化为我们在定量结果中观察到的优越的推荐性能。

Table 4: 消融研究比较了在 All Beauty、Movies 和 TV 以及 LastFM (NDCG@10) 上的不同架构中使用基本 LLM 主干 (Base) 与使用 FreLLM4Rec (+Fre) 的差异。

LLM Backbone	All Beauty		Movies and TV		LastFM	
	Base	+Fre	Base	+Fre	Base	+Fre
Qwen2.5-7B	0.5564	0.6287	0.6025	0.6311	0.2236	0.3327
Llama3.1-8B	0.5671	0.5886	0.6188	0.6402	0.2341	0.3174
Mistral-7B-v0.3	0.5794	0.6117	0.5961	0.6444	0.1776	0.3225

Table 5: 使用来自不同顺序推荐器的嵌入进行传统基于 ID 的模型 (Trad.) 和 FreLLM4Rec (Fre.) 性能 (NDCG@10) 比较。

ID Network	All Beauty		Movies and TV		LastFM	
	Trad.	Fre.	Trad.	Fre.	Trad.	Fre.
SASRec	0.5774	0.6287	0.3566	0.6311	0.1996	0.3327
MAERec	0.5772	0.6020	0.4109	0.6012	0.1702	0.2793
BSARec	0.5634	0.6112	0.3722	0.5983	0.2214	0.3062

5.4 跨架构的鲁棒性分析 (RQ3)

表 4 表明，频谱衰减是 LLM 架构固有限制，而不是特定模型的产物。我们的跨架构评估显示，虽然 Llama-3.1-8B 和 Mistral-7B 之间的基线性能有所不同，但两种架构都遭受类似的频谱降解模式。FreLLM4Rec 在所测试的所有基础模型中始终缓解了这一问题，实现了统一的 4-7 % 性能提升，同时保持 G-LPF 和 TFM 组件的稳定贡献比率，这表明我们的频域方法解决了核心架构限制，而不是实施特定的低效问题。

表 5 显示了对不同协作信号来源的鲁棒性。无论使用来自 SASRec、MAERec 还是 BSARec 的嵌入，FreLLM4Rec 始终提高了 3-6 % 的表现。这证实了我们的方法解决的是架构性问题，而不是特定嵌入方法的产物。

6 相关工作

6.1 序列推荐系统

序列推荐通过学习物品表示来预测用户未来可能交互的物品。从物品表示学习的角度来看，现有的方法大致可以分为三类：基于 ID 的方法、基于文本的方法和混合 ID-文本的方法。

基于 ID 的方法为每个项目分配一个唯一标识符，并使用各种序列建模技术来学习相应的表示 [15, 19, 34, 37]。这些基于 ID 的表示主要通过序列轨迹中建模多跳共同出现模式来编码协同过滤信号 [12, 44, 46, 47]。尽管这些方法在捕捉行为模式方面效果显著，但它们无法处理训练期间未见的项目或领域，从根本上限制了它们的泛化能力 [20, 51]。

基于文本的方法使用预训练的语言模型从项目的元数据中派生出文本嵌入来表示项目 [17, 24, 56]。然而，仅靠文本嵌入无法捕捉推荐所需的协同模式——它们错过了如项目受欢迎程度、用户社区偏好以及从集体互动中产生的行为模式等关键信号 [18, 58]。

混合 ID-文本方法通过各种融合策略尝试结合这两种模式：连接 ID 和文本嵌入 [52, 53]，使用文本增强 ID 表示 [29, 42]，或基于注意力的融合机制 [10]。虽然这些方法相对于单一模式的方法显示了改进，但它们仍然需要在目标领域数据上进

行训练以学习有效的表示。更重要的是，它们将融合视为静态组合，未考虑这些不同信号在深度神经架构中如何交互和转变 [14, 36]。

6.2 大型语言模型在推荐中

将大型语言模型应用于推荐系统代表了在如何实现个性化推荐方面的范式转变 [8, 21, 45]。这一演变经历了不同阶段的发展，每个阶段都解决了整合挑战的不同方面。最初的方法着重于利用基于文本的方法进行语义理解，将物品视为文本描述，并将推荐视为语言生成任务 [9, 24, 56]。随后的发展引入了嵌入作为令牌的范式，使用户序列的处理更为高效，同时保留了预训练嵌入中编码的结构信息 [16, 18, 55]。

最近的研究表明，虽然语义信息能够捕捉项目属性和内容特征，但仅靠语义信息并不足以完全捕捉到有效推荐所必需的协同关系 [22, 29, 42]。这一认识导致了通过各种融合策略将语义理解与协同信号结合的混合方法的出现 [4, 14, 23, 35, 39, 52, 58]。这些发展强调了一个关键的见解：最有效的基于 LLM 的推荐系统是那些成功平衡了语义理解与协同信息保存的系统。然而，现有的方法主要把语言模型当作黑箱对待，专注于输入输出关系，而没有研究协同信息在模型内部机制中如何演变。

6.3 推荐中的谱方法和频域分析

图信号处理表明，协作信息主要存在于表示平滑的、社区级模式的低频分量中，而高频分量通常编码噪声 [12, 28, 33, 43]。这种频谱视角表明，有效的图神经网络本质上作为低通滤波器 [27]，能够保留重要的协作信号。

近期的研究已经明确将频域分析纳入推荐体系结构中 [7, 32, 60]。在知识蒸馏中，频率感知方法揭示了不同的频谱成分具有不同的重要性 [61]。尽管传统推荐模型在这些方面取得了进展，但基于 LLM 的推荐系统的频谱特性仍未被探索。我们的工作填补了这一关键空白，首次识别出 LLM 中的层内频谱衰减现象，并提出 FreLLM4Rec，这是一种频率感知方法，通过目标频谱校正来保留协作信号。

7 结论与未来工作

结论。在本文中，我们识别出基于 LLM 推荐系统中的层内光谱衰减现象，这种现象系统性地削弱了对于有效推荐至关重要的协作信号。我们的光谱分析揭示，LLM 逐渐削弱编码社区偏好和交互模式的低频成分，这解释了为何复杂的语言模型在推荐任务中常常表现不佳。提出的 FreLLM4Rec 方法通过 G-LPF 和 TFM 模块引入频率感知的矫正，以净化输入信号并在整个网络中保持光谱完整性，从而成功减轻了光谱衰减现象。限制和未来工作。虽然 FreLLM4Rec 在各种推荐任务中展示了有效性和普适性，但它仍然需要对频域超参数 α 和 ω_c 进行调整。开发能够根据数据特性自动调整过滤参数的自适应频率控制机制留待未来研究。此外，在其他领域探索 LLMs 处理结构化信息时的光谱衰减现象也是一个潜在的研究方向。

References

- [1] Keqin Bao, Jizhi Zhang, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. 2023. TALLRec: An effective and efficient tuning framework to align large language model with recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems*. 1007–1014.
- [2] Stephen Boyd and Lieven Vandenberghe. 2004. *Convex optimization*. Cambridge university press.
- [3] Iván Cantador, Peter Brusilovsky, and Tsvi Kuflik. 2011. Second workshop on information heterogeneity and fusion in recommender systems (HetRec2011). In *Proceedings of the fifth ACM conference on Recommender systems*. 387–388.
- [4] Yang Chen, Yixuan Wang, Ting Dong, and Xiangnan He. 2025. Text-like Encoding of Collaborative Information in Large Language Models for Recommendation. In *Findings of the Association for Computational Linguistics: ACL 2025*. 3456–3470.
- [5] Zhiqian Chen, Feng Chen, Lei Zhang, Taoran Ji, Kaidi Fu, Liang Zhao, Fanglan Chen, Lingfei Wu, Charu Aggarwal, and Chang-Tien Lu. 2022. Revisiting the role of similarity based graph filtering in graph neural networks. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 121–131.
- [6] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. 2016. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in neural information processing systems*, Vol. 29.
- [7] Xinyu Du, Huanhuan Yuan, Pengpeng Zhao, Jianfeng Qu, Fuzhen Zhuang, Guan-feng Liu, Yanchi Liu, and Victor S Sheng. 2023. Frequency enhanced hybrid attention network for sequential recommendation. *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2023), 78–88.
- [8] Wenqi Fan, Zihui Zhao, Jiatong Li, Yunqing Liu, Xiaowei Mei, Yiqi Wang, Jiliang Tang, and Qing Li. 2023. Recommender systems in the era of large language models (LLMs). *arXiv preprint arXiv:2307.02046* (2023).
- [9] Shijie Geng, Shuchang Liu, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2022. Recommendation as language processing (RLP): A unified pretrain, personalized prompt & predict paradigm (P5). In *Proceedings of the 16th ACM Conference on Recommender Systems*. 299–315.
- [10] Jesse Harte, Wouter Zorgdrager, Panos Louridas, Asterios Katsifodimos, Dietmar Jannach, and Marios Fragkoulis. 2024. Leveraging Large Language Models for Sequential Recommendation. In *Proceedings of the 18th ACM Conference on Recommender Systems*. 1096–1108.
- [11] Ruining He and Julian McAuley. 2016. Fusing similarity models with markov chains for sparse sequential recommendation. In *2016 IEEE 16th international conference on data mining (ICDM)*. IEEE, 191–200.
- [12] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. 2020. LightGCN: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*. 639–648.
- [13] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*. 173–182.
- [14] Yingzhi He, Xiaohao Liu, An Zhang, Yunshan Ma, and Tat-Seng Chua. 2025. LLM2Rec: Large Language Models Are Powerful Embedding Models for Sequential Recommendation. arXiv:2506.21579 [cs.LG] <https://arxiv.org/abs/2506.21579>
- [15] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2015. Session-based recommendations with recurrent neural networks. In *arXiv preprint arXiv:1511.06939*.
- [16] Yupeng Hou, Shanlei Li, Zhoujun Liu, Xiaopeng Yao, Yuan Yao, and Jinyang Huai. 2024. E4SRec: An Elegant Effective Efficient Extensible Solution for Large Language Model Based Sequential Recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 792–802.
- [17] Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian McAuley, and Wayne Xin Zhao. 2024. Large language models are zero-shot rankers for recommender systems. In *European Conference on Information Retrieval*. Springer, 364–381.
- [18] Panfeng Ji, Shuai Wang, Yupeng Hou, and Fuli Feng. 2024. GenRec: Generative Sequential Recommendation with Large Language Models. In *Advances in Neural Information Processing Systems*, Vol. 37.
- [19] Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 197–206.
- [20] Hongyu Li, Mao Zhang, Fei Chen, and Xin Xia. 2024. LLM-Enhanced Cross-Domain Sequential Recommendation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 2317–2326.
- [21] Lei Li, Yongfeng Zhang, and Li Chen. 2024. Large Language Models for Generative Recommendation: A Survey and Visionary Discussions. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 4321–4331.
- [22] Jiayi Liao, Sihang Li, Zhengyi Yang, Jiancan Wu, Yancheng Yuan, Xiang Wang, and Xiangnan He. 2024. LLaRA: Large Language-Recommendation Assistant. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2132–2136.
- [23] Bin Liu, Tao Zhu, Jianxun Pan, and Fuli Feng. 2025. Triple Modality Fusion: Aligning Visual, Textual, and Graph Data with Large Language Models for Multi-Behavior Recommendations. In *Proceedings of the ACL 2025 Conference*. 789–803.
- [24] Peng Liu, Lemei Zhang, and Jon Atle Gulla. 2023. Pre-train, prompt and recommendation: A comprehensive survey of language modelling paradigm adaptations in recommender systems. *Transactions of the Association for Computational Linguistics* 11 (2023), 1553–1571.

- [25] Julian McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*. 165–172.
- [26] Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*. 188–197.
- [27] Hoang NT and Takanori Maehara. 2019. Revisiting graph neural networks: All we have is low-pass filters. In *arXiv preprint arXiv:1905.09550*.
- [28] Antonio Ortega, Pascal Frossard, Jelena Kovačević, José MF Moura, and Pierre Vanderghenst. 2018. Graph signal processing: Overview, challenges, and applications. *Proc. IEEE* 106, 5 (2018), 808–828.
- [29] Xubin Ren, Wei Wei, Lianghao Xia, Lixin Su, Suqi Cheng, Junfeng Wang, Dawei Yin, and Chao Huang. 2024. Representation Learning with Large Language Models for Recommendation. In *Proceedings of the ACM Web Conference 2024*. 3464–3475.
- [30] Zhaopeng Ren, Zihan Wang, Zijian Ke, Fuzhen Zhuang, Zhenya Liu, Qing Zhu, and Qi Liu. 2025. Large Language Models meet Collaborative Filtering: An Efficient All-round LLM-based Recommender System. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2341–2351.
- [31] Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. 2010. Factorizing personalized markov chains for next-basket recommendation. In *Proceedings of the 19th international conference on World wide web*. 811–820.
- [32] Yehjin Shin, Jeongwhan Choi, Hyowon Wi, and Noseong Park. 2024. An Attentive Inductive Bias for Sequential Recommendation beyond the Self-Attention. *arXiv:2312.10325 [cs.LG]* <https://arxiv.org/abs/2312.10325>
- [33] David I Shuman, Sunil K Narang, Pascal Frossard, Antonio Ortega, and Pierre Vanderghenst. 2013. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE Signal Processing Magazine* 30, 3 (2013), 83–98.
- [34] Fei Sun, Jun Liu, Jian-Yun Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. 2019. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 1441–1450.
- [35] Zhongxiang Sun, Zihua Si, Xiaoxue Zang, Kai Zheng, Yang Song, Xiao Zhang, and Jun Xu. 2024. Large Language Models Enhanced Collaborative Filtering. *arXiv preprint arXiv:2403.17688* (2024).
- [36] Youhua Tan, Chang Liu, Zhiqi Wei, Guannan Chen, and Xiaochun Zheng. 2024. ID vs. Modality: Which is Better for Sequential Recommendation?. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 3421–3430.
- [37] Jiayi Tang and Ke Wang. 2018. Personalized Top-N Sequential Recommendation via Convolutional Sequence Embedding. *arXiv:1809.07426 [cs.LR]* <https://arxiv.org/abs/1809.07426>
- [38] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).
- [39] Lingfeng Wang, Jiayu Zhang, Xiang Chen, Yujie Ma, and Defu Lian. 2025. Order-agnostic Identifier for Large Language Model-based Generative Recommendation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 123–132.
- [40] Shoujin Wang, Liang Hu, Yan Wang, Longbing Cao, Quan Z Sheng, and Mehmet Orgun. 2021. A survey on deep learning for recommender systems. *Knowledge and Information Systems* 61, 1 (2021), 1–50.
- [41] Xiang Wang, Xiangnan He, Meng Wang, Fuli Feng, and Tat-Seng Chua. 2019. Neural graph collaborative filtering. In *Proceedings of the 42nd international ACM SIGIR conference on Research and development in Information Retrieval*. 165–174.
- [42] Wei Wei, Xubin Ren, Jiabin Tang, Qinyong Wang, Lixin Su, Suqi Cheng, Junfeng Wang, Dawei Yin, and Chao Huang. 2024. LLMRec: Large Language Models with Graph Augmentation for Recommendation. In *Proceedings of the 18th ACM International Conference on Web Search and Data Mining*. 806–815.
- [43] Felix Wu, Amauri Souza, Tianyi Zhang, Christopher Fifty, Tao Yu, and Kilian Weinberger. 2019. Simplifying graph convolutional networks. In *International conference on machine learning*. PMLR, 6861–6871.
- [44] Jiancan Wu, Xiang Wang, Fuli Feng, Xiangnan He, Liang Chen, Jianxun Lian, and Xing Xie. 2021. Self-supervised graph learning for recommendation. In *Proceedings of the 44th international ACM SIGIR conference on Research and development in information retrieval*. 726–735.
- [45] Likang Wu, Zhi Zheng, Zhaopeng Qiu, Hao Wang, Hongchao Gu, Tingjia Shen, Chuan Qin, Chen Zhu, Hengshu Zhu, Qi Liu, et al. 2023. A survey on large language models for recommendation. *arXiv preprint arXiv:2305.19860* (2023).
- [46] Shu Wu, Yuyuan Tang, Yanqiao Zhu, Liang Wang, Xing Xie, and Tieniu Tan. 2019. Session-based recommendation with graph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 346–353.
- [47] Xu Xie, Fei Sun, Zhaoyang Liu, Shiwen Wu, Jinyang Gao, Jiandong Zhang, Bolin Ding, and Bin Cui. 2022. Contrastive learning for sequential recommendation. In *2022 IEEE 38th international conference on data engineering (ICDE)*. IEEE, 1259–1273.
- [48] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 Technical Report. *arXiv preprint arXiv:2407.10671* (2024).
- [49] Yaowen Ye, Lianghao Xia, and Chao Huang. 2023. Graph Masked Autoencoder for Sequential Recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23)*. ACM, 321–330. doi:10.1145/3539618.3591692
- [50] Lu Yu, Jun Xie, and Xiang Zheng. 2023. The relationship between graph Fourier transform (GFT) and discrete cosine transform (DCT) for 1D signal and 2D image. *Signal, Image and Video Processing* 17, 2 (2023), 445–451.
- [51] Haipeng Yuan, Shuai Wang, Zhen Li, and Xiangnan He. 2024. Cross-Domain Sequential Recommendation via Neural Process. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 9876–9884.
- [52] Xiaoyu Yuan, Yang Liu, Xin Xia, and Jiliang Yin. 2024. Customizing Language Models with Instance-wise LoRA for Sequential Recommendation. In *Advances in Neural Information Processing Systems*, Vol. 38.
- [53] Zheng Yuan, Fajie Yuan, Yu Song, Youhua Li, Junchen Fu, Fei Yang, Yunzhu Pan, and Yongxin Ni. 2023. Where to Go Next for Recommender Systems? ID- vs. Modality-based Recommender Models Revisited. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (Taipei, Taiwan) (SIGIR '23)*. Association for Computing Machinery, New York, NY, USA, 2639–2649. doi:10.1145/3539618.3591932
- [54] Jiaqi Zhai, Lucy Liao, Xing Liu, Yueming Wang, Rui Li, Xuan Cao, Leon Gao, Zhaojie Gong, Fangda Gu, Michael He, et al. 2024. Actions speak louder than words: Trillion-parameter sequential transducers for generative recommendations. *arXiv preprint arXiv:2402.17152* (2024).
- [55] Juntao Zhang, Yiqun Xie, Hao Zhang, Menghan Yu, Shuguang Chen, Bo Shen, Menghan Zhang, Weixiang Chen, Chi-min Wu, and Peng Cui. 2024. IDGenRec: LLM-RecSys Alignment with Textual ID Learning. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 812–822.
- [56] Zeyu Zhang, Xiaocan Feng, Yong Wang, and Defu Lian. 2024. Ranking-based Prompting for Large Language Models in Recommendation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 8901–8915.
- [57] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223* (2023).
- [58] Xinyu Zhao, Hansi Ren, Lei Li, and Yue Wang. 2024. Lifelong Sequential Recommendation with Large Language Models. In *Proceedings of the International Conference on Learning Representations*.
- [59] Kun Zhou, Hui Wang, Wayne Xin Zhao, Yutao Zhu, Sirui Wang, Fuzheng Zhang, Zhongyuan Wang, and Ji-Rong Wen. 2020. S3-rec: Self-supervised learning for sequential recommendation with mutual information maximization. In *Proceedings of the 29th ACM international conference on information & knowledge management*. 1893–1902.
- [60] Kun Zhou, Hui Yu, Wayne Xin Zhao, and Ji-Rong Wen. 2022. Filter-enhanced MLP is all you need for sequential recommendation. In *Proceedings of the ACM Web Conference 2022*. 2388–2399.
- [61] Zhangchi Zhu and Wei Zhang. 2025. Exploring Feature-based Knowledge Distillation for Recommender System: A Frequency Perspective. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*. 2182–2193.

在本文中，我们考虑以下基线：

- GRU4Rec [15]：一种开创性的基于 RNN 的方法，该方法将门控循环单元应用于基于会话的推荐，通过循环连接来建模序列模式。
- Caser [37]：采用卷积神经网络使用水平和垂直卷积滤波器捕捉点级和联合级的序列模式。
- SASRec [19]：利用自注意力机制从用户的行为历史中识别相关项目，使模型能够自适应地捕捉长距离依赖关系。
- BERT4Rec [34]：通过 Cloze 任务将双向 Transformer 架构用于序列推荐，使模型能够利用左右两侧的上下文。
- MoRec [53]：研究基于 ID 与基于模态的推荐范式的有效性，提出了一种统一这两种方法的端到端训练框架。
- FMLPRec [60]：用频域中的可学习滤波器替换自注意力，展示了简单的频域操作可以以更低地计算成本实现有竞争力的性能。

- BSARec [32]：引入一种注意归纳偏置机制，该机制超越了自注意力，结合频率感知组件以更好地捕捉序列模式。
- SR-GNN [46]：将会话序列建模为图结构，并通过门控图神经网络应用图神经网络，以捕捉复杂的项目转换。
- MAERec [49]：使用图遮蔽自动编码器进行顺序推荐，通过对项目转换图进行自监督预训练来学习稳健的表示。
- LLaMA-3 [38]：一个开源的大型语言模型，我们通过零样本提示将其调整用于推荐任务，利用其预训练知识来理解项目。
- LLARA [22]：一种大型语言推荐助手，通过将大型语言模型与协同信号相结合的混合架构，弥合语言理解与推荐之间的差距。
- E4SRec [16]：提出了一种基于 LLM 的序列推荐框架，通过一种新颖的索引策略有效处理项目 ID，使 LLM 能够直接处理和生成项目标识符，同时保持推荐准确性和计算效率。
- IDGenRec [55]：通过文本化 ID 学习将 LLMs 与推荐系统对齐，从而更好地将语义理解与协同过滤信号结合。
- LLM2Rec [14]：展示了大型语言模型可以通过利用其预训练的表示，在顺序推荐中充当强大的嵌入模型。

A 超参数分析与敏感性

图 5 展示了 G-LPF 强度 α 的影响。随着过滤强度增加到 $\alpha \approx 0.5$ ，性能提高，这证实了去除高频噪声是有帮助的。在这之后，过度过滤开始去除有用信号，导致性能下降。最佳范围 (0.2-0.5) 表明了鲁棒性。

图 6 分析了 TFM 截止频率 ω_c 。在没有过滤 ($\omega_c = 1.0$) 的情况下，频谱衰减会影响性能。当我们减少 ω_c 时，通过保留低频，性能得到改善。最佳范围 (0.1-0.5) 在保持协作模式的同时维持必要的顺序动态之间取得平衡。

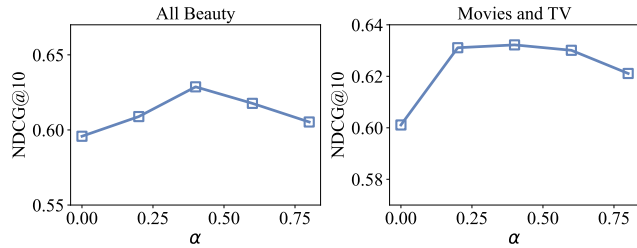


Figure 5: G-LPF 滤波强度分析。适度滤波 ($\alpha \in [0.2, 0.5]$) 最佳地平衡了噪声去除和信号保留。

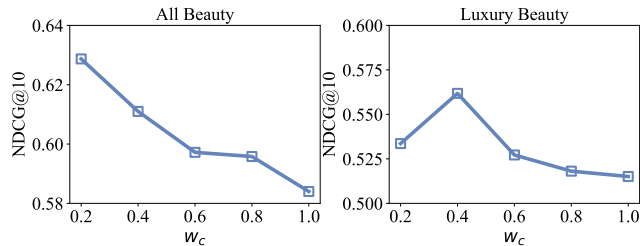


Figure 6: TFM 截止频率分析。较低的截止频率更好地保留协作信号，其最佳性能在 $\omega_c \in [0.1, 0.5]$ 。

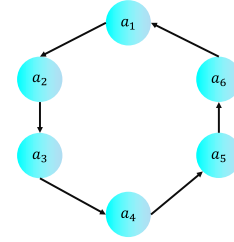


Figure 7: 具有 XMATHX_N 个节点的环状图。

B 连接时间域和图频域

离散傅里叶变换 (DFT) 对于一个时间信号 $x = [x_0, x_1, \dots, x_{T-1}]^T$ 被定义为：

$$\mathcal{F}(x)[k] = \sum_{t=0}^{T-1} x_t e^{-j2\pi kt/T}, \quad (10)$$

，其中 $k = 0, 1, \dots, T-1$ 代表频率索引。以矩阵形式，这变为 $\hat{x} = Fx$ ，其中 DFT 矩阵 F 的条目为 $F_{kn} = \frac{1}{\sqrt{T}} e^{-j2\pi kn/T}$ 。

当我们认识到 DFT 可以被解释为有向循环图上的图傅里叶变换 (GFT) 时，基本联系就出现了，这种图通常被称为环图 (如图 7 所示)。在这个环结构中，每个节点连接到其时间上的后继节点，形成一个循环链。该环图的拉普拉斯特征向量正是 DFT 的基向量 [50]，这表明当应用于适当结构化的数据时，时间频率分析和图谱分析在数学上是等价的。

这种等价性揭示了 DFT 中的时间频率 k 对应于图的特征值：

$$\lambda_k = 2 - 2 \cos(2\pi k/T). \quad (11)$$

低时间频率 (小 k) 对应于小特征值，表示序列中缓慢变化的模式。相反，高时间频率 (大 k) 对应于大特征值，捕捉相邻元素之间的快速波动。

对于推荐系统，这种关联具有深远的影响。当我们处理代表用户交互历史的一系列项目嵌入时，时间平滑操作会鼓励序列中的相邻项目具有相似的表示。在合理假设用户在时间上连续地与相似项目交互作为偏好局部性的表现的情况下 [19]，时间平滑直接转化为增强项目共现图上的平滑性。

C 定理 ?? 的证明

PROOF. 该定理指出，对信号应用时间低通滤波器可以增强其图平滑性，进而使信号的能量集中在图谱的低频带。我们将分两个主要部分来证明这一点。第 1 部分：时间低通滤波提高图平滑性。设 $f \in \mathbb{R}^T$ 是一个一维时间序列 (图 $\mathcal{G}$ 节点上的信号)，并设 $f' = \text{LPF}_{\text{time}}(f)$ 是应用时间低通滤波器后的信号。信号的图平滑性由拉普拉斯二次型 $S_G(f) = f^T L f$ 测量，其可以展开为：

$$S_G(f) = \sum_{i=1}^T \sum_{j=1}^T w_{ij} (f_i - f_j)^2 \quad (12)$$

其中 w_{ij} 是图的邻接矩阵 W 的元素。我们的目标是证明 $S_G(f') \leq S_G(f)$ 。时间低通滤波器是一种卷积操作，作用类似于局部平均函数。在时刻 t 处，经过滤的信号值 f'_t 是围绕 t 的局部邻域中原始信号值的加权平均。这本质上减少了相邻点之间的差异。根据假设 1 (时空局部性)，对于在时间序列中接近的节点 i 和 j (即 $|i-j|$ 小)，邻接权重 w_{ij} 较大。这些较大的 w_{ij} 项在 $S_G(f)$ 的和中占主导地位。时间滤波器的局部平均效应系

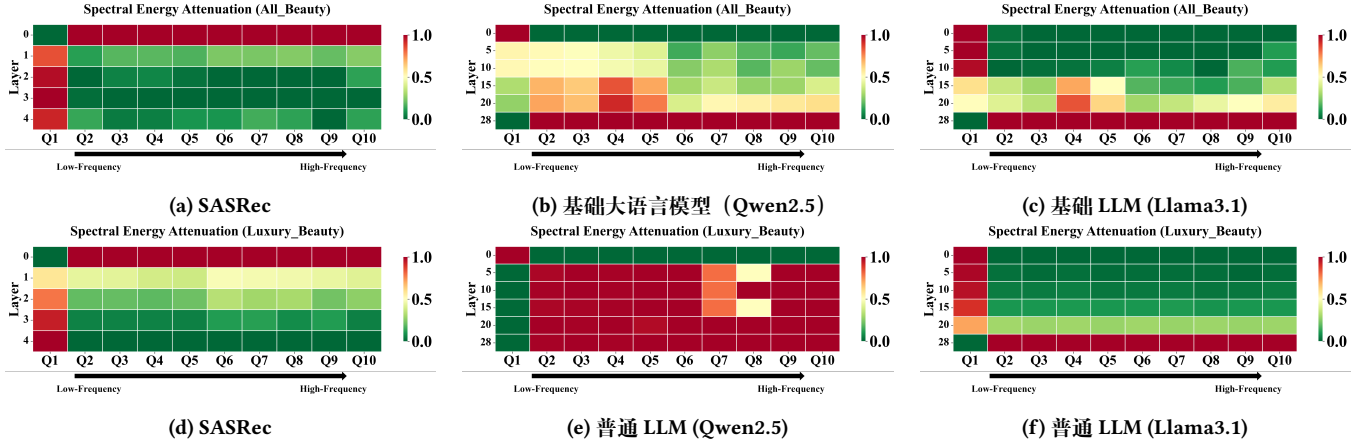


Figure 8: 在 All Beauty 和 Luxury Beauty 数据集上不同 LLMs 的光谱衰减现象

统地减少了这些占主导地位的对中 $|i - j|$ 小的差异 $(f'_i - f'_j)^2$ 。更正式地说，经过滤信号中的差异 $(f'_i - f'_j)^2$ 是原始信号中一组局部差异的函数。令低通滤波器由线性算子 H 表示，因此 $f' = Hf$ 。差异转换为 $(f'_i - f'_j)^2 = (\sum_k h_{i,k} f_k - \sum_k h_{j,k} f_k)^2$ 。由于卷积核 H 的平滑性质，结果差异被削弱。应用 Jensen 的不等式 [2] 对于凸函数 $\phi(x) = x^2$ 提供了方差缩减的正式基础。在所有由 w_{ij} 加权的 (i, j) 对中求和，我们得出总图平滑度减小：

$$S_G(f') = \sum_{i,j} w_{ij} (f'_i - f'_j)^2 \leq \sum_{i,j} w_{ij} (f_i - f_j)^2 = S_G(f) \quad (13)$$

这建立了定理的第一部分。部分 2：增强的图平滑度意味着图低频谱中的能量集中。从 GSP，我们有一个将图平滑度与谱域关系的恒等式：

$$S_G(f) = f^T L f = \hat{f}^T \Lambda \hat{f} = \sum_{k=1}^T \lambda_k |\hat{f}_k|^2 \quad (14)$$

其中 $\hat{f}_k$ 是 f 的第 k 个 GFT 系数， λ_k 是相应的图频率（特征值）。从部分 1，我们有不等式 $S_G(f') \leq S_G(f)$ 。应用上面提到的恒等式，我们得到：

$$\sum_{k=1}^T \lambda_k |\hat{f}'_k|^2 \leq \sum_{k=1}^T \lambda_k |\hat{f}_k|^2 \quad (15)$$

这表明经过滤后，总能量随着图频率加权后减少。此外，GFT 的 Parseval 定理指出，信号的总能量在谱域中保持恒定： $\|f\|_2^2 = \sum_{k=1}^T |\hat{f}_k|^2$ 。尽管时间低通滤波器可能略微减少总信号能量，我们可以假设它基本保持恒定，或者信号可以被重新归一化使得：

$$\sum_{k=1}^T |\hat{f}'_k|^2 \approx \sum_{k=1}^T |\hat{f}_k|^2 \quad (16)$$

我们现在有两个条件：频率加权能量之和减少（方程 15），同时能量总和保持大致恒定（方程 16）。鉴于图的频率 λ_k 按减少顺序排序（ $0 \leq \lambda_1 \leq \dots \leq \lambda_T$ ），满足这两个条件的唯一方法是将能量分布 $|\hat{f}'_k|^2$ 从具有大 λ_k （高频率）的项转移到具有小 λ_k （低频率）的项。因此，应用时间低通滤波器会导致信

号在图谱域中的能量重新分布。这实现了对低频带的更多集中。 □