计 算 机 研 究 与 发 展
Journal of Computer Research and Development
https://crad.ict.ac.cn
63(8): 2169−2181, 2026

基于大语言模型的多模态自适应融合推荐算法

肖光义[1] 侯 旻[1] 廖宇欣[1] 武 晗[1] 白晨希[1] 魏 思[2] 吴 乐[1]
[1](智能互联系统安徽省实验室(合肥工业大学) 合肥 230009)
[2](科大讯飞股份有限公司人工智能研究院 合肥 230088)
(479245378@qq.com)

Large Language Model-Based Multimodal Recommendation Algorithm with Adaptive Modality Fusion

Xiao Guangyi[1], Hou Min[1], Liao Yuxin[1], Wu Han[1], Bai Chenxi[1], Wei Si[2], and Wu Le[1]
[1] (*Intelligent Interconnected Systems Laboratory of Anhui Province* (*Hefei University of Technology*), *Hefei* 230009)
[2] (*Artificial Intelligence Research Institute*, *iFLYTEK Co.*, *Ltd.*, *Hefei* 230088)

Abstract Multimodal recommendation methods have introduced features such as text and images to enrich item representations, significantly enhancing recommender system performance. In recent years, large language models (LLMs), with their extensive world knowledge and powerful reasoning capabilities, have shown great potential in the field of recommendation. However, effectively integrating LLMs with multimodal information to construct LLM-based multimodal recommender systems still faces several challenges. Firstly, there is a significant semantic gap between the textual input format of LLMs and the multimodal information, which limits the model's ability to fully understand multimodal features. Secondly, there exists an information imbalance between different modalities, where naive feature fusion can lead to performance degradation. To address these challenges, we propose an LLM-based multimodal recommendation algorithm with adaptive modality fusion (AMFRec). The framework introduces multimodal hybrid prompt templates and projection modules, aligning non-textual features with the LLM token space and recommendation objectives through instruction tuning. To mitigate modality imbalance, a multimodal weight dynamic allocation strategy based on gating mechanism is designed, enabling adaptive adjustment of modality weights by combining mixture-of-experts projection and gating network. In addition, the algorithm requires training only the projection module and gating network, without updating the LLM itself, thereby significantly reducing training costs. Extensive experiments on multiple datasets demonstrate the effectiveness of the proposed algorithm.

Key words large language model; multimodal recommendation; mixture-of-experts; adaptive modality fusion; modality imbalance

摘 要 多模态推荐方法引入了文本、图像等模态特征丰富物品表示,提升了推荐系统的性能。近年来,大语言模型(large language models, LLMs)凭借其强大的世界知识与推理能力,在推荐领域展现出巨大潜力。然而,如何充分结合大语言模型与多模态信息,构建基于大语言模型的多模态推荐系统仍面临诸多挑战。首先,大语言模型的文本输入与多模态信息之间存在显著语义鸿沟,导致多模态特征无法被模

收稿日期: 2025-11-24; **修回日期**: 2026-04-22
基金项目: 新一代人工智能国家科技重大专项(2023ZD0121103); 国家自然科学基金项目(U23B2031, 62402159); 中央高校基本科研业务费专项资金(JZ2025HGPB0248, PA2025IISL0105, PA2025IISL0108)
This work was supported by the National Science and Technology Major Project (2023ZD0121103), the National Natural Science Foundation of China (U23B2031, 62402159), and the Fundamental Research Funds for the Central Universities of China (JZ2025HGPB0248, PA2025IISL0105, PA2025IISL0108).
通信作者: 侯旻(hmhoumin@gmail.com)

型充分理解;其次,不同模态间存在信息不平衡问题,对多模态信息的简单融合可能会造成推荐性能的下降。为此,提出基于大语言模型的多模态自适应融合推荐算法,设计多模态混合提示模板与投影模块,通过指令微调实现非文本模态特征与大语言模型词元空间及推荐任务对齐。针对模态不平衡问题,提出基于门控机制的多模态权重动态分配策略,结合混合专家投影与门控网络自适应调整各模态特征权重。此外,该算法仅训练投影模块和门控网络,无需更新大语言模型,显著降低了训练成本。在多个数据集上的实验验证了该算法的有效性。

关键词 大语言模型;多模态推荐;混合专家;自适应融合;模态不平衡

中图法分类号 TP39

DOI: 10.7544/issn1000-1239.202550836 **CSTR:** 32373.14.issn1000-1239.202550836

随着互联网技术的迅猛发展,信息爆炸时代已然到来。在这一背景下,推荐系统应运而生,成为各类以用户为中心的互联网应用的核心组成部分。推荐系统通过学习用户历史行为数据建模用户偏好,对未来时刻用户可能感兴趣的物品进行精准预测,从而有效缓解了信息过载问题。传统的推荐方法将历史交互记录中的物品表示为ID,即仅建模协同过滤信息,难以全面捕捉用户的多元兴趣和行为模式。为此,近年来许多多模态推荐方法[1-4]通过引入物品的多模态信息(如标题文本、图片、视频)来丰富物品表示。这类方法增强了模型对物品的理解,实现了对用户偏好的精准刻画。

自21世纪20年代初以来,大语言模型(large language models, LLMs)引发了持续性的技术革新,其影响力与日俱增。凭借其丰富的世界知识和强大的推理能力,大语言模型为推荐系统领域的发展[5-10]带来了前所未有的机遇。当前,主流的基于大语言模型的推荐系统研究工作的核心是将推荐任务重塑为语言生成任务,即将用户的历史交互记录转换成文本形式的提示模板来作为输入,将大语言模型作为推荐系统生成未来用户可能会购买的物品。提示模板通常采用如下格式:“某用户曾经购买过$item_1$, $item_2$, …, $item_n$,请预测该用户未来可能购买的物品。”在这种范式下,如何表征提示模板中的物品(即$item_k$)至关重要。现有方法主要分为2类,第1类方法使用纯文本形式表征物品[5-8],例如使用标题或是将其表示成文本形式的ID词元(token)(如:用“5”表示物品“索尼耳机”)。这种方法的优势在于简洁直观,可让推荐任务适配语言模型的输入格式。然而,纯文本表征难以捕捉协同过滤信号、多模态特征(如图像信息)等关键属性,从而限制了模型对物品及用户的全面理解。为克服这一限制,近期研究人员们提出了第2类基于跨模态对齐的方法。该类方法通过基于对比学习或基于距离的损失函数,将传统推荐模型中提取的协同过滤特征或物品图像特征与相对应的文本表征进行对齐,从而赋予大语言模型理解跨模态信息的能力。这类方法将基于大语言模型的推荐范式进行了跨模态扩展,有效地提升了推荐性能。

构建基于大语言模型的多模态推荐系统具有重要的研究价值和广泛的应用前景。然而,该方向仍面临诸多挑战:1)大语言模型的文本输入与物品的多模态信息之间存在显著的语义鸿沟。现有方法通常通过将多模态特征向文本对齐,使大语言模型能够“理解”非文本模态的信息。例如,A-LLMRec[11]通过构建基于最小均方误差的损失函数,将物品的协同过滤表征与其标题文本对齐,使大语言模型能够将对齐后的协同过滤信号作为输入;TEA-GLM[12]则采用基于对比学习的损失函数,实现图节点表征与大语言模型词元的对齐。本质上,这类方法依赖于不同模态间的信息一致性进行跨模态转换。然而,将多模态信息映射至文本空间不可避免地会导致模态特有信息的丢失,这种显式对齐策略更适用于模态间信息一一对应的任务场景,如跨模态问答与检索等。而在推荐系统领域,各模态数据的特有信息尤为关键。如图1所示,物品的文本模态主要刻画其性能参数,图像模态则强调其外观特性,而协同过滤信息则隐含了物品之间的潜在关联。这些多模态信息互相补充,从不同维度共同描绘了物品的全貌。2)此外,对于多模态推荐任务来说,不同模态间存在信息不平衡问题[13-15]。具体来说,用户在购买物品时对不同模态数据的侧重程度不同,例如在购买耳机时,用户更关注图像模态体现的外观特性,而对于另一些物品,用户更在意其跟自己以前购买的电子产品的兼容程度。由于模态信息的不平衡问题,研究表明[13],对多模态特征进行简单的拼接、相加等操作进行融合,在所有模态上优化同一目标会导致弱模

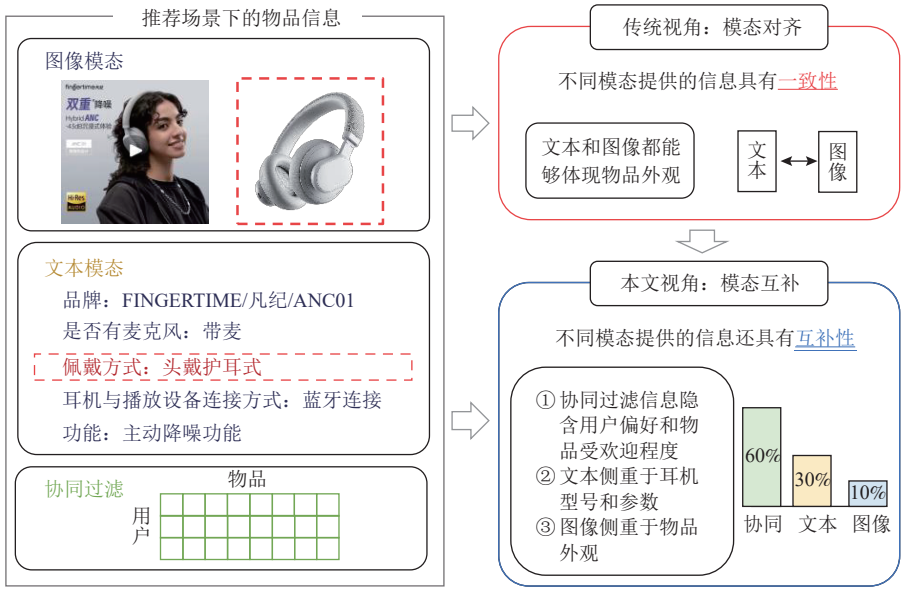


Fig. 1 Schematic diagram of multimodal recommendation scenarios based on large language models

图 1 基于大语言模型的多模态推荐场景示意图

态欠优化，从而导致推荐模型性能下降。

为解决上述挑战，本文提出基于大语言模型的多模态自适应融合推荐算法（LLM-based multimodal recommendation algorithm with adaptive modality fusion，AMFRec）。该算法旨在充分发挥多模态信息间的互补性优势，构建高效的基于大语言模型的多模态推荐系统。具体而言，首先，本文设计了多模态混合提示模板，将物品的文本模态、协同过滤信号、图像模态信息进行融合，作为大语言模型的输入。为此，本文利用预训练的传统推荐模型（如 SASRec[16]）提取物品的协同过滤特征，并采用图像表征模型（例如 AlexNet[17]）抽取图像特征。针对不同的模态信息，本文构建了投影模块（如可训练的线性投影矩阵），将非文本模态信息投影至大语言模型词元空间中。通过在推荐领域指令微调数据上训练投影模块，使得非文本模态信息不仅能够和大语言模型词元空间对齐，还能与推荐任务紧密结合，从而保留推荐任务所需的互补信息，避免由于显式地将非文本模态转化为文本而导致的信息损失。

在此基础上，针对模态信息不平衡问题，本文提出基于门控机制的多模态权重动态分配策略。通过引入混合专家投影机制和门控网络，动态计算不同模态特征在融合过程中的权重比例，实现多模态特征的自适应融合与用户模态偏好的个性化建模，增强了推荐系统在根据用户对模态信息关注点差异时的适应性。此外，在模型训练过程中，采用专家负载均衡策略以防止模型过度依赖单一模态信息，同时引入 dropout 策略鼓励各模态信息的均衡优化，缓解由信息不平衡引发的弱模态欠优化问题。值得注意的是，本文提出的算法仅需训练投影模块和门控网络，而无需对大语言模型进行更新，从而大幅降低了模型训练和推理成本，提高了系统在实际应用中的适应性。

本文提出算法的贡献主要包括：

1）提出了一种基于大语言模型的多模态自适应融合推荐算法，设计了多模态混合提示模板及投影模块，将物品多模态特征进行有效融合，不仅避免了显式文本转换带来的信息损失，而且实现了与推荐任务的深度结合。

2）提出了基于门控机制的多模态权重动态分配策略，引入混合专家投影机制与门控网络，自适应分配多模态特征融合过程中的权重，精准刻画用户对不同模态信息的偏好。此外，该算法仅需训练投影模块与门控网络，无需对大语言模型本体进行更新，从而大幅降低了训练与推理成本。

3）通过在多个推荐数据集上的广泛实验，验证了所提算法在热启动、冷启动等多种推荐场景下的有效性。

1 相关工作

1.1 多模态推荐方法

多模态推荐系统[1-4]利用物品多模态特征（如图像、声音等）中所蕴含的丰富语义信息，将从预训练

模型中提取的物品多模态表征纳入经典的协同过滤推荐范式，缓解了协同过滤推荐模型受限于数据稀疏的问题。早期方法(如 VBPR[1])通过将 ID 表征与从物品图像中提取的视觉表征融合，取得了良好的推荐性能。受图卷积网络最新进展的启发，一些多模态推荐方法通过图卷积网络[2,4,18-19]使用信息传播机制将高阶语义注入用户和物品表示学习中。例如，在微视频推荐任务中，MMGCN[2]为每个模态构建用户-物品二分图，并通过图卷积聚合邻居节点的表征来丰富物品表征。GRCN[18]通过识别并修剪假阳性噪声边来完善用户-物品二分图的结构。LATTICE[19]显式构建了每种模态的物品-物品关系图，并设计模态感知的图结构学习层从多模态特征中学习物品表示。LGMRec[4]利用具有局部拓扑信息的局部嵌入和具有超图依赖的全局嵌入进行建模。其中，局部图嵌入模块学习与协同相关和模态相关的局部用户兴趣，全局超图嵌入模块用于挖掘全局用户兴趣。

还有一些方法专注于挖掘用户对特定模态的细粒度偏好。例如，SLMRec[20]设计 3 种多模态数据增强策略以生成单个物品的多个视图，通过对比学习将同一物品的不同模态的表征进行对齐。BM3[3]使用 dropout 机制增强用户和物品表示的潜在对比视图，不需要负样本，降低了计算成本。由于不同模态的特征在表示上存在明显差距，多模态特征对齐方法[21-22]将不同模态的特征对齐到统一的空间进行融合。FETTLE[21]根据估计的用户反馈自动确定每对模态之间的物品级对齐方向，从用户和物品的角度实现集群级对齐，以实现更稳定的推荐。AlignRec[22]设计了 3 个对齐目标，包括视觉和文本模态之间的对齐、多模态表示和基于 ID 表示之间的对齐、交互的用户和物品之间的对齐，弥补了在多模态推荐中多模态内容特征和 ID 特征之间的差距。

虽然上述模型均取得了良好的推荐效果，但多模态特征信息不平衡[13-15]依然是悬而未决的问题。

1.2 基于大语言模型的推荐方法

大语言模型具备强大的知识表达能力和推理能力，为推荐系统的发展带来了全新的研究范式——生成式推荐[23]，在学术界和工业界引起了广泛关注。传统推荐模型[24-27]大多为判别式推荐，计算用户和物品的相似度，找出与用户最相似的物品来作为用户可能交互的物品。序列推荐[16,28-30]将用户的交互序列作为输入，学习序列中物品转换模式，挖掘用户偏好演变，输出用户接下来交互的物品表示。而基于大语言模型的推荐方法通常包括以下 2 种范式：第 1 种是将大语言模型作为增强器，利用大语言模型中丰富的预训练知识来增强推荐系统中用户和物品的表示(比如 LLMRec[31]，RLMRec[32]，SeRALM[33])。第 2 种是将大语言模型作为推荐器(比如 P5[5]，POD[6]，TALLRec[7]，ReLLa[8]，LLM-TRSR[9]，CLLM4Rec[10]，RecCocktail[34]，WeaveRec[35])，将用户的历史交互记录转换为文本指令，对大语言模型进行指令微调。为了进一步提升大语言模型的推荐性能，一些方法将协同信息输入到大语言模型中，比如 LLaRA[36]和 A-LLMRec[11]。

当前，基于大语言模型的多模态推荐方法将预训练图像编码器提取的物品图像特征输入到大语言模型中，使得大语言模型充分理解物品的特征，为用户提供精准的推荐。VIP5[37]在 P5[5]推荐范式下，同时考虑视觉和文本 2 种模态，统一各种模态和推荐任务。NoteLLM-2[38]利用大语言模型的上下文学习能力将笔记内容压缩为 2 个不同的模态压缩词并利用多模态门控融合机制将视觉编码器的视觉表示和大语言模型的表示结合，生成最终的多模态表示。MLLM-MSR[39]利用基于多模态大语言模型的物品总结器提取给定物品的图像特征，并采用循环用户偏好总结生成范式，捕捉用户偏好的动态变化。LargeTR[40]利用大语言模型强大的语义理解能力过滤多模态数据中的噪声，对多模态用户偏好进行建模，实现可信多模态推荐。DPRec[41]利用大语言模型进行多模态数据特征表示，并同时考虑记忆与扩展的双视角编码，利用超图与双视角编码得到的模态嵌入进行多级精准跨模态对齐，从而在模态融合层实现精确融合。

然而，这些方法并未全面考虑物品的文本、图像特征和协同信息这 3 种模态信息。协同信息和图像特征的有效性不同，且都能提升推荐系统性能。为此，本文提出基于大语言模型的多模态自适应融合推荐算法，将标题文本、协同信息和图像特征联合嵌入到大语言模型的词元空间中，使大语言模型充分理解用户偏好和物品属性，提升多模态推荐的预测准确性。

2 问题定义

2.1 序列推荐

给定用户 u 的历史交互 $S_u=\{item_1,item_2,\cdots,item_n\}$，其中 $item_j$ 是用户在第 j 步交互的物品。序列推荐的主要任务是根据用户的历史交互序列预测与用户偏

好对齐的下一个物品。在基于大语言模型的序列推荐中，输入用户的历史交互序列，通过在候选集中输出潜在的下一个物品$item_{n+1}$的概率分布来训练序列推荐模型，使其最大化下一个真实交互物品的概率。

2.2 多模态序列推荐

给定用户u、用户的历史交互序列$S_u=\{item_1, item_2, \cdots, item_n\}$和候选物品$C=\{item'_1, item'_2, \cdots, item'_m\}$，每个物品包含物品的文本数据$item_i^{\text{title}}$、协同过滤信息$item_i^{\text{cf}}$和图像数据$item_i^{\text{img}}$。多模态序列推荐的目标是根据用户的历史交互序列和物品的多模态特征，从候选集C中预测用户下一个交互物品$item_{n+1}$。本文提出基于大语言模型的多模态自适应融合推荐算法，将不同模态的特征统一嵌入到大语言模型中，并利用多模态权重动态分配策略有效结合多模态特征。具体算法在第3节中介绍。

3 多模态自适应融合推荐算法

3.1 整体过程

AMFRec的整体过程如图2所示。首先，基于多模态自适应融合编码提取物品的多模态特征。具体而言，文本模态的物品描述按照大语言模型的词表进行词向量映射；交互模态的物品协同信息通过协同过滤推荐模型进行提取；物品的图像特征使用预训练图像编码器进行提取。随后，将不同模态的特征统一映射到大语言模型的词元空间中，使用基于门控机制的多模态权重动态分配策略进行多模态特征自适应融合。融合后的多模态特征作为大语言模型的输入，进行推荐任务微调。特别地，在多模态自适应融合编码模块中，通过设计混合专家投影机制，将物品的协同信息和图像特征分别通过不同的专家网络对齐到大语言模型的词元空间，减少多模态特征与大语言模型之间的差距；并设计多模态权重动态分配策略，通过门控网络平衡不同模态特征的作用，将多模态特征进行有效结合。在微调期间预训练模型保持不变，只需训练投影模块与门控网络，提高训练和推理效率。

3.2 多模态特征提取

本算法首先提取物品的多模态特征。针对文本特征来说，其包含物品的标题或描述，能够直观体现物品的文本属性。大语言模型在预训练期间获得了文本理解能力，能够通过文本特征理解物品属性，并利用大语言模型的通用知识增强对文本表示的建模。

对于物品的文本元数据，文本表示的提取方式为

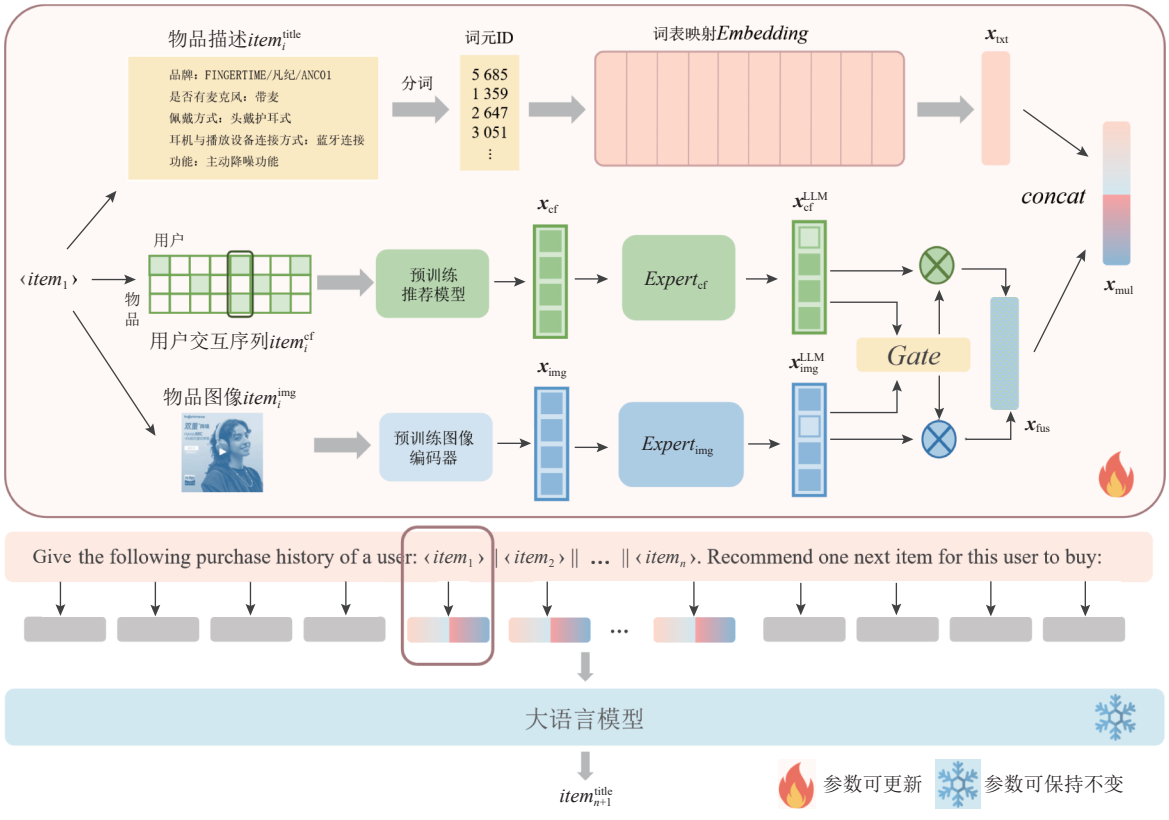


Fig. 2 Framework of AMFRec

图2 AMFRec框架

$$token_i = Tokenizer(item_i^{title}), \quad (1)$$

其中$Tokenizer(\cdot)$表示大语言模型的分词器，$token_i$为物品标题对应的词元序列。再通过大语言模型的嵌入层将词元序列映射成词向量：

$$\boldsymbol{x}_{txt} = Embedding(token_i), \quad (2)$$

其中$Embedding(\cdot)$表示大语言模型的嵌入层，$\boldsymbol{x}_{txt}$为映射成的词向量，作为物品的文本特征，并与大语言模型的词元空间兼容。

协同信息能够体现用户的交互行为，根据用户的历史交互预测用户的偏好。传统的序列推荐模型[16,28-30]能够根据用户的历史交互，训练得到基于ID的物品行为表示。为此，本文使用传统的推荐模型提取协同信息，然后将得到的协同信息对齐到大语言模型的词元空间中。其中，协同信息的提取方式为

$$\boldsymbol{x}_{cf} = BehaviorEncoder(item_i^{cf}), \quad (3)$$

其中$item_i^{cf}$表示物品的ID，$BehaviorEncoder(\cdot)$表示预训练协同过滤推荐模型，$\boldsymbol{x}_{cf}$表示物品的协同信息。

除了物品的协同信息外，还需提取物品的图像特征来构建物品的多模态表示。使用预训练图像编码器提取物品的图像特征：

$$\boldsymbol{x}_{img} = ImageEncoder(item_i^{img}), \quad (4)$$

其中$item_i^{img}$表示物品的图像，$ImageEncoder(\cdot)$表示预训练图像编码器，$\boldsymbol{x}_{img}$表示物品的图像特征。

3.3 基于门控机制的多模态权重动态分配策略

在得到物品的多模态表示后，将其投影到大语言模型的词元空间，并设计推荐任务指令微调数据集来微调投影矩阵。物品的文本标题、协同信息和图像特征共同构成多模态混合指令模板。需要说明的是，本文的门控融合机制仅作用于协同信息与图像特征，而未将文本模态纳入门控网络。这一设计基于以下考虑：1)文本特征通过大语言模型分词器和词嵌入层直接映射为词向量，已天然位于大语言模型的词元空间中，无需额外的投影对齐；而协同过滤特征和图像特征作为外部异构特征，需要通过专家投影模块映射到大语言模型空间，二者在对齐过程中可能存在贡献度差异，因此需要门控机制进行动态权重调节。2)在基于大语言模型的推荐范式中，文本是用户理解物品的核心载体，本文将其作为语义锚点，协同信息和图像特征则作为补充信息增强物品表示，这与现有研究A-LLMRec[11]，VIP5[37]的设计思路一致。

为了使得协同信息和图像特征有效融合，先将物品的协同信息和图像特征通过模态混合专家投影对齐到大语言模型的词元空间中，再利用基于门控机制的多模态权重动态分配策略，通过门控网络动态计算不同模态特征在融合过程中的权重比例，实现多模态特征的自适应融合与用户模态偏好的个性化建模，增强推荐系统在根据用户对不同模态信息关注点差异时的适应性。物品的多模态特征自适应融合过程如图2所示，其公式如下：

$$\begin{aligned}
&\boldsymbol{x}_{cf}^{LLM} = Expert_{cf}(\boldsymbol{x}_{cf}),\\
&\boldsymbol{x}_{img}^{LLM} = Expert_{img}(\boldsymbol{x}_{img}),\\
&g = Gate\left([\boldsymbol{x}_{cf}^{LLM};\boldsymbol{x}_{img}^{LLM}]\right),\\
&[a,b] = Softmax(g),\\
&\boldsymbol{x}_{fus} = a\cdot\boldsymbol{x}_{cf}^{LLM} + b\cdot\boldsymbol{x}_{img}^{LLM},
\end{aligned} \quad (5)$$

其中协同过滤专家网络$Expert_{cf}(\cdot)$将物品的协同信息$\boldsymbol{x}_{cf}$投影到大语言模型的词元空间得到$\boldsymbol{x}_{cf}^{LLM}$，图像专家网络$Expert_{img}(\cdot)$将物品的图像特征$\boldsymbol{x}_{img}$投影到大语言模型的词元空间得到$\boldsymbol{x}_{img}^{LLM}$，然后将投影后的特征通过门控网络$Gate(\cdot)$，并经过$Softmax(\cdot)$计算2种模态的权重$a$，$b$，这里门控网络为单层感知机。最后将投影后的特征进行自适应融合得到协同信息和图像特征的融合表示$\boldsymbol{x}_{fus}$。

在得到物品的融合表示后，与文本标题结合，构成物品的多模态表示：

$$\boldsymbol{x}_{mul} = concat(\boldsymbol{x}_{txt}, \boldsymbol{x}_{fus}), \quad (6)$$

其中$concat(\cdot,\cdot)$表示将2种特征进行拼接，实现对物品的全面描述。

将得到的物品多模态特征集成到指令模板中，对大语言模型进行指令微调，如图3所示。指令模

指令输入： You are a helpful recommendation assistant. Given the following purchase history of a user: **[HISTORY (Item Titles, Item Embs)]**. Recommend one next item for this user to buy next from the following item title set: **[CANDIDATE (Item Titles, Item Embs)]**. The recommendation is
指令输出： **[Next Item Title]**

Fig. 3 Instruction prompt template

图3 指令提示模板

板的输入包含了指令任务(要执行的多模态序列推荐任务)、用户的历史交互序列以及候选物品。根据指令提示对大语言模型进行微调，并根据用户的历史交互序列从候选物品中挑选出用户接下来要交互的物品。大语言模型的输出为下一个物品的标题。

3.4 参数高效微调

由于时间和成本的限制，微调大语言模型的所有参数是非常耗时的。受参数高效微调方法[42]的启发，模板中引入可以训练的连续提示，通过对部分参数进行微调，提高参数微调效率并取得相当的性能。在微调过程中，只微调多模态自适应融合编码模块，协同过滤推荐模型、预训练图像编码器和大语言模型的参数保持不变。此外，由于不同模态特征在训练过程中收敛时间存在差异[13]，会对推荐性能产生影响。鉴于此，对专家网络施加 dropout 正则化，以保证模态训练平衡。具体的模态平衡分析实验见 4.4 节。

损失函数采用语言模型中常见的方式[5-6,11]，根据输入的文本指令直接输出用户接下来可能交互的物品，将推荐问题转化为条件文本生成问题：

$$L=-\sum_{(x,y)}\sum_{i=1}^{|y|}\ln P_\theta(y_i|x,y_{<i})+\lambda\cdot\delta\,, \tag{7}$$

其中x表示输入的提示，y_i表示输出的第i个词元，$y_{<i}$表示第i个词元前面的词元。θ表示需要训练的参数。δ为负载平衡损失，λ为系数，确保不同专家网络之间的负载均衡。其中$\delta=\dfrac{\sigma}{\mu}$，$\sigma$为门控权重的标准差，$\mu$为门控权重的均值。

4 实　　验

4.1 实验设置

1)数据集

本文使用的数据均来自亚马逊评论数据集①，包括 Beauty, Toys_and_Games, Baby 这 3 个类别。其中图像特征来自预处理的 4 096 维特征向量。所有数据集均为 5 次评分数据集，以保证每个用户至少包含 5 次交互。数据集中包含的用户数、物品数和交互次数如表 1 所示。遵循大语言模型在序列推荐中的常见方式[5-6,11]，将数据集划分为训练集、验证集和测试集。设用户的交互序列长度为 n，前$n-2$个物品用于训练，第$n-1$个物品用于验证，最后 1 个物品用于测试。

Table 1　Statistics of Amazon Datasets

表 1　亚马逊数据集统计

数据集	用户数	物品数	交互次数
Beauty	22 363	12 101	198 502
Toys_and_Games	19 412	11 924	167 597
Baby	19 445	7 050	160 792

2)实现细节

实验采用 Qwen2-7B-Instruct 作为基础模型，具体实现细节如下：基于预训练模型 SASRec[16] 提取物品的协同信息。协同信息的维度为 50，通过专家投影将特征映射到维度为 3 584 的大语言模型词元空间。对于物品的图像特征，使用 AlexNet[17] 提取的 4 096 维的图像特征，经过专家投影对齐到大语言模型的词元空间。对于物品的文本特征，根据大语言模型的词表将文本标题映射成词向量。随机初始化投影层的参数并使用 ReLU 作为激活函数。基于 Adam 优化器进行训练，学习率大小设置为 1E－4，λ设为 0.1。利用广泛使用的评估指标命中率(hit ratio, HR)来评估算法的有效性。

3)对比方法

将提出的 AMFRec 与多模态推荐方法和大语言模型推荐方法进行比较。为保证公平比较，采用统一的多模态推荐框架 MMRec[43] 测试多模态推荐方法。多模态推荐方法包括 VBPR[1], MMGCN[2], BM3[3], LGMRec[4]。大语言模型推荐方法 A-LLMRec[11] 利用文本标题和协同信息共同构成物品的表示。协同信息来自于协同过滤推荐模型得到的物品表示，并将其投影到大语言模型的词元空间，与文本标题的词嵌入共同构成物品的表示。使用的协同过滤推荐模型为 SASRec[16]，为使比较方法更具说服力，同时将 SASRec 作为一种对比方法。

4.2 对比实验

整体的对比实验结果如表 2 所示。AMFRec 的 HR@i 指标明显优于常见的多模态推荐方法，包括 VBPR, MMGCN, BM3, LGMRec。大语言模型在预训练期间获得了大量的通用知识，具备更强大的多模态特征的融合能力和理解能力，能够理解不同类型的数据(文本特征、协同信息和图像特征)。AMFRec 也优于大语言模型推荐方法 A-LLMRec。A-LLMRec 使用文本标题和协同信息来表示物品，但没有考虑物品的图像特征以及不同模态特征之间的有效性差

① https://cseweb.ucsd.edu/~jmcauley/datasets/amazon/links.html

Table 2 Comparative Experiments

表 2 对比实验

数据集	评价指标	VBPR	MMGCN	BM3	LGMRec	SASRec	A-LLMRec	AMFRec（本文）
Beauty	HR@1	0.313 4	0.297 5	0.336 1	0.357 1	0.351 3	0.330 7	0.417 3
	HR@3	0.511 8	0.497 9	0.545 1	0.557 8	0.556 6	0.530 5	0.618 9
Toys_and_Games	HR@1	0.270 0	0.263 6	0.288 1	0.332 2	0.346 6	0.322 9	0.407 8
	HR@3	0.463 0	0.470 5	0.487 3	0.517 8	0.549 4	0.538 7	0.611 8
Baby	HR@1	0.213 2	0.204 1	0.227 3	0.233 1	0.257 1	0.203 0	0.270 7
	HR@3	0.413 3	0.400 8	0.417 3	0.423 5	0.483 0	0.409 1	0.503 4

异。图像特征能够直观地展示物品的外观、颜色、纹理等视觉信息，而协同信息能够体现用户的交互行为。AMFRec 同时使用了物品的文本特征、协同信息和图像特征，并且多模态权重动态分配策略通过门控权重分配机制，平衡不同模态特征之间的作用。这有助于多模态特征的有效结合，构建更加全面的物品表示，提供更准确的推荐。此外，大语言模型推荐方法中使用的协同信息均来自于预训练模型 SASRec 得到的物品表示，为使比较方法更具说服力，也将 SASRec 作为一种对比方法。AMFRec 通过基于门控机制的多模态权重动态分配策略将多模态特征进行有效结合，其中通过门控网络计算不同模态特征的权重，平衡不同模态特征的关系。这有助于构建更加全面的多模态表示，增强推荐系统对物品的理解，生成更符合用户期望的推荐。

4.3 消融实验

为了验证 AMFRec 中不同成分的作用，进行了 4 种不同的实验设置，如表 3 所示。第 1 种设置 [Title][CF] 包含物品的文本标题和协同信息。第 2 种设置 [Title][Image] 包含物品的文本标题和图像特征。第 3 种设置 [Title][CF][Image] 包含物品的文本标题、协同信息和图像特征，并将这 3 种特征进行串联拼接。第 4 种设置即本文提出的 AMFRec，将物品的协同信息与图像特征通过多模态自适应融合编码模块进行结合，再拼接到标题的后面。从实验结果可以看出，仅在标题后面增加图像特征的方法（对应第 2 种设置）并没有取得最好的推荐性能，并且将协同信息与图像特征直接拼到标题后面的方法（对应第 3 种设置）也不如直接在标题后面拼接协同信息（对应第 1 种设置）。第 3 种设置直接将文本特征、协同信息和图像特征进行拼接，没有考虑到不同模态特征之间的有效性差异，无法平衡多模态特征之间的作用。协同信息和图像特征来自于不同的编码器，并且在维度和语义上存在显著差距。物品的协同信息能够根据用户的交互行为捕获用户的行为偏好，物品的图像特征能够直观反映物品的外观和形状并捕获用户的视觉偏好，结合这 2 种不同模态的特征需要考虑多模态特征之间的有效性差异。AMFRec 将文本、图像和协同信息统一嵌入到大语言模型的词元空间中，并使用不同的专家网络分别处理不同模态的特征，减小多模态特征与大语言模型之间的差距。同时，使用基于门控机制的多模态权重动态分配策略计算协同信息和图像特征的权重，平衡不同模态特征的作用，并将多模态特征有效融合，进一步提升了多模态推荐性能。

Table 3 Ablation Experiments

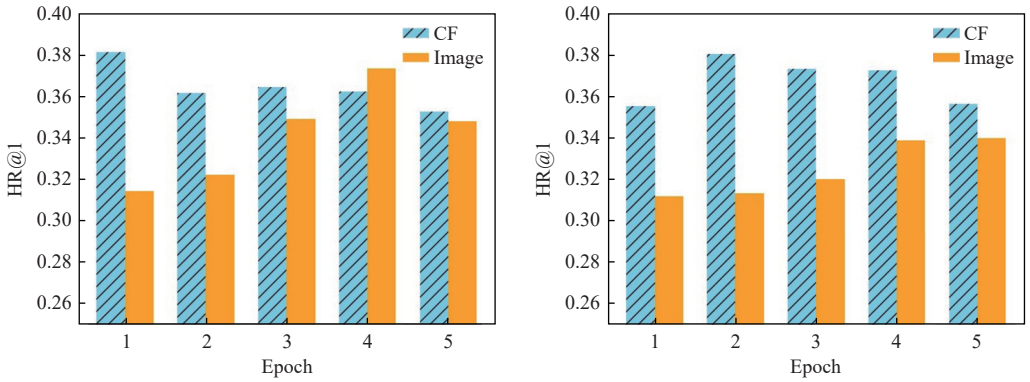
表 3 消融实验

数据集	评价指标	实验设置			
		[Title][CF]	[Title][Image]	[Title][CF][Image]	AMFRec（本文）
Beauty	HR@1	0.364 4	0.348 7	0.359 2	0.417 3
	HR@3	0.566 4	0.532 8	0.567 6	0.618 9
Toys_and_Games	HR@1	0.372 8	0.319 9	0.356 8	0.407 8
	HR@3	0.579 8	0.506 8	0.548 1	0.611 8
Baby	HR@1	0.263 1	0.182 2	0.254 5	0.270 7
	HR@3	0.504 4	0.365 2	0.479 0	0.503 4

4.4 模态不平衡分析实验

协同信息和图像特征属于不同模态的特征，存在模态维度和语义上的差距。协同信息来自协同过滤推荐模型 SASRec，图像特征来自预训练的深度卷积神经网络。由于模态之间的差距，协同信息和图像特征存在明显的模态不平衡问题[13]，即这 2 种不同模态的特征收敛时间不一致。这种模态不平衡现象使得无法有效利用多模态特征，严重影响多模态推荐性能。本文首先定量验证了这种普遍存在的模态不平衡现象，然后通过对专家投影网络施加 dropout 正则化缓解这一问题。在 Beauty 和 Toys_and_Games 数据集上进行单模态实验验证，单独训练每种模态

特征并观察其收敛速度，设计了 2 种不同的实验设置。第 1 种设置对应 4.3 节消融实验中的第 1 种设置，可训练的参数仅为协同过滤专家网络。第 2 种设置对应 4.3 节消融实验中的第 2 种设置，可训练的参数仅为图像专家网络。在这 2 种设置下的实验结果如图 4 所示，用 HR@1 来评估模型的推荐性能。从实验结果可以看出，协同信息收敛较快，收敛后出现过拟合现象，之后性能开始下降。而图像特征收敛较慢，推荐性能缓慢提升。因此这 2 种特征的训练存在明显的模态不平衡问题。



(a) Beauty (b) Toys_and_Games

Fig. 4 Performance analysis of single-modality recommendation

图 4 单模态推荐性能分析

为了获得更好的推荐性能，需保证协同信息和图像特征在训练过程同时达到最好的推荐性能，再通过混合专家网络将这 2 种特征进行结合。因此，在协同信息投影网络中增加 dropout 来缓解协同信息过拟合。分别设置 dropout 的参数为 0.3, 0.5, 0.7, 验证在这 3 种条件下的推荐性能。实验结果如表 4 所示，可以看出，依次增加 dropout 的大小，模型在 Beauty 和 Toys_and_Games 数据集上的性能逐渐提高，表明协同信息存在明显的过拟合问题。对协同信息投影网络施加 dropout 正则化，可以缓解协同信息的过拟合问题。协同信息投影网络中的 dropout 设置为 0.7 时性能达到最优。

Table 4 Modality Imbalance Analysis Experiments

表 4 模态不平衡分析实验

数据集	评价指标	dropout			
		0	0.3	0.5	0.7
Beauty	HR@1	0.366 3	0.383 1	0.387 0	0.417 3
	HR@3	0.560 7	0.570 0	0.582 5	0.618 9
Toys_and_Games	HR@1	0.356 2	0.370 3	0.388 3	0.407 8
	HR@3	0.556 4	0.572 5	0.589 0	0.611 8

4.5 冷启动实验

在推荐系统中，当用户的交互记录很少或者引入新用户时，难以提供准确的推荐。因此，物品的多模态特征成为冷启动推荐的关键信息。由于在预训练期间获得了丰富的通用知识，大语言模型可以很好地应用在冷启动推荐场景中。为了进一步验证 AMFRec 在冷启动推荐场景下的有效性，选择有 3 个交互记录的新用户作为冷启动用户，其中用户交互的最后一个物品用来测试。测试方法包括 VBPR, MMGCN, BM3, LGMRec, A-LLMRec, AMFRec, 如表 5 所示。从实验结果可以看出，在冷启动推荐场景下，AMFRec 显著优于其他方法，这源于大语言模型在预训练期间获得了丰富的通用知识，这些知识可以增强对物品表示的理解。即使对未在训练集中出现的交互序列，AMFRec 也能利用上下文学习能力理解用户意图，生成准确的推荐。

Table 5 Cold Start Experiments

表 5 冷启动实验

数据集	评价指标	VBPR	MMGCN	BM3	LGMRec	A-LLMRec	AMFRec（本文）
Beauty	HR@1	0.098 4	0.108 3	0.096 6	0.070 6	0.338 3	0.455 3
	HR@3	0.228 1	0.252 5	0.212 7	0.177 5	0.548 3	0.642 4
Toys_and_Games	HR@1	0.089 2	0.096 8	0.080 5	0.081 1	0.402 1	0.464 4
	HR@3	0.211 6	0.233 8	0.184 1	0.189 2	0.609 2	0.656 8
Baby	HR@1	0.162 2	0.153 2	0.145 0	0.130 5	0.231 1	0.297 6
	HR@3	0.341 6	0.323 6	0.298 1	0.288 8	0.455 2	0.531 9

4.6 门控网络权重分析

为验证多模态自适应融合编码模块如何对协同信息和图像特征进行有效的特征选择，对每个专家网络的权重进行分析。在训练集上进行训练，分别在所有物品、冷物品（与用户交互次数比较少的物品）

和热物品(与用户交互次数比较多的物品)上计算门控网络中协同信息和图像特征的门控权重，如表 6 所示。从实验结果可以看出，在所有物品的计算中，Beauty 和 Baby 数据集中协同信息和图像特征的权重几乎相同。热物品与用户交互的次数比较多，协同信息在门控网络中计算得到的权重偏大，图像特征计算的权重偏小。这表明协同信息在推荐中起到主要作用，图像特征在推荐中起到辅助作用。而冷物品与用户交互次数比较少，协同信息的权重偏小，图像特征的权重偏大。这表明图像特征起到主要作用，协同信息起到辅助作用。对于 Toys_and_Games 数据集，从整体来看，协同信息的权重偏小，图像特征的权重偏大。而在热物品中，协同信息的权重略有增加，图像特征的权重略有减小。在冷物品中，协同信息的权重略有减少，图像特征的权重略有增加。由此可以看出，对于不同的数据集，门控网络通过调整不同模态特征的权重进行有效的特征融合。对于在不同场景下的相同数据集，门控网络中的权重也可以进行动态调整，计算出合适的不同模态特征的融合比例来适应不同场景的需求，进行准确的推荐。

Table 6 Gate Control Weight Analysis

表 6 门控权重分析

数据集	所有物品		热物品		冷物品	
	协同信息	图像特征	协同信息	图像特征	协同信息	图像特征
Beauy	0.498 9	0.501 1	0.505 0	0.495 0	0.497 1	0.502 9
Toys_and_Games	0.440 4	0.559 6	0.461 6	0.538 4	0.432 7	0.567 3
Baby	0.500 7	0.499 3	0.520 1	0.479 9	0.489 7	0.510 3

4.7 不同大语言模型推荐性能比较

为比较 AMFRec 在不同类型大语言模型上的性能差异，将多模态大语言模型 Qwen2-VL-7B-Instruct 以及纯文本大语言模型 Qwen2-7B-Instruct 和 Opt-6.7b 的实验结果列在表 7 中。从实验结果可以看出，多模态大语言模型 Qwen2-VL-7B-Instruct 取得了最好的推荐性能，优于纯文本大语言模型 Qwen2-7B-Instruct 和 Opt-6.7b。多模态大语言模型在预训练期间获得了图像相关的知识，具备了图像理解能力，可以更好地应用于多模态推荐中。而纯文本大语言模型不能充分理解图像特征，取得的推荐性能不如多模态大语言模型。因此，多模态大语言模型的跨模态理解能力使得其能够在多模态场景中拥有广泛应用，并在多模态推荐领域也能取得很好的推荐性能。

Table 7 Performance Comparison of Different LLMs

表 7 不同大语言模型性能比较

大语言模型	Beauty		Toys_and_Games		Baby	
	HR@1	HR@3	HR@1	HR@3	HR@1	HR@3
Opt-6.7b	0.366 4	0.583 2	0.372 9	0.580 8	0.277 0	0.460 0
Qwen2-7B-Instruct	0.417 3	0.618 9	0.407 8	**0.611 8**	0.270 7	0.503 4
Qwen2-VL-7B-Instruct	**0.425 0**	**0.628 9**	**0.417 7**	0.608 9	**0.294 3**	**0.525 0**

注：黑体数值表示最优值。

4.8 案例研究

为了进一步表明不同物品多模态特征的权重分配差异，图 5 列举一个用户的交互物品序列，并分别展示每个物品对应的协同信息和图像特征的权重分配情况。从图 5 可以看出，与该用户交互的第 1 个、第 4 个、第 5 个和第 6 个物品的图像特征的权重分数高于协同信息，表明这 4 个物品的图像特征比协同信息起到更重要的作用。而第 2 个和第 3 个物品中的协同信息的权重分数高于图像特征，表明协同信息占有主要的作用，图像特征起到辅助作用。多模态权重动态分配策略为每个物品的多模态特征分配不同的权重，并自适应融合不同模态的特征，进一步提升多模态推荐性能。

5 结 论

大语言模型凭借其在预训练期间获得的丰富知识，提升对多模态特征的理解，充分感知用户意图并提升个性化推荐性能。针对现有模型未考虑模态互补性和不同模态之间的有效性差异，本文提出基于大语言模型的多模态自适应融合推荐算法 AMFRec，



	$user_{73}$	$item_{7670}$	$item_{7}$	$item_{1788}$	$item_{6208}$	$item_{6702}$	$item_{7739}$
协同信息		0.499 4	0.520 2	0.507 7	0.493 7	0.461 6	0.455 8
图像特征		0.500 6	0.479 8	0.492 3	0.506 3	0.538 4	0.544 2

Fig. 5 Multimodal feature weight allocation for different items

图 5 不同物品多模态特征权重分配

将物品的文本特征、协同信息和图像特征统一对齐到大语言模型的词元空间中进行推荐任务微调，充分理解用户偏好和物品属性，进一步提升多模态推荐性能。首先使用协同过滤推荐模型和图像编码器分别提取物品的协同信息和图像特征，然后基于门控机制的多模态权重动态分配策略将多模态特征有效结合，显著提升多模态推荐性能。在训练时，只需对多模态自适应融合编码模块进行微调，无需微调预训练编码器和大语言模型的参数，降低了计算成本。在多个真实世界数据集上的实验证明了 AMFRec 的有效性。

作者贡献声明：肖光义、侯旻提出了算法思路和实验方案，并撰写论文；廖宇欣、武晗、白晨希参与写作与实验，并修改论文；魏思、吴乐提供指导意见。

参 考 文 献

[1] He Ruining, McAuley J. VBPR: Visual Bayesian personalized ranking from implicit feedback[C]//Proc of the 29th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2015: 144−150

[2] Wei Yinwei, Wang Xiang, Nie Liqiang, et al. MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video[C]//Proc of the 27th ACM Int Conf on Multimedia. New York: ACM, 2019: 1437−1445

[3] Zhou Xin, Zhou Hongyu, Liu Yong, et al. Bootstrap latent representations for multi-modal recommendation[C]//Proc of the 32nd ACM Web Conf. New York: ACM, 2023: 845−854

[4] Guo Zhiqiang, Li Jianjun, Li Guohui, et al. LGMRec: Local and global graph learning for multimodal recommendation[C]//Proc of the 38th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2024: 8454−8462

[5] Geng Shijie, Liu Shuchang, Fu Zuohui, et al. Recommendation as language processing (RLP): A unified pretrain, personalized prompt & predict paradigm (P5)[C]//Proc of the 16th ACM Conf on Recommender Systems. New York: ACM, 2022: 299−315

[6] Li Lei, Zhang Yongfeng, Chen Li. Prompt distillation for efficient LLM-based recommendation[C]//Proc of the 32nd ACM Int Conf on Information and Knowledge Management. New York: ACM, 2023: 1348−1357

[7] Bao Keqin, Zhang Jizhi, Zhang Yang, et al. TALLRec: An effective and efficient tuning framework to align large language model with recommendation[C]//Proc of the 17th ACM Conf on Recommender Systems. New York: ACM, 2023: 1007−1014

[8] Lin Jianghao, Shan Rongjie, Zhu Chenxu, et al. ReLLa: Retrieval-enhanced large language models for lifelong sequential behavior comprehension in recommendation[C]//Proc of the 33rd ACM on Web Conf. New York: ACM, 2024: 3497−3508

[9] Zheng Zhi, Chao WenShuo, Qiu Zhaopeng, et al. Harnessing large language models for text-rich sequential recommendation[C]//Proc of the 33rd ACM on Web Conf. New York: ACM, 2024: 3207−3216

[10] Zhu Yaochen, Wu Liang, Guo Qilnli, et al. Collaborative large language model for recommender systems[C]//Proc of the 33rd ACM on Web Conf. New York: ACM, 2024: 3162−3172

[11] Kim S, Kang H, Choi S, et al. Large language models meet collaborative filtering: An efficient all-round LLM-based recommender system[C]//Proc of the 30th ACM SIGKDD Conf on Knowledge Discovery and Data Mining. New York: ACM, 2024: 1395−1406

[12] Wang Duo, Zuo Yuan, Li Fengzhi, et al. LLMs as zero-shot graph learners: Alignment of GNN representations with LLM token embeddings[C]//Proc of the 38th Neural Information Processing Systems. New York: Curran Associates, 2024: 5950−5973

[13] Zhang Jinghao, Liu Guofan, Liu Qiang, et al. Modality-balanced learning for multimedia recommendation[C]//Proc of the 32nd ACM Int Conf on Multimedia. New York: ACM, 2024: 7551−7560

[14] Peng Xiaokang, Wei Yake, Deng Andong, et al. Balanced multimodal learning via on-the-fly gradient modulation[C]//Proc of the 39th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 8228−8237

[15] Wang Weiyao, Tran D, Feiszli M. What makes training multi-modal classification networks hard?[C]//Proc of the 37th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 12692−12702

[16] Hidasi B, Karatzoglou A, Baltrunas L, et al. Session-based recommendations with recurrent neural networks[J]. arXiv preprint, arXiv: 1511.06939, 2015

[17] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2012, 60: 84−90

[18] Wei Yinwei, Wang Xiang, Nie Liqiang, et al. Graph-refined convolutional network for multimedia recommendation with implicit feedback[C]//Proc of the 28th ACM Int Conf on multimedia. New York: ACM, 2020: 3541−3549

[19] Zhang Jinghao, Zhu Yanqiao, Liu Qiang, et al. Mining latent structures for multimedia recommendation[C]//Proc of the 29th ACM Int Conf on Multimedia. New York: ACM, 2021: 3872−3880

[20] Tao Zhulin, Liu Xiaohao, Xia Yewei, et al. Self-supervised learning for multimedia recommendation[J]. IEEE Transactions on Multimedia, 2023, 25: 5107−5116

[21] Li Yang, Zhao Qi'ao, Lin Chenxi, et al. Who to align with: Feedback-oriented multi-modal alignment in recommendation systems[C]//Proc of the 47th Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2024: 667−676

[22] Liu Yifan, Zhang Kangning, Ren Xiangyuan, et al. AlignRec: Aligning and training in multimodal recommendations[C]//Proc of the 33rd ACM Int Conf on Information and Knowledge Management. New York: ACM, 2024: 1503−1512

[23] Hou Min, Wu Le, Liao Yuxin, et al. A survey on generative recommendation: Data, model, and tasks[J]. arXiv preprint, arXiv:

2510.27157, 2025

[24] He Xiangnan, Deng Kuan, Wang Xiang, et al. LightGCN: Simplifying and powering graph convolution network for recommendation[C]// Proc of the 43rd Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2020: 639−648

[25] Wu Jiancan, Wang Xiang, Feng Fuli, et al. Self-supervised graph learning for recommendation[C]//Proc of the 44th Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2021: 726−735

[26] Yu Junliang, Yin Hongzhi, Xia Xin, et al. Are graph augmentations necessary?: Simple graph contrastive learning for recommendation[C]//Proc of the 45th Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2022: 1294−1303

[27] Cai Xuheng, Huang Chao, Xia Lianghao, et al. LightGCL: Simple yet effective graph contrastive learning for recommendation[C/OL]//Proc of the 11th Int Conf of Learning Representation. 2023[2026-01-29]. https://openreview.net/forum?id=FKXVK9dyMM

[28] Kang Wangcheng, McAuley J. Self-attentive sequential recommendation[C]//Proc of the 34th IEEE Int Conf on Data Mining. Piscataway, NJ: IEEE, 2018: 197−206

[29] Sun Fei, Liu Jun, Wu Jian, et al. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer[C]// Proc of the 28th ACM Int Conf on Information and Knowledge Management. New York: ACM, 2019: 1441−1450

[30] Xie Xu, Sun Fei, Liu Zhaoyang, et al. Contrastive learning for sequential recommendation[C]//Proc of the 38th Int Conf on Data Engineering. Piscataway, NJ: IEEE, 2022: 1259−1273

[31] Wei Wei, Ren Xubin, Tang Jiabin, et al. LLMRec: Large language models with graph augmentation for recommendation[C]//Proc of the 17th ACM Int Conf on Web Search and Data Mining. New York: ACM, 2024: 806−815

[32] Ren Xubin, Wei Wei, Xia Lianghao, et al. Representation learning with large language models for recommendation[C]//Proc of the 33rd ACM on Web Conf. New York: ACM, 2024: 3464−3475

[33] Ren Yankun, Chen Zhongde, Yang Xinxing, et al. Enhancing sequential recommenders with augmented knowledge from aligned large language models[C]//Proc of the 47th Annual Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2024: 345−354

[34] Hou Min, Bai Chenxi, Wu Le, et al. RecCocktail: A generalizable and efficient framework for LLM-based recommendation[J]. arXiv preprint, arXiv: 2502.08271, 2025

[35] Hou Min, Liu Xin, Wu Le, et al. WeaveRec: An LLM-based cross-domain sequential recommendation framework with model merging[J]. arXiv preprint, arXiv: 2510.26546, 2025

[36] Liao Jiayi, Li Sihang, Yang Zhengyi, et al. LLaRA: Large language-recommendation assistant[C]//Proc of the 47th Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2024: 1785−1795

[37] Geng Shijie, Tan Juntao, Liu Shuchang, et al. VIP5: Towards multimodal foundation models for recommendation[C]//Proc of the 2023 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2023: 9606−9620

[38] Zhang Chao, Zhang Haoxin, Wu Shiwei, et al. NoteLLM−2: Multimodal large representation models for recommendation[C]// Proc of the 31st ACM SIGKDD Conf on Knowledge Discovery and Data Mining. New York: ACM, 2025: 2815−2826

[39] Ye Yuyang, Zheng Zhi, Shen Yishan, et al. Harnessing multimodal large language models for multimodal sequential recommendation[C]//Proc of the 39th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2025: 13069−13077

[40] Yan Meng, Xu Cai, Huang Haibin, et al. Large language model-based trusted multi-modal recommendation algorithm[J]. Journal of Computer Research and Development, 2025, 62(7): 1611−1621 (in Chinese)
(闫萌, 徐偲, 黄海槟, 等. 基于大语言模型的可信多模态推荐算法[J]. 计算机研究与发展, 2025, 62(7): 1611−1621)

[41] Li Yiming, Yu Yaxin, Yu Zhisheng, et al. Dual-perspective multi-level cross-modal recommendation based on large language models[J]. Journal of Computer Research and Development, 2026, 63(1): 147−161 (in Chinese)
(李佚名, 于亚新, 于之晟, 等. 基于大语言模型的双视角多级跨模态推荐[J]. 计算机研究与发展, 2026, 63(1): 147−161)

[42] Li X L, Liang P. Prefix-tuning: Optimizing continuous prompts for generation[C]//Proc of the 59th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2021: 4582−4597

[43] Zhou Xin. MMRec: Simplifying multimodal recommendation[C]// Proc of the 5th ACM Int Conf on Multimedia in Asia Workshops. New York: ACM, 2023: 1−2



Xiao Guangyi, born in 1994. PhD candidate. His main research interests include large language model and recommender system.
肖光义, 1994 年生。博士研究生。主要研究方向为大语言模型、推荐系统。



Hou Min, born in 1995. PhD, lecturer. Her main research interests include user modeling, recommender system, and data mining.
侯 旻, 1995 年生。博士, 讲师。主要研究方向为用户建模、推荐系统、数据挖掘。



Liao Yuxin, born in 2001. PhD candidate. Her main research interests include user modeling and personalized recommendation.
廖宇欣, 2001 年生。博士研究生。主要研究方向为用户建模、个性化推荐。



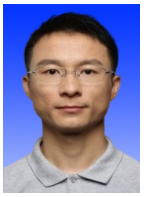
Wu Han, born in 1993. Postdoc. Her main research interests include knowledge evolution analysis, recommender system, and data mining.

武　晗, 1993 年生。博士后。主要研究方向为知识演化分析、推荐系统、数据挖掘。



Bai Chenxi, born in 2001. Master candidate. His main research interest includes recommender system based on large language models.

白晨希, 2001 年生。硕士研究生。主要研究方向为基于大模型推荐系统。



Wei Si, born in 1982. PhD. Senior Member of CCF. His main research interest includes key core technologies of artificial intelligence such as intelligent voice, natural language processing, and general large language models.

魏　思, 1982 年生。博士。CCF 高级会员。主要研究方向为智能语音、自然语言处理、通用大语言模型等人工智能关键核心技术。



Wu Le, born in 1988. PhD, professor, PhD supervisor. Her main research interests include data mining, personalized recommendation, and causal reasoning and decision making.

吴　乐, 1988 年生。博士, 教授, 博士生导师。主要研究方向为数据挖掘、个性化推荐、因果推理与决策。