计 算 机 研 究 与 发 展
Journal of Computer Research and Development
https://crad.ict.ac.cn
63(1): 147−161, 2026

基于大语言模型的双视角多级跨模态推荐

李佚名　于亚新　于之晟　司一廷　叶育松
（东北大学计算机科学与工程学院　沈阳　110169）
（医学影像智能计算教育部重点实验室（东北大学）　沈阳　110169）
（2372016@stu.neu.edu.cn）

Dual-Perspective Multi-Level Cross-Modal Recommendation Based on Large Language Models

Li Yiming, Yu Yaxin, Yu Zhisheng, Si Yiting, and Ye Yusong
（*School of Computer Science and Engineering*, *Northeastern University*, *Shenyang* 110169）
（*Key Laboratory of Intelligent Computing in Medical Image* (*Northeastern University*), *Ministry of Education*, *Shenyang* 110169）

Abstract　Multimodal recommendation systems aim to deliver more accurate and personalized recommendations by leveraging diverse data modalities such as text and images. Despite their potential, existing approaches face several critical challenges. 1) Feature distortion. The input embeddings are processed by relatively small pre-trained language models and deep convolutional neural networks, which may lead to inaccurate and incomplete feature representations. 2) Single encoding perspective. The multimodal encoding layers of current models only consider encoding from a single perspective that is either memory-based or expansion-based, which limits representational capacity and leads to information loss. 3) Poor multimodal alignment. Embeddings from different modalities are distributed in distinct and heterogeneous semantic spaces, and need to be effectively mapped into a shared semantic space to achieve accurate alignment. However, existing methods typically rely on simple operations such as direct product of behavioral information, which cannot adequately capture the complex and high-order relationships among modalities. As a result, precise alignment across multiple modalities is difficult to achieve. To address these issues, we propose a novel model named DPRec. This model considers encoding from both memory and expansion perspectives and introduces hypergraphs for multi-level precise cross-modal alignment. The model is evaluated through extensive experiments on three real-world datasets. Experimental results confirm the effectiveness of DPRec, demonstrating its superiority over state-of-the-art approaches in improving recommendation accuracy.

Key words　multimodal recommendation; feature representations; encoding perspective; hypergraph; cross-modal alignment

摘　要　多模态推荐系统旨在提供更为精准和个性化的推荐服务。然而，现有研究仍存在以下问题：1）特征失真。由于输入的嵌入均由小型预训练语言模型和深层卷积神经网络等模型进行处理，导致得到的特征表示不准确。2）编码视角单一。目前模型的多模态编码层只考虑在单一的记忆或扩展视角进行编码，造成信息缺失。3）多模态对齐效果欠佳。不同模态嵌入分布在不同空间中，需将其映射至同一空间以实现对齐。而现有方法通过简单的行为信息乘积无法捕捉模态之间的复杂关系，导致多种模态无法精

收稿日期：2025-01-15；**修回日期**：2025-08-27
基金项目：国家自然科学基金项目（62471122, 61871106, 61973059）；国家重点研发计划项目（2016YFC0101500）
This work was supported by the National Natural Science Foundation of China (62471122, 61871106, 61973059) and the National Key Research and Development Program of China (2016YFC0101500).
通信作者：于亚新（yuyx@mail.neu.edu.cn）

确对齐。基于上述问题,提出了一个新颖的模型 DPRec。该模型同时考虑了记忆与扩展的双视角编码,并引入超图进行多级精准跨模态对齐。所提模型在 3 个真实数据集上进行了扩展实验,实验结果验证了所提模型的有效性。

关键词 多模态推荐;特征表示;编码视角;超图;跨模态对齐

中图法分类号 TP391

DOI: 10.7544/issn1000-1239.202550040 **CSTR:** 32373.14.issn1000-1239.202550040

推荐系统是基于海量数据挖掘的一种智能平台,它根据用户的个人信息和物品特征(如用户的兴趣、历史购买行为和物品的材质、价格等),利用机器学习等人工智能技术建立模型,从而预测用户对新物品的评价和喜好[1]。然而,在大规模的用户和物品空间中,用户与物品之间的交互数据往往非常稀疏,因此存在数据稀疏、冷启动等问题。此外,传统推荐往往仅依赖于用户的历史交互信息,忽视了其他可能影响用户偏好的丰富信息源。

为了解决数据稀疏和单一信息源等问题,多模态推荐系统被提出,如图 1 所示。多模态推荐系统可以整合不同模态的信息,例如图片、文本、音频和视频等。通过将多种模态信息进行编码和融合,得到用户和物品的嵌入,从而进行推荐。多模态推荐方法致力于将多模态嵌入与用户行为嵌入进行精确融合,以准确捕捉用户的偏好和物品属性信息。

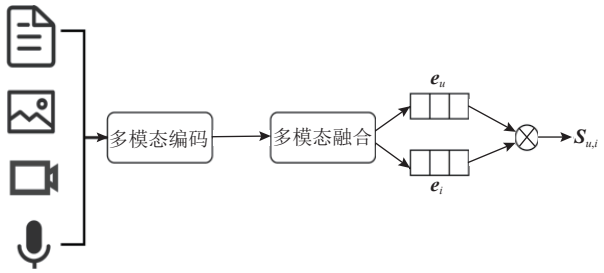


Fig. 1 A brief schematic diagram of multimodal recommendation

图 1 多模态推荐简要示意图

目前,许多研究已将多模态信息集成到推荐系统中。例如,VBPR[2] 和 CKE[3] 通过整合视觉特征扩展了矩阵分解(MF)[4] 方法。图卷积网络(GCN)[5-7] 因其能捕获高阶连接并强化偏好特征,正逐渐成为研究热点。例如,MMGCN[8] 将不同模态的信息融合到多个用户-项目图中,并通过连接学习到的模态表示来对用户偏好进行建模。GRCN[9] 在 MMGCN 的基础上利用多模态特征进一步细化了用户-项目图,旨在剔除虚假交互。DualGNN[10] 构建了一个额外的同构用户-用户图,以探索普遍的用户偏好模式。这些基于 GCN 的方法取得了巨大成功,并达到了最先进的性能水平。

多模态推荐相比于传统推荐,能够更全面地理解和刻画用户需求与物品特征,从而显著提升推荐的准确性和个性化程度。例如,在衣服推荐的场景中,如图 2 所示,多模态系统能够深入挖掘衣服的图片信息和描述文本,甚至结合用户对衣服的评论反馈,通过融合这些多模态信息,系统可以捕捉到用户对时尚的独特偏好,比如用户既喜欢某款衣服的明亮色彩和时尚剪裁,又注重其舒适材质和适合日常穿搭的实用性。因此,系统能够为用户推荐出既符合其审美又满足实际穿着需求的多样化衣服选项,极大地丰富了用户的购物体验,提高了推荐的满意度和转化率。

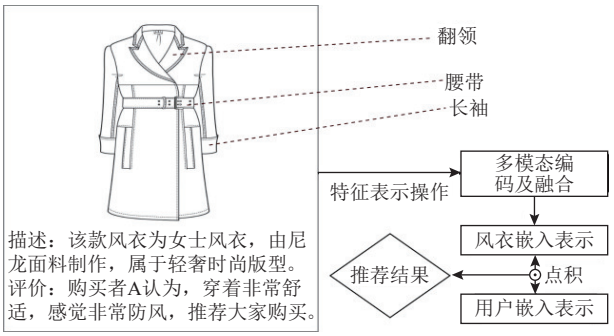


Fig. 2 Clothing recommendation example

图 2 衣服推荐示例

现有的多模态推荐方法虽然取得了显著成功,但仍存在 3 种局限性。1)目前,多模态模型在处理文本和图像特征嵌入时,分别使用小型预训练语言模型和深层 CNN 等模型处理。例如 sentence2Vec[11],VGG16[12] 等。由于该类模型参数规模相对较小,它们并不具备强大的自然语言处理能力以及图片处理能力,这可能导致特征失真,进而影响整体推荐性能。2)目前对于多模态编码层的研究[8-10] 主要集中于从单一记忆视角进行编码,也有少数研究[13] 尝试从扩展视角进行编码。这 2 种视角的编码分别通过构建用户-项目图及扩展用户-项目图来实现。然而,这种

单一视角编码容易导致编码信息的缺失。正确的做法应是将记忆信息作为主要编码信息，同时将扩展信息作为辅助补充，这样才能获得信息丰富的嵌入。换句话说，目前研究得到的多模态编码层信息不充分，导致多模态融合层无法得到精确的用户项目嵌入。3）现有研究通过简单地利用行为信息与不同模态嵌入进行乘积等相关运算，使其映射至与行为信息同一空间之中。但是，简单的向量拼接、元素级相乘等方法缺乏表达能力，无法充分捕捉不同模态之间的复杂关系。

为了解决上述问题，提出了一种融合记忆与扩展的 GCN 推荐模型 DPRec。该模型的输入是经过大语言模型（large language model, LLM）处理的嵌入向量，并构建了一个同时考虑记忆与扩展的双视角信息编码层，以获取更加丰富的模态信息。同时，利用超图进行多级跨模态精准对齐，将多种模态嵌入映射至同一空间中，实现模态对齐。

本文的主要贡献有 4 个方面：

1）提出了利用 LLM 进行多模态数据特征表示。同时，从不同模态信息中抽取实体词，并将其与用户行为信息相结合。该方法避免了特征失真，并从更细粒度的层面提升了模型的记忆能力。

2）提出了同时考虑记忆与扩展的双视角信息编码。该模块通过记忆与扩展 2 个图神经网络，有效抽取了记忆性与扩展性信息，避免了信息缺失。

3）提出了利用超图与双视角编码得到的模态嵌入进行多级跨模态精准对齐的方法。该方法可有效捕捉模态之间的复杂关系，使多种模态嵌入映射至同一空间，从而在模态融合层实现精确融合。

4）在 3 个公开数据集上进行了扩展实验，实验结果表明，所提模型在性能指标上均优于其他对比模型，验证了本文方法的有效性。

1 相关工作

1.1 LLM

随着模型参数量和训练数据量的不断增加，语言模型的文本建模困惑度以可预测的方式降低，在各类自然语言处理任务上的表现也持续提升[14]。基于 BERT[15] 和 Transformer[16] 的大语言模型通过自注意力机制和深度学习架构，极大地提升了文本处理效果，使其在推荐系统领域的应用越来越广泛。例如，LLMRec[17] 通过 3 种简单但有效的基于 LLM 的图增强策略来增强推荐系统。IDGenRec[18] 使用人类语言标记将每个项目表示为一个独特、简洁、语义丰富且平台无关的文本 ID，通过训练一个文本 ID 生成器与基于 LLM 的生成器协同工作，使得个性化推荐能够无缝集成到自然语言生成中。TALLRec[19] 通过推荐数据对 LLM 进行调整，构建了一个大型推荐语言模型，用于将 LLM 与推荐任务进行对齐。

以上推荐方法主要是在推荐过程中应用 LLM，但并没有从根本上解决特征失真问题。DPRec 模型通过使用 2 种不同的 LLM 来对多模态数据进行特征表示，从而获得多模态数据的精确嵌入。同时，利用 LLM 对多模态数据进行实体词抽取，并将这些实体词转换为嵌入向量与用户行为信息相结合。这有助于更好地获得记忆视角的信息，从而提高整体推荐效果。

1.2 多模态推荐

多模态推荐已成为网络媒体平台的重要应用之一。通过分析海量的多模态信息（如图像、文本）以及用户的历史行为数据，能够为用户提供高度个性化的推荐服务。这一过程不仅提升了用户体验，还能增强平台的用户粘性和满意度。例如，VPBR 和 VPOI[20] 利用预训练的 CNN 模型来提取深度视觉特征，以此丰富项目表示。ACF[21] 通过注意力机制自适应地捕获用户对每种模态的偏好。LATTICE[22] 引入了项目语义图，以揭示项目之间的潜在相关信号。FREEDOM[23] 在 LATTICE 基础上，通过降低用户-项目交互图的噪声，并在初始化后固定项目语义图，进一步提升了推荐精度。最近，MKGAT[24] 引入了图注意力机制，以平衡模态融合过程中不同模态的权重。然而，现有方法在多模态编码层仍拘泥于使用单一视角的用户-项目图或扩展用户-项目图，并且多模态对齐问题没有得到很好的解决，从而限制了推荐系统的整体表现。

1.3 超 图

超图作为一种能够捕捉复杂关系的数据结构，在推荐系统中展现出了独特的应用潜力。超图卷积神经网络本质上是将 GCN 显式或隐式地扩展到超图上，或者引入注意力机制，设计出适用于超图的卷积操作[25]。在推荐系统领域，超图可以用来构建用户与商品之间的多维关系模型，例如，用户对商品的评分、浏览历史、购买行为等，这些都可以作为超边连接多个节点的依据。超图神经网络（HGNN）[26] 是图神经网络的扩展，它能够处理超图结构数据，更精准地捕捉用户和商品之间的复杂交互模式。HGNN 通过超边卷积操作利用高阶数据相关性进行表示学习，

在处理多模态、异构数据时显示出了优势。为了避免构建超图时产生过多虚假关联，通常使用函数对超图进行参数化，例如 Gumbel-Softmax[27]。此外，超图对比学习也被提出用于推荐系统，通过构建超图结构并进行对比学习，可以提高推荐系统在处理噪声和倾斜分布数据时的性能。最近的研究还提出了自监督超图 Transformer(self-supervised hypergraph Transformer，SHT)[28]，这是一种新框架，它通过探索全局协作关系来增强用户嵌入，并使用跨视图生成自监督学习组件进行数据增强，以增强推荐系统的鲁棒性。

DPRec 利用超图学习到的模态嵌入与双视角编码得到的模态嵌入进行多级跨模态精准对齐，实现了将多种模态嵌入映射到同一空间的目标，从而提高模型的整体性能。

2 方法介绍

在本节中，将详细描述所提出的模型 DPRec。表 1 中记录了本文出现的主要符号的描述。

2.1 定义及总体架构

定义 1. 记忆视角。记忆视角指根据用户集 U 与项目集 I 的交互矩阵 $\boldsymbol{R}\in\{0,1\}^{|U|\times|I|}$，构建对应的二分图 G。而后将 $\boldsymbol{E}_{i,\mathrm{id}}$，$\boldsymbol{E}_{\mathrm{entity}}$，$\boldsymbol{E}_{i,m}$ 等输入图 G 中进行图卷积操作得到 $\boldsymbol{E}_{m_\mathrm{rep}}$，其中 $m\in\{\mathrm{id,t,v}\}$。记忆视角编码 $Memery$ 表示为 $\sum_{m\in\{\mathrm{id,t,v}\}} E_{m_\mathrm{rep}}$。

定义 2. 扩展视角。扩展视角指在二分图 G 基础上，利用项目 I 之间的相似度对 G 进行扩展，得到扩展图 $\tilde{G}$。将 $\boldsymbol{E}^{\wedge}_{u,\mathrm{id}}$，$\boldsymbol{E}^{\wedge}_{i,\mathrm{id}}$，$\boldsymbol{E}^{+}_{i,\mathrm{v}}$，$\boldsymbol{E}^{+}_{i,\mathrm{t}}$，$\boldsymbol{E}_{u,\mathrm{v}}$，$\boldsymbol{E}_{u,\mathrm{t}}$ 等输入 $\tilde{G}$ 中进行图卷操作得到 $\boldsymbol{E}^{\sim}_{m}$，其中 $m\in\{\mathrm{id,t,v}\}$。扩展视角编码 $Expand$ 表示为 $\sum_{m\in\{\mathrm{id,t,v}\}} E_{m_\mathrm{rep}}$。

定义 3. 双视角。双视角指在多模态编码层搭建用户-项目图 G 与扩展图 $\tilde{G}$，同时考虑在记忆与扩展这 2 种视角进行信息编码。双视角(DP)编码可以表示为 $Memery\cup Expand$，其中 $\cup$ 表示将二者沿着维度为 1 进行拼接。

定义 4. 多级跨模态对齐。多级跨模态对齐是指在 3 个不同级别上进行模态之间的对齐。划分为 3 个级别进行模态对齐，旨在从更细粒度的层面提高模态对齐效果。3 个级别分别表示为

$$L_1=\boldsymbol{E}^{\mathrm{t},H}\leftrightarrow\boldsymbol{E}_{\mathrm{v_rep}},\tag{1}$$

$$L_2=\boldsymbol{E}^{\mathrm{v},H}\leftrightarrow\boldsymbol{E}_{\mathrm{t_rep}},\tag{2}$$

Table 1 Description of the Main Notations

表 1 主要符号的描述

符号	描述
$U=\{u\}$	用户集
$I=\{i\}$	项目集
M	模态集合
m	模态
G	二分图
V	节点集
E	边集
d	ID 嵌入维度
d^m	LLM 处理模态嵌入的维度
$\boldsymbol{R}\in\{0,1\}^{\lvert U\rvert\times\lvert I\rvert}$	交互矩阵
$\boldsymbol{A}\in\mathbb{R}^{(\lvert U\rvert+\lvert I\rvert)\times(\lvert U\rvert+\lvert I\rvert)}$	根据交互矩阵 $\boldsymbol{R}$ 构造的邻接矩阵
$\boldsymbol{E}_{i,m}$	LLM 处理的模态嵌入
$\boldsymbol{E}_{\mathrm{entity}}$	LLM 处理的实体词嵌入
$\boldsymbol{E}^{m,H}$	超图得到的模态嵌入
$\boldsymbol{E}^{H}$	超图得到的混合模态嵌入
$\boldsymbol{V}^{m}$	模态超边初始化嵌入
$\boldsymbol{E}_{m_\mathrm{rep}}$	记忆视角得到的模态嵌入
$\boldsymbol{E}^{\sim}_{m}$	扩展图得到的模态嵌入
$\boldsymbol{E}^{\wedge}_{u,\mathrm{id}}$	扩展视角输入用户 ID 初始化嵌入
$\boldsymbol{E}^{\wedge}_{i,\mathrm{id}}$	扩展视角输入项目 ID 初始化嵌入
$\boldsymbol{E}^{+}_{i,m}$	项目同质图得到的模态嵌入
$\boldsymbol{E}_{u,m}$	用户模态初始化嵌入
$\boldsymbol{H}^{m}_{i}$	项目超边依赖矩阵
$\boldsymbol{H}^{m}_{u}$	用户超边依赖矩阵

$$L_3=\boldsymbol{E}^{H}\leftrightarrow\boldsymbol{E}_{\mathrm{id_rep}}+\boldsymbol{E}^{H}\leftrightarrow\tilde{\boldsymbol{E}}_{\mathrm{t}}+\boldsymbol{E}^{H}\leftrightarrow\tilde{\boldsymbol{E}}_{\mathrm{v}},\tag{3}$$

其中 L_1，L_2，L_3 表示多级别对齐产生的对齐损失，$\leftrightarrow$ 表示在不同嵌入之间参数化计算模态之间的距离，H 表示图卷积总层数。

根据定义 1~4 可大致描述本文要解决的问题及简要过程。在多模态编码层只考虑记忆视角得到 $Memery$ 或只考虑扩展视角得到 $Expand$ 均会导致编码信息不充分。基于此，提出在编码层构建一个双视角编码由 $Memery\cup Expand$ 得到，既考虑了记忆信息也考虑了扩展信息。除此之外，由于不同模态之间分布差异巨大，不可直接进行多模态融合。基于此，提出利用超图学习到的丰富信息与双视角编码层得到的信息进行多级跨模态对齐操作，即上述 $\boldsymbol{E}^{\mathrm{t},H}\leftrightarrow\boldsymbol{E}_{\mathrm{v_rep}}$，$\boldsymbol{E}^{\mathrm{v},H}\leftrightarrow\boldsymbol{E}_{\mathrm{t_rep}}$，$\boldsymbol{E}^{H}\leftrightarrow\boldsymbol{E}_{\mathrm{id_rep}}+\boldsymbol{E}^{H}\leftrightarrow\tilde{\boldsymbol{E}}_{\mathrm{t}}+\boldsymbol{E}^{H}\leftrightarrow\tilde{\boldsymbol{E}}_{\mathrm{v}}$。经上述操作之后，将得到同时富含记忆与扩展且分布在同一空间中的多种模态信息。便于之后的模态

融合得到准确的用户项目嵌入，提高推荐效果。

模型的总体架构如图 3 所示，其中主要包含 2 个模块：

1）多模态双视角信息编码层。在保留传统用户-项目图的同时引入扩展用户-项目图，旨在同时抽取记忆性与扩展性信息，得到富含双视角信息的用户项目嵌入，避免信息缺失，如图 4 所示。

2）多模态对齐层。利用 HGNN 学习到的不同模态嵌入、混合模态嵌入与双视角编码得到的不同模态嵌入进行 3 种级别的跨模态精准对齐，如图 5 所示。

2.2 多模态特征表示

在本研究中，摒弃了之前使用小型预训练语言模型和 CNN 进行数据处理的方法，转而利用 LLM 将多模态数据处理为嵌入表示，以发挥其本身的自然语言处理能力和图片处理方面的优势。具体来说，使用了 2 种预训练 LLM 来分别处理文本和图像数据。此外，还利用 LLM 执行其擅长的实体词抽取任务。在抽取实体词方面，通过设计合适的提示词模板从文本和图像中抽取关键实体词。对于文本数据，设计了如下提示词：“从该段文本中抽取出你认为所有可能的实体词”。对于图像数据，使用了描述性的提示词，如“请提取出该图片中的实体词，只需提取出关键主题词，一些微小因素可进行适当忽略。例如，如果图片中有一个人，描述他们的服装、姿势和表情等相关词语”。通过这些提示词模板，LLM 能够自动提取出品牌名、产品类型等关键实体。同时，将抽取出来的实体词转换为嵌入向量，并将其与用户行为信息相结合，旨在更细粒度地捕捉用户兴趣偏好，丰富项目属性信息，从而增强推荐系统的准确性。

2.3 多模态信息编码

现有的多模态方法面临着难以精确获取记忆信息和扩展信息的挑战。鉴于此，本文创新性地提出

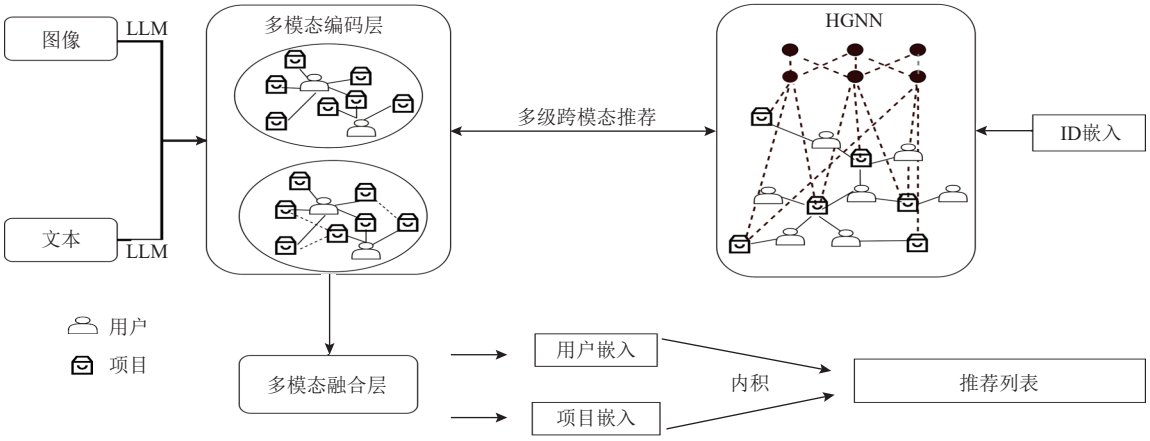


Fig. 3 Overall framework diagram of system

图 3 系统整体框架图

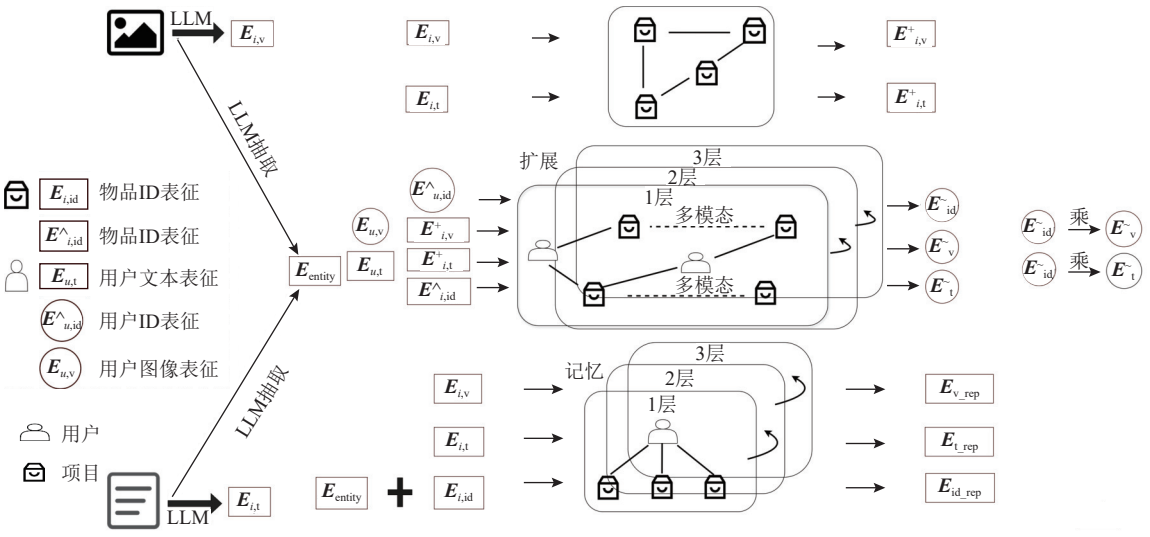


Fig. 4 Multimodal encoding layer

图 4 多模态编码层

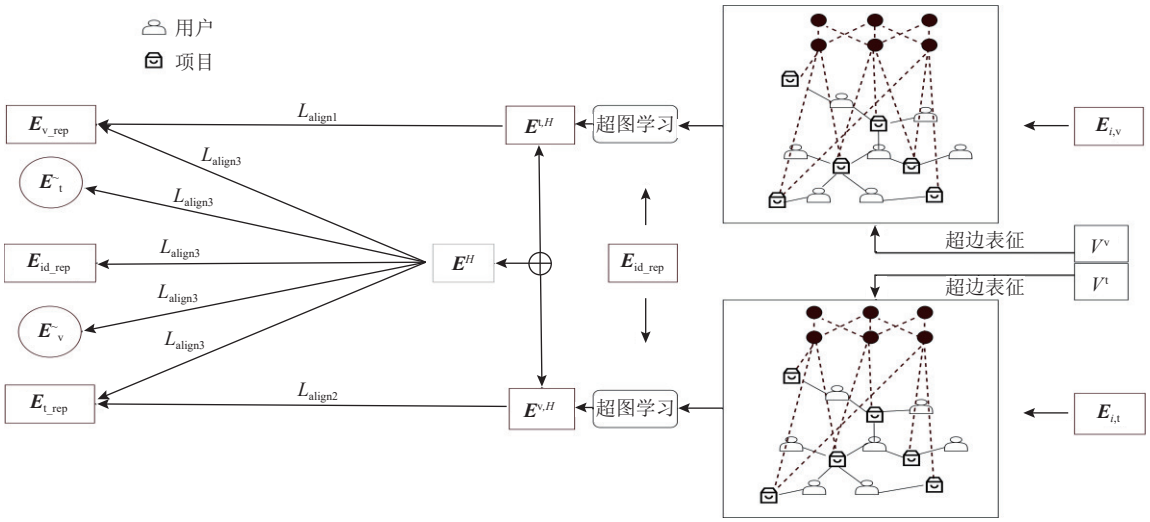


Fig. 5 Multimodal alignment

图 5 多模态对齐

了一种同时考虑记忆与扩展的双视角多模态信息编码器，旨在获取更丰富的多模态用户项目嵌入，以解决现有方法中嵌入信息缺失的问题。

2.3.1 扩展视角下的编码

在基于 GCN 的推荐系统中，盲目引入高阶节点可能会导致虚假信息在传播过程中被引入。受 GUME[13] 的启发，应采取一种更为精细化的方法：仅为每个项目保留相似度得分最高的前 K 个邻居来构建扩展用户-项目图。这一方法有助于识别出在多种模式下与目标项目相似的项目，但 GUME 仅将扩展图得到的用户与项目模态嵌入进行串联拼接并未将行为扩展信息引入其中。本文的扩展视角在得到扩展的模态与行为嵌入后，利用行为嵌入与模态嵌入相乘，引入行为信息。

首先，采用 KNN 算法为每种模态 m 构建一个项目-项目图，目的在于挖掘项目之间的多模态关系。之后，通过计算项目对 $(i,j)\in I$ 的原始模态特征 $\boldsymbol{e}_i^m$ 和 $\boldsymbol{e}_j^m$ 的余弦相似度来计算相似度得分 $S_{i,j}^m$。

$$S_{i,j}^m=\frac{\left(\boldsymbol{e}_i^m\right)^{\mathrm{T}}\boldsymbol{e}_j^m}{\|\boldsymbol{e}_i^m\|\|\boldsymbol{e}_j^m\|}。\tag{4}$$

只保留相似度得分最高的前 K 个邻居来捕获最相关的特征。

$$\bar{S}_{i,j}^m=\begin{cases}S_{i,j}^m, & S_{i,j}^m\in topK\left(\left\{S_{i,p}^m,p\in I\right\}\right),\\ 0, & \text{其他},\end{cases}\tag{5}$$

其中 $\bar{S}_{i,j}^m$ 表示模态 m 内项目 i 和项目 j 之间的边权重，$\{S_{i,p}^m,p\in I\}$ 表示第 i 个项目的邻居分数。同时，为了减缓梯度爆炸或者消失问题，将邻接矩阵标准化为：

$$\tilde{\boldsymbol{S}}^m=(\boldsymbol{D}^m)^{-\frac{1}{2}}\bar{\boldsymbol{S}}^m(\boldsymbol{D}^m)^{-\frac{1}{2}},\tag{6}$$

其中 $\boldsymbol{D}^m$ 是 $\bar{\boldsymbol{S}}^m$ 的对角线矩阵，在初始化后便固定每个模态项图。

然后，将这些项目定义为目标项目的语义邻居。项目集及其语义邻居可以表示为

$$C=\bigcup_{i\in I}\{(i,j)|j\in \boldsymbol{N}_i\},\tag{7}$$

其中 N_i 表示项目 i 的语义邻居。然后，通过在项目和它们的语义邻居之间添加边来扩展用户-项目图。形式上：

$$\boldsymbol{\gamma}=\begin{pmatrix}\boldsymbol{0} & \boldsymbol{R}\\ \boldsymbol{R}^{\mathrm{T}} & \boldsymbol{C}\end{pmatrix},\tag{8}$$

其中 $\boldsymbol{0}$ 是全 0 矩阵，$\boldsymbol{R}$ 是用户-项目交互矩阵。

扩展图成功构建后，对项目同质图执行图卷积操作以传播和更新模态特征：

$$\boldsymbol{E}_{i,m}^{+}=\tilde{\boldsymbol{S}}^m\boldsymbol{E}_{i,m}。\tag{9}$$

同时，在扩展用户-项目图上传播特定模态的嵌入，以提取特定于模态的扩展兴趣嵌入，这同样适用于 ID 嵌入。具体来说，在图卷积的第 l 层，用户和项目嵌入可以表示为

$$\boldsymbol{E}_m^{\sim(l)}=\left(\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{\gamma}\boldsymbol{D}^{-\frac{1}{2}}\right)\boldsymbol{E}_m^{\sim(l-1)},\tag{10}$$

其中 $\boldsymbol{E}_m^{\sim(l)}$ 表示图卷积第 l 层用户和项目的特定模态嵌入，$\boldsymbol{D}$ 是 $\boldsymbol{\gamma}$ 的对角度矩阵，$\boldsymbol{E}_m^{\sim(0)}$ 是 $\boldsymbol{E}_{u,m}$ 和 $\boldsymbol{E}_{i,m}^{+}$ 的串联。

通过聚合多层邻居信息，可以获得最终特定模态的扩展兴趣特征：

$$\boldsymbol{E}_m^{\sim}=\frac{1}{L+1}\sum_{i=0}^{L}\boldsymbol{E}_m^{\sim(l)},\tag{11}$$

其中 L 是用户-项目图层数。

最后，利用扩展视角得到的用户行为信息与其

余模态信息进行相乘，目的是进行特征增强，将扩展行为信息引入其余扩展模态嵌入中。

2.3.2 记忆视角下的编码

为了捕获记忆视角信息，首先为每个模态分别构建一个用户-项目图G。之后对$\boldsymbol{E}_{i,\mathrm{id}}$进行更新，利用LLM抽取的实体词使其用户行为信息更加丰富。

$$\boldsymbol{E}_{i,\mathrm{id}} = \boldsymbol{E}_{i,\mathrm{id}} + \boldsymbol{E}_{\mathrm{entity}}, \tag{12}$$

其中$\boldsymbol{E}_{\mathrm{entity}}$是由LLM进行实体词抽取后转化得到的嵌入向量。

在第l个图卷积层上用户和项目表示可以公式化为

$$\boldsymbol{E}_m^{(l)} = \sum_{i\in N_u} \frac{1}{\sqrt{|N_u||N_i|}} \boldsymbol{E}_m^{(l-1)}, \tag{13}$$

其中N_u和N_i分别代表图G中u和i的1跳邻居，$\boldsymbol{E}_m$为$\boldsymbol{E}_{i,m}$和记忆视角用户嵌入串联获得，记忆视角用户嵌入由聚合邻居项目嵌入得到。

每种模态的最终嵌入都是通过逐元素求和来计算。形式上：

$$\boldsymbol{E}_{m_\mathrm{rep}} = \sum_{l=0}^{L} \boldsymbol{E}_m^{(l)}, \tag{14}$$

其中L是用户-项目图层数。

然后将$\boldsymbol{E}_{i,m}$放进G中进行图卷积操作，利用式(13)(14)进行计算操作，最终可以得到$\boldsymbol{E}_{m_\mathrm{rep}}$。其中$m \in \{\mathrm{id}, \mathrm{t}, \mathrm{v}\}$。

2.4 跨模态对齐

本节利用超图得到的丰富模态嵌入与双视角编码得到的嵌入信息进行多级跨模态精准对齐，目的是将不同模态的嵌入映射到同一空间中。

2.4.1 超图构建及消息传递

LGMRec[29]将超图学习到的文本与图像模态嵌入利用对比学习方法拉近二者之间距离，并将二者相加，与该模型记忆视角得到的嵌入相结合得到最终的嵌入。本文受其超图应用的启发，在保留消息传递得到2种模态嵌入之外，将二者相加得到超图混合模态嵌入，并弃用利用对比学习直接拉近其距离的方式。

首先，本文将可学习的隐式属性向量$\boldsymbol{v}_a^m \in \mathbb{R}^{d_m}$，$a = 1,2,\cdots,N$定义为模态$m$的超边，自适应地学习隐式属性以及项目与用户之间的依赖关系，其中N是超边数量。具体来说，通过以下方式获得低维嵌入空间中的超图依赖矩阵：

$$\boldsymbol{H}_i^m = \boldsymbol{E}_{i,m} \cdot ((\boldsymbol{V}^m)^{\mathrm{T}}), \boldsymbol{H}_u^m = \boldsymbol{A}_u \cdot ((\boldsymbol{H}_i^m)^{\mathrm{T}}), \tag{15}$$

其中$\boldsymbol{H}_i^m \in \mathbb{R}^{|I|\times N}$和$\boldsymbol{H}_u^m \in \mathbb{R}^{|U|\times N}$是项目和用户的超边依赖矩阵，$\boldsymbol{V}^m = \left(\boldsymbol{V}_1^m, \boldsymbol{V}_2^m, \cdots, \boldsymbol{V}_N^m\right) \in \mathbb{R}^{N\times d_m}$为超边向量矩阵，$\boldsymbol{A}_u \in \mathbb{R}^{|U|\times|I|}$是从$\boldsymbol{A}$中提取的与用户相关的邻接矩阵。

为避免构建超图时产生虚假关联，采用Gumbel-Softmax重新进行参数化，以确保每个项目尽可能仅附加到1个超边：

$$\tilde{\boldsymbol{h}}_{i,*}^m = softmax\left(\frac{\ln\delta - \ln(1-\delta) + \boldsymbol{h}_{i,*}^m}{\tau}\right), \tag{16}$$

其中$\boldsymbol{h}_{i,*}^m \in \mathbb{R}^N$是项目超边依赖矩阵中第$i$行向量，反映了项目$i$与所有超边之间的关系。$\boldsymbol{\delta} \in \mathbb{R}^N$是噪声向量，其中，每个值$\delta_j \sim Uniform(0,1)$，$\tau$是温度超参数。经过以上操作，可以得到增强项目属性超图依赖矩阵$\tilde{\boldsymbol{H}}_i^m$。同理，也可以得到增强用户属性超图依赖矩阵$\tilde{\boldsymbol{H}}_u^m$。

通过将属性超边作为中间枢纽，可以实现超图消息传递，将全局信息传递给用户和项目，而不受几跳距离的限制。形式上：

$$\boldsymbol{E}_i^{m,h+1} = drop(\tilde{\boldsymbol{H}}_i^m) \cdot drop\left((\tilde{\boldsymbol{H}}_i^m)^{\mathrm{T}}\right), \tag{17}$$

其中$\boldsymbol{E}_i^{m,h}$是第h超图层中项目全局嵌入矩阵，$drop$代表dropout函数。使用记忆视角得到的项目协作嵌入$\boldsymbol{E}_{\mathrm{id_rep}}$作为$h=0$时的全局嵌入矩阵初始化。类似地，全局用户嵌入矩阵可以计算为

$$\boldsymbol{E}_u^{m,h+1} = drop(\tilde{\boldsymbol{H}}_u^m) \cdot drop\left((\tilde{\boldsymbol{H}}_u^m)^{\mathrm{T}}\right) 。 \tag{18}$$

然后，可以通过聚合所有模态的全局嵌入来获得全局嵌入矩阵$\boldsymbol{E}^H$，

$$\boldsymbol{E}^H = \alpha \cdot norm\left(\sum_{m\in M} \boldsymbol{E}^{m,H}\right), \tag{19}$$

$$\boldsymbol{E}^{m,H} = \left(\boldsymbol{E}_u^{m,H}, \boldsymbol{E}_i^{m,H}\right), \tag{20}$$

其中$\boldsymbol{E}_u^{m,H} \in \mathbb{R}^{|U|\times d}$，$\boldsymbol{E}_i^{m,H} \in \mathbb{R}^{|I|\times d}$分别是在模态$m$下第$H$超图层中获得的用户$u$与项目$i$的全局嵌入矩阵，$H$为超图卷积总层数，$norm$是一个标准化函数，用于减轻嵌入之间的值尺度差异，$\alpha$是一个可调节因子，用于控制全局嵌入的结合程度。

2.4.2 多级跨模态精准对齐

模态对齐一直是热点讨论问题。MENTOR[30]中提出了一个4级跨模态对齐方式，包括行为直接指导、行为间接指导、模态直接对齐和模态间接对齐，旨在将记忆视角得到的多种模态嵌入实现多种不同程度的对齐以拉近不同模态之间的距离。受其启发，本文利用超图得到的3种模态嵌入与双视角得到的嵌入来实现跨模态对齐，对齐分为3个级别：1)超图文本嵌入与记忆视角图像嵌入对齐；2)超图图像嵌入与记忆视角文本嵌入对齐；3)超图混合模态嵌入与记忆视角行为嵌入以及扩展视角文本与图像嵌入

均进行对齐。为高效实现跨模态对齐,需采用一种方法来量化和比较不同模态之间的相似性和差异性。高斯分布参数化是一种自然且有效的选择,其由均值和方差2个参数完全确定。通过参数化模态嵌入为高斯分布,可直观地衡量不同模态之间的差异。此外,分布之间的相似性可通过KL散度来量化,值越小表示2个分布越接近,这为多级对齐提供了理论基础。假设有2个嵌入分别为$\boldsymbol{E}_1$, $\boldsymbol{E}_2$,将其参数化为$\boldsymbol{E}_1 \sim N\left(\mu_1,\sigma_1^2\right)$, $\boldsymbol{E}_2 \sim N\left(\mu_2,\sigma_2^2\right)$,二者之间的KL散度为

$$D_{\mathrm{KL}} = \ln\left(\frac{\sigma_2}{\sigma_1}\right) + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2}, \tag{21}$$

可等效简化为

$$D_{\mathrm{KL}} \leftrightarrow |\mu_1 - \mu_2| + |\sigma_1 - \sigma_2| \text{。} \tag{22}$$

由此可以得到所提3级对齐的表示方式。

1)第1级别为:

$$\boldsymbol{E}^{\mathrm{t},H} \sim N\left(\mu_{\mathrm{ct}},\sigma_{\mathrm{ct}}^2\right), \boldsymbol{E}_{\mathrm{v_rep}} \sim N\left(\mu_{\mathrm{v}},\sigma_{\mathrm{v}}^2\right), \tag{23}$$

$$L_{\mathrm{align1}} = |\mu_{\mathrm{ct}} - \mu_{\mathrm{v}}| + |\sigma_{\mathrm{ct}} - \sigma_{\mathrm{v}}|, \tag{24}$$

其中$(\mu_{\mathrm{ct}},\sigma_{\mathrm{ct}})$与$(\mu_{\mathrm{v}},\sigma_{\mathrm{v}})$分别表示超图文本嵌入与记忆视角图像嵌入的高斯分布。

2)第2个级别为:

$$\boldsymbol{E}^{\mathrm{v},H} \sim N\left(\mu_{\mathrm{cv}},\sigma_{\mathrm{cv}}^2\right), \boldsymbol{E}_{\mathrm{t_rep}} \sim N\left(\mu_{\mathrm{t}},\sigma_{\mathrm{t}}^2\right), \tag{25}$$

$$L_{\mathrm{align2}} = |\mu_{\mathrm{cv}} - \mu_{\mathrm{t}}| + |\sigma_{\mathrm{cv}} - \sigma_{\mathrm{t}}|, \tag{26}$$

其中$(\mu_{\mathrm{cv}},\sigma_{\mathrm{cv}})$与$(\mu_{\mathrm{t}},\sigma_{\mathrm{t}})$分别表示超图图像嵌入与记忆视角文本嵌入的高斯分布。

3)第3级别为:

$$\boldsymbol{E}^{H} \sim N\left(\mu_{\mathrm{ch}},\sigma_{\mathrm{ch}}^2\right), \tilde{\boldsymbol{E}}_{\mathrm{t}} \sim N\left(\mu_{\mathrm{et}},\sigma_{\mathrm{et}}^2\right), \tag{27}$$

$$\boldsymbol{E}_{\mathrm{id_rep}} \sim N\left(\mu_{\mathrm{id}},\sigma_{\mathrm{id}}^2\right), \tilde{\boldsymbol{E}}_{\mathrm{v}} \sim N\left(\mu_{\mathrm{ev}},\sigma_{\mathrm{ev}}^2\right), \tag{28}$$

$$\begin{aligned} L_{\mathrm{align3}} = {} & |\mu_{\mathrm{ch}} - \mu_{\mathrm{et}}| + |\sigma_{\mathrm{ch}} - \sigma_{\mathrm{et}}| + |\mu_{\mathrm{ch}} - \mu_{\mathrm{id}}| + \\ & |\sigma_{\mathrm{ch}} - \sigma_{\mathrm{id}}| + |\mu_{\mathrm{ch}} - \mu_{\mathrm{ev}}| + |\sigma_{\mathrm{ch}} - \sigma_{\mathrm{ev}}|, \end{aligned} \tag{29}$$

其中$(\mu_{\mathrm{ch}},\sigma_{\mathrm{ch}})$, $(\mu_{\mathrm{et}},\sigma_{\mathrm{et}})$, $(\mu_{\mathrm{id}},\sigma_{\mathrm{id}})$, $(\mu_{\mathrm{ev}},\sigma_{\mathrm{ev}})$分别表示超图混合模态嵌入、扩展文本嵌入、记忆视角行为信息嵌入、扩展视角图像嵌入的高斯分布。

最终计算总体多级跨模态对齐损失L_{align}。形式上:

$$L_{\mathrm{align}} = \lambda_{\mathrm{align}}\left(L_{\mathrm{align1}} + L_{\mathrm{align2}} + L_{\mathrm{align3}}\right), \tag{30}$$

其中λ_{align}是平衡超参数。

2.5 模态融合

首先,基于每种模态项目-项目图的最终嵌入来增强记忆视角的项目模态嵌入,表示为:

$$\boldsymbol{E}_{m_\mathrm{rep}} = \left\{\boldsymbol{E}_{\mathrm{v},m_\mathrm{rep}} \| \boldsymbol{E}_{i,m_\mathrm{rep}} + \boldsymbol{E}_{i,m}^{+}\right\}, \tag{31}$$

其中‖表示串联操作。

然后,以不同的注意力水平融合记忆视角的ID、图像、文本模态。

$$\boldsymbol{E}_{\mathrm{rep}} = \{\eta \times \boldsymbol{E}_{\mathrm{v_rep}} \| \beta \times \boldsymbol{E}_{\mathrm{t_rep}} \| \kappa \times \boldsymbol{E}_{\mathrm{id_rep}}\}, \tag{32}$$

其中η, β, κ是可训练参数,且$\eta+\beta+\kappa=1$。

同理,对于扩展视角的模态嵌入,也使用不同的注意力水平融合图像、文本模态:

$$\boldsymbol{E}_{\mathrm{extend_rep}} = \{\varpi \times \tilde{\boldsymbol{E}}_v \| \rho \times \tilde{\boldsymbol{E}}_t\}, \tag{33}$$

其中ϖ, ρ是可训练参数,且$\varpi+\rho=1$。

最后,将上述得到的记忆视角模态嵌入与扩展视角模态嵌入进行横向拼接,得到富含双视角信息的用户项目嵌入。

$$\boldsymbol{E}_{\mathrm{final_rep}} = \left\{\boldsymbol{E}_{\mathrm{rep}} \oplus \boldsymbol{E}_{\mathrm{extend_rep}}\right\}, \tag{34}$$

其中⊕表示横向拼接。

2.6 模态预测和优化

为了优化模型的推荐效果,采用贝叶斯个性化排名(BPR)损失作为主要的优化函数。BPR的目标是扩大每个三元组$(u,p,n) \in D$的正样本和负样本之间的预测偏好差距,其中D表示训练集。正样本p是指用户u交互过的项目,而负样本n是从用户u从未互动过的物品集合中随机选择的。BPR函数定义为:

$$L_{\mathrm{bpr}} = \sum_{(u,p,n)\in D} -\ln(\sigma(y_{u,p} - y_{u,n})), \tag{35}$$

其中$y_{u,p}$和$y_{u,n}$是用户u对正样本p和负样本n的评分,由$\boldsymbol{E}_{u_\mathrm{final_rep}}{}^{\mathrm{T}}\boldsymbol{E}_{p_\mathrm{final_rep}}$和$\boldsymbol{E}_{u_\mathrm{final_rep}}{}^{\mathrm{T}}\boldsymbol{E}_{n_\mathrm{final_rep}}$得到。$\sigma$是Sigmoid函数。

然后,结合BPR损失和多级跨模态对齐损失来更新用户和项目表示。如下所示:

$$L_{\mathrm{total}} = L_{\mathrm{bpr}} + L_{\mathrm{align}} + \lambda \|\Theta\|_2^2, \tag{36}$$

其中λ是控制L_2正则化的超参数,Θ是模型参数集。

3 性能测试与评价

在本节中,使用全面的实验来评估DPRec模型在3个广泛使用的现实数据集上的性能,实验中主要验证4个研究问题:RQ1, RQ2, RQ3, RQ4。RQ1:与当前先进的多模态推荐方法相比,本文模型性能是否最优。RQ2:本文模型的各个组件对模型的性能有什么影响。RQ3:本文模型的参数对性能的影响。RQ4:本文模型具体计算开销。

3.1 实验设置

3.1.1 数据集描述

为了评估本文提出的DPRec模型在top-k项目推荐任务中的性能,在亚马逊数据集[31]的3个类别

Baby, Sports, Clothing 上进行了实验。这些数据集同时提供了产品的描述和图像，本文利用 2 个预训练好的 LLM 对数据集进行处理，分别提取 512 维视觉特征和 1 536 维文本特征。这些数据集的统计数据如表 2 所示。

Table 2　Statistics of Datasets

表 2　数据集统计

数据集	用户数量	项目数量	交互数量	稀疏性/%
Baby	19 445	7 050	160 792	99.88
Sports	35 598	18 357	296 337	99.95
Clothing	39 387	23 033	278 677	99.97

3.1.2　评价指标

为了公平地评估性能，本文采用了 2 个广泛使用的指标：*Recall@k* 和 *NDCG@k*。在训练、验证和测试阶段，数据集采用 8:1:1 的随机数据分割。最后，通过报告测试集中所有用户在 top-*k* 条件下的平均指标来评估不同方法的推荐性能，其中 *k* 值根据经验设置为 10 和 20。

3.1.3　基线

为了证明 DPRec 模型的有效性，将其与以下最先进的方法进行比较，这些方法可以分为传统推荐方法和多模态推荐方法 2 组。

1）传统的推荐方法

① MF-BPR[32]。该方法利用 BPR 损失函数，有效提升了基于矩阵分解的推荐系统性能。

② LightGCN[33]。该方法基于 GCN 去除了冗余组件，从而简化了模型结构。

③ NCL[34]。该模型是一个图协同过滤框架，通过结合结构邻居和语义邻居来改进神经图协同过滤。

2）多模态推荐方法

① VBPR[2]。该方法通过集成每个项目的视觉和 ID 嵌入作为其表示来增强 MF-BPR 的方法，并将其输入到矩阵分解框架中。为了确保评估过程的公正性，进一步将 VBPR 模型扩展到了文本领域。

② MMGCN[8]。该方法利用 GCN 为每种模态构建特定的图。最终的预测结果是通过融合所有模态来计算。

③ LATTICE[22]。该方法构建了一个额外的项目语义图来捕获潜在的语义相关信号。

④ MMGCL[35]。该模型通过图对比学习来增强微视频推荐中的多模态嵌入学习。

⑤ SLMRec[36]。该模型将 SSL（自监督学习）融入基于图神经网络（GNN）的推荐模型中，并提出了 3 种数据增强方法，旨在揭示多模态信息中的潜在模式。

⑥ BM3[37]。该模型通过采用潜在表示丢失机制代替图增强来生成对比视图，简化了自监督多模态推荐模型。

⑦ DiffMM[38]。该模型将模态感知图扩散模型与跨模态对比学习相结合，更好地对齐多模态特征信息与协作关系建模。

⑧ MENTOR[30]。该模型提出了一个多层次的自监督学习任务，以解决标签稀疏问题和模态对齐问题。

⑨ LGMRec[29]。该模型提出了利用超图捕获全局用户项目嵌入，同时将得到的全局嵌入与用户-项目图得到的局部嵌入结合，以提高准确性和鲁棒性。

⑩ IDGenRec[18]。该模型通过人类语言标记将每个项目表示为一个独特、简洁、语义丰富且与平台无关的文本 ID，从而能够将个性化推荐无缝集成到自然语言生成中。

3.1.4　实施细节

DPRec 以及对比方法均基于 PyTorch 实现。对于一般设置，本文使用维度为 64 的 Xavier[39] 初始化嵌入，并使用 Adam[40] 作为优化器。为了实现公平评估，按照已发表的论文对每个基线方法进行完整的网格搜索，以找到最佳设置。对于提出的 DPRec，平衡超参数 $\lambda_{\text{align}} \in \{0.1, 0.2, 0.3\}$、超边数量 $N \in \{1, 2, 4, 8, 16, 32, 64, 128, 256\}$、可调节因子 $\alpha \in \{0.1, 0.4, 0.7, 1.0\}$及邻居数量 $k \in \{5, 10, 15, 20, 25, 30\}$中进行了网格搜索。学习率固定为 1E−4，异构图中的 GCN 层数设为 $L = 2$。由于考虑收敛，早停阈值固定为 20。同时，利用验证集上的 *Recall*@20 来作为指标更新最佳纪录。

3.2　性能比较（针对 RQ1）

本文在 3 个数据集上验证了模型性能，表 3 显示了 DPRec 和其他基线在 3 个数据集上的性能比较。表 3 揭示了以下观察结果：DPRec 模型在多个指标上取得了优异性能，超越了传统推荐模型和多模态推荐模型。具体来说，在 Baby, Clothing 数据集的 *Recall*@20 方面，DPRec 分别比最佳基线高出 3.1 个百分点、0.6 个百分点。就 *NDCG*@20 而言，分别显示出 1.1 个百分点、1.2 个百分点的提高。就 *Recall*@10 而言，分别显示出 3.1 个百分点、1.1 个百分点的改进。最后，对于 *NDCG*@10 而言，显示出 2.3 个百分点、0.6 个百分点的提高。对于 Sports 数据集，DPRec 的表现与目前最优模型接近。以上结果验证了 DPRec 的有效性。

从表 3 中可以观察到，传统推荐方法的效果远不及 DPRec，其原因主要是：1）DPRec 整合了多种类型

Table 3 Overall Performance Comparison of the Models

表 3 模型的整体性能比较

模型	Baby				Sports				Clothing			
	R@10	*R*@20	*N*@10	*N*@20	*R*@10	*R*@20	*N*@10	*N*@20	*R*@10	*R*@20	*N*@10	*N*@20
MF-BPR	0.035 7	0.057 5	0.019 2	0.024 9	0.043 2	0.065 3	0.024 1	0.029 8	0.018 7	0.027 9	0.010 3	0.012 6
LightGCN	0.047 9	0.075 4	0.025 7	0.032 8	0.056 9	0.086 4	0.031 3	0.038 7	0.034 0	0.052 6	0.018 8	0.023 6
NCL	0.053 8	0.083 6	0.029 2	0.036 9	0.061 6	0.094 0	0.033 9	0.042 1	0.041 0	0.060 7	0.022 8	0.027 5
VBPR	0.042 3	0.066 3	0.022 3	0.028 4	0.055 8	0.085 6	0.030 7	0.038 4	0.028 1	0.041 5	0.015 8	0.019 2
MMGCN	0.037 8	0.061 5	0.020 0	0.026 1	0.037 0	0.060 5	0.019 3	0.025 4	0.037 8	0.061 5	0.020 0	0.024 7
LATTICE	0.054 7	0.085 0	0.029 2	0.037 0	0.062 0	0.095 3	0.033 5	0.042 1	0.049 2	0.073 3	0.026 8	0.033 0
MMGCL	0.052 2	0.077 8	0.028 9	0.035 5	0.066 0	0.099 4	0.036 2	0.044 8	0.043 8	0.066 9	0.023 9	0.029 7
SLMRec	0.054 0	0.081 0	0.028 5	0.035 7	0.067 6	0.101 7	0.037 4	0.046 2	0.054 0	0.081 0	0.028 5	0.035 7
BM3	0.056 4	0.088 3	0.030 1	0.038 3	0.065 6	0.098 0	0.035 5	0.043 8	0.042 2	0.062 1	0.023 1	0.028 1
DiffMM	0.062 4	0.097 5	0.032 3	0.041 1	0.067 9	0.101 7	0.037 1	0.045 8	0.051 2	0.079 3	0.027 6	0.032 5
MENTOR	0.065 3	0.103 6	0.034 9	0.044 7	0.073 7	0.109 5	0.039 6	0.049 7	0.060 9	0.089 8	0.033 2	0.040 2
LGMRec	0.064 4	0.100 2	0.034 9	0.044 0	0.072 0	0.106 8	0.039 0	0.048 0	0.055 5	0.082 8	0.030 2	0.037 1
IDGenRec	0.065 1	0.103 5	0.034 7	0.044 3	0.068 3	0.106 2	0.037 2	0.047 5	0.059 3	0.087 9	0.032 7	0.039 1
DPRec(本文)	**0.067 3**	**0.104 5**	**0.035 7**	**0.045 2**	0.071 2	0.108 3	0.038 7	0.048 3	**0.061 6**	**0.090 3**	**0.033 4**	**0.040 7**

注：*R*@10，*R*@20，*N*@10，*N*@20 分别代表 *Recall*@10，*Recall*@20，*NDCG*@10，*NDCG*@20；黑体数值表示本文模型达到最好的实验效果。

的数据，比单一数据源的传统推荐更丰富。2）DPRec 等多模态方法通过不同模态理解用户的意图和物品特征，提高了匹配准确性。3）DPRec 可以更好地解决新用户和新物品的推荐问题。4）DPRec 的不同模态信息可以相互补充，提高推荐系统的准确性。

除了传统推荐方法之外，从表 3 还可看出，DPRec 与表中的多模态推荐方法相比也表现出更好的性能。虽然它们均是基于 GCN 的多模态推荐方法，但其他方法大都存在些许不足。例如，LATTICE 根据多模态特征相似性动态学习项目之间的潜在结构。然而，动态生成潜在结构是不必要的，并且占用大量内存，这会影响计算效率。BM3 只学习用户的 ID 表示，而忽略了用户的多模态表示，这是对用户多模态偏好的忽视。SLMRec 通过简单聚合物品多模态特征来获取用户多模态特征，该方法不利于捕捉用户多模态偏好。同时，SLMRec 基于预定义不同模态之间层次相关性生成增强视图，可能会影响其在各种多模态推荐数据集上自监督信号的有效性。VBPR 采用了带有偏好向量的简单设计，更擅长处理较少的多模态特征和密集交互，当项目数量增加时，VBPR 会变得不稳定。MMGCN 模型采用统一的方式融合不同模态的用户偏好，忽略了用户对不同模态偏好程度的差异，每个模态均等的权重可能会导致模型性能变差。MMGCL 中直接屏蔽模态特征的方法可能会导致重要信息丢失。

DPRec 主要优势在于：避免了动态生成项目间的潜在结构，对于用户嵌入表示不仅学习 ID 表示也学习多模态表示，并通过自注意力机制学习每种模态信息对于用户和项目的重要性，考虑到了不同模态偏好程度差异。除此之外，DPRec 的多模态编码层搭建了同时考虑记忆与扩展信息的双视角信息编码层。多模态对齐方面，创新性地提出了使用超图学习到的丰富嵌入表示与双视角信息编码层的嵌入表示进行多级跨模态精准对齐，将多种模态嵌入映射至同一空间之中进行融合，提高整体推荐效果。

3.3 消融实验（针对 RQ2）

在本节中，将通过实验分析 DPRec 的不同组成部分，并开发了 5 个变体，以更好地验证其有效性。在 3 个数据集上的实验结果如表 4 所示，DPRec 优于其 5 个变体：DPRec-o，DPRec-e，DPRec-m，DPRec-c，DPRec-c1。这表明，通过使用 LLM 进行特征表示操作确实可以得到更加准确的特征表示，从而提高推荐性能。同时，仅去掉第 3 级别的模态对齐相比于完全去掉跨模态对齐操作性能有明显提升。这表明，利用超图学习到的嵌入信息进行多级跨模态对齐操作能有效地将不同模态映射至同一空间中。此外，还观察到，当舍弃记忆视角信息编码时，整体推荐效果出现了大幅度下滑，而舍弃扩展视角信息编码时，效果则出现了小范围的下降。这表明，利用超图学习到的嵌入信息进行多级跨模态对齐操作能有效地

Table 4 Effect of Model Performance of Each Module

表 4 各个模块的模型性能影响

数据集	Baby				Sports				Clothing			
	R@10	*R*@20	*N*@10	*N*@20	*R*@10	*R*@20	*N*@10	*N*@20	*R*@10	*R*@20	*N*@10	*N*@20
DPRec	0.067 3	0.104 5	0.035 7	0.045 2	0.071 2	0.108 3	0.038 7	0.048 3	0.061 6	0.090 3	0.033 4	0.040 7
DPRec-o	0.061 9	0.101 6	0.033 6	0.042 1	0.063 4	0.103 7	0.035 6	0.045 3	0.052 6	0.081 2	0.027 5	0.036 6
DPRec-e	0.066 0	0.102 0	0.035 1	0.044 9	0.067 1	0.104 2	0.036 2	0.045 3	0.056 0	0.084 3	0.029 7	0.037 5
DPRec-m	0.023 5	0.038 2	0.012 5	0.016 7	0.031 7	0.046 2	0.021 0	0.025 3	0.012 2	0.027 6	0.011 3	0.019 5
DPRec-c	0.066 5	0.103 9	0.035 1	0.044 6	0.069 5	0.107 5	0.037 9	0.047 5	0.059 2	0.088 5	0.032 5	0.038 9
DPRec-c1	0.066 9	0.104 1	0.035 4	0.044 9	0.070 4	0.107 9	0.038 2	0.048 0	0.060 1	0.089 6	0.033 0	0.039 2

注：*R*@10，*R*@20，*N*@10，*N*@20 分别代表 *Recall*@10，*Recall*@20，*NDCG*@10，*NDCG*@20。DPRec-o 表示使用原始嵌入替换掉由 LLM 处理得到的嵌入；DPRec-e 表示在编码层舍弃扩展视角编码；DPRec-m 表示在编码层舍弃记忆视角编码；DPRec-c 表示去掉超图跨模态对齐操作；DPRec-c1 表示保留前 2 个级别的跨模态对齐，去掉第 3 个级别的跨模态对齐操作。

将不同模态映射至同一空间中。此外，还观察到，当舍弃记忆视角信息编码时，整体推荐效果出现了大幅度下滑，而舍弃扩展视角信息编码时，效果则出现了小范围的下降。这说明用户的历史交互信息仍是捕获用户兴趣偏好的重点，但通过对原始图进行一定范围的图扩展操作，确实也可以捕捉到一些用户额外感兴趣的信息和可能感兴趣的物品。

3.4 模型的参数分析（针对 RQ3）

在本节中，将对 DPRec 方法的超参数进行分析，这些超参数包括项目邻居数量K、平衡超参数λ_{align}、超边数量N和可调节因子α。

3.4.1 项目邻居数量K

为了避免传播不相关项目信息，仅利用最相似的K个项目构建项目-项目图。结果如图 6 所示，可以发现，对于 Baby 数据集和 Sports 数据集，K=5 是项目邻居数量的最佳值。然而，对于 Clothing 数据集，K=10 则是最佳值。这表明K的最优值根据场景不同而异，但可以得到一个确切的结论：盲目使用较大K值构建项目-项目图，会引入很多不相关的项目邻居，造成大量噪声传播，从而导致推荐结果下降。

3.4.2 平衡超参数λ_{align}

在图 7 中展示的 DPRec 模型性能分析中，超参数$\lambda_{\text{align}} \in \{0.1, 0.2, 0.3\}$三个值中进行选择。结果表明，除了 Baby 数据集的 *NDCG*@20 在$\lambda_{\text{align}} = 0.1$时达到最佳外，其余所有数据集的性能评价指标均在$\lambda_{\text{align}} = 0.2$时达到最优。这表明，当$\lambda_{\text{align}} = 0.2$时，模型整体性能最佳，说明模态对齐达到了最好的效果。而当λ_{align}值过大时，可能会导致跨模态对齐效果欠佳，无法将多种模态嵌入有效地映射至同一空间中。

3.4.3 超边数量N与可调节因子α

从图 8(a)中可以观察到，随着超边数量从 1 增至 4，模型性能持续提升。然而，当超边数量继续增加时，模型性能出现了轻微下降。当超边进一步提高至 128 条或更多时，模型性能达到最佳。这一现象表明，当超边数量较少时，模型学习到的嵌入可能不够稳定；而当超边达到一定阈值时，超图卷积学习能够捕捉到各个模态的丰富嵌入信息，通过与编码层的对齐，使其映射至同一空间，从而变得更加丰富和准确。

从图 8(b)可以看出，当可调节因子$\alpha = 0.4$时，可以得到最佳的超图混合模态嵌入。随着α的增加，模型的整体性能开始下降。这表明适当增加α的值有助于获得更准确的融合嵌入，但过大的α可能会导致模态嵌入融合过程混乱，影响模态对齐的准确性，进而导致整体性能的降低。

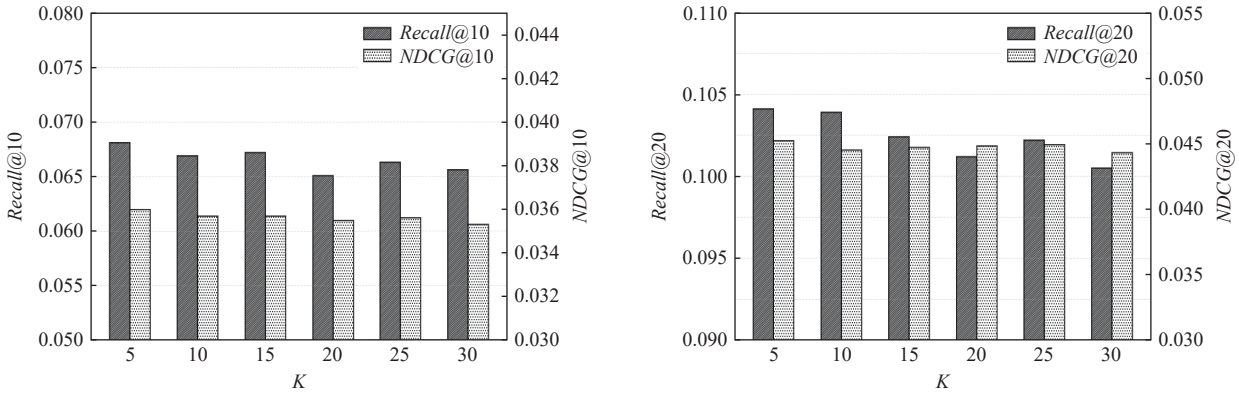
3.5 模型计算开销（针对 RQ4）

本文中模型计算开销主要由 3 部分组成：利用 LLM 将文本与图像转为模态嵌入、构建超图进行消息传递，以及利用超图得到的嵌入与双视角嵌入进行的跨模态对齐操作。以 Clothing 数据集为例，其包括 23 033 个项目，下面讨论其开销。

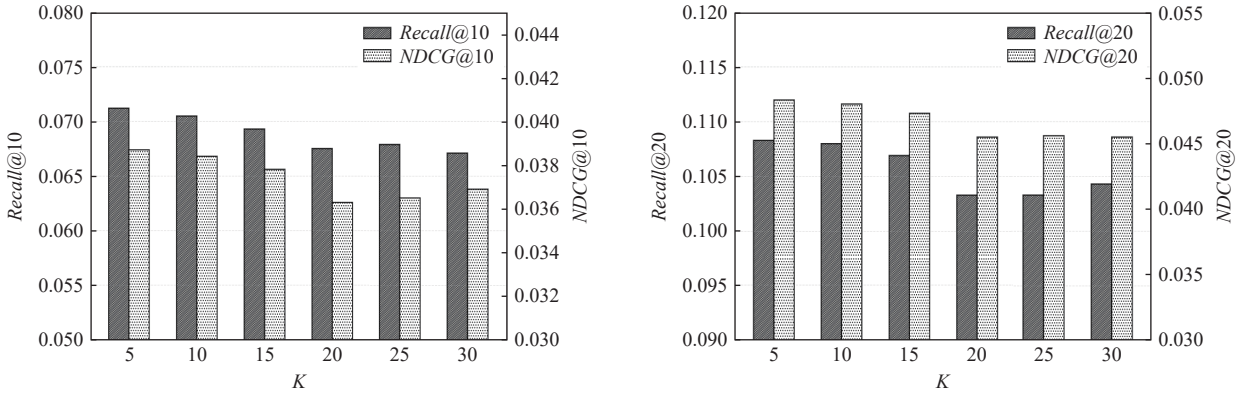
1)对于数据处理部分。利用通义千问(Qwen2.5-Max)对单条数据进行处理时，大概耗时 50 ms，将所有数据处理进行批量并行处理后，总耗时为 19 h 左右(离线完成)。尽管 LLM 的数据处理成本较高，但其带来的特征表示优势，可显著提升推荐表现。

2)对于超图构建及消息传递部分。由式(15)(16)可知，超图初始化阶段的总耗时为 4.3 s，其中超边依赖矩阵计算通过稀疏矩阵乘法优化后，内存占用约为 7.2 GB。当使用 Gumbel-Softmax 进行重新参数化后，可大幅降低项目超边关联矩阵的计算复杂度。由式(17)(18)可知，消息传递阶段在 dropout 对矩阵密度的有效控制下耗时约为 4.8 s。

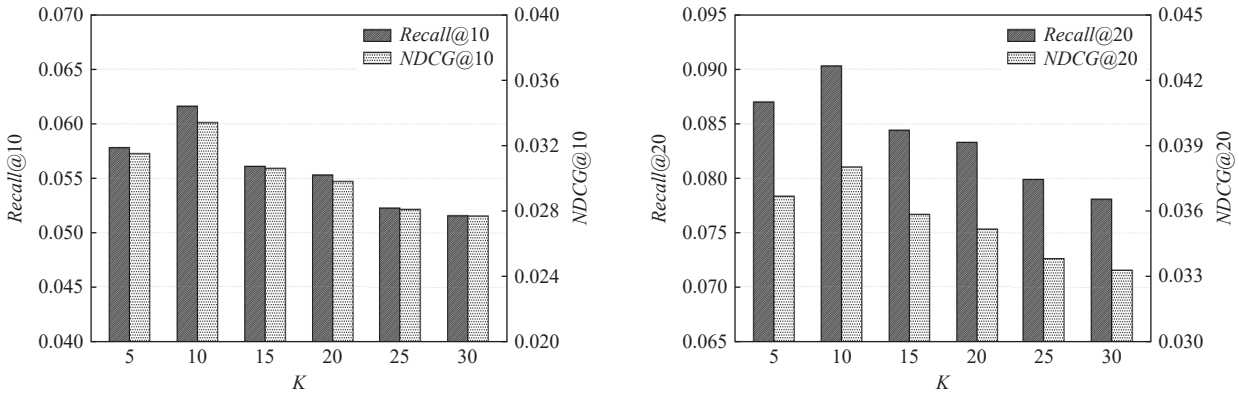
3)跨模态对齐部分。其主要开销在于 3 级 KL



(a) Baby



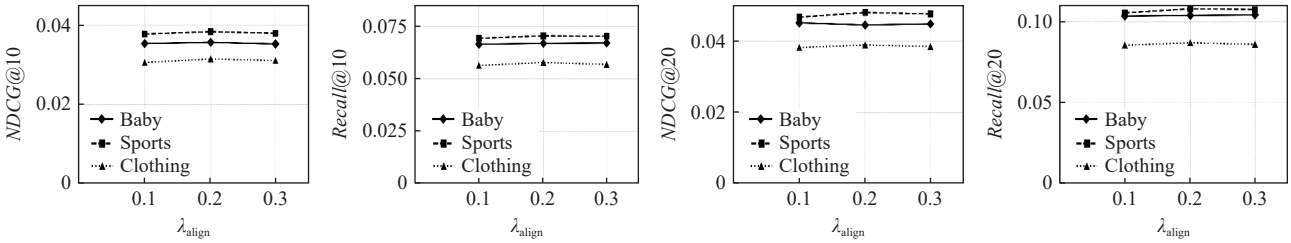
(b) Sports



(c) Clothing

Fig. 6 Performance comparison under neighbor K in different datasets

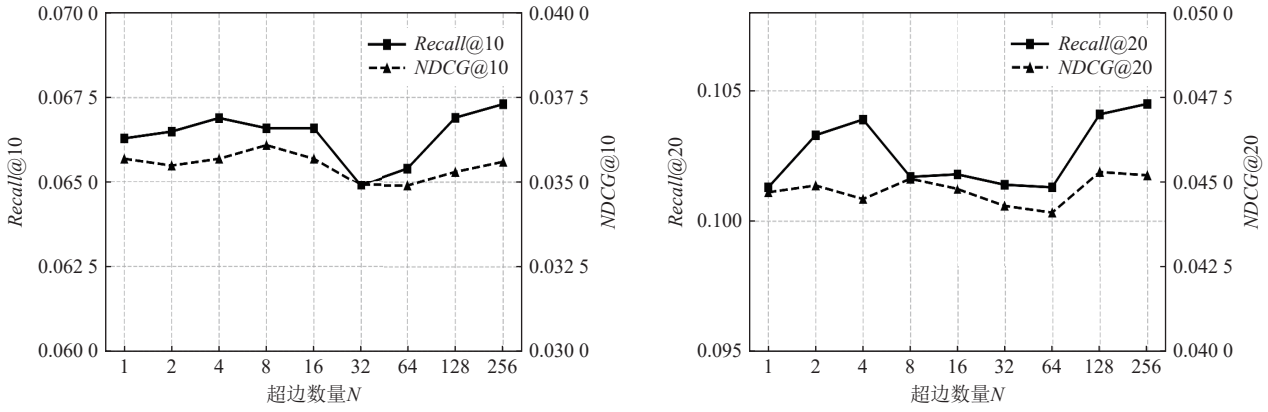
图 6 不同数据集的邻居K下的性能比较



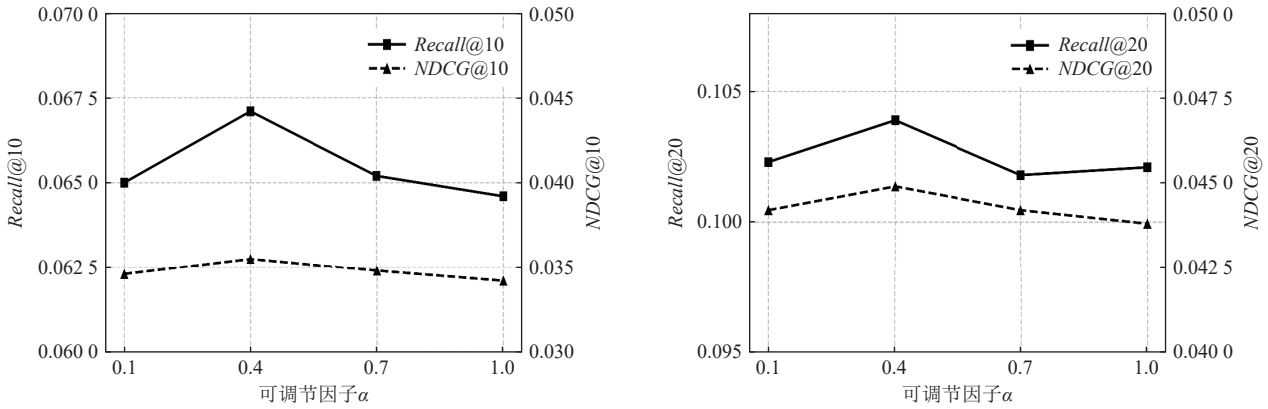
(a) *NDCG*@10的比较 (b) *Recall*@10的比较 (c) *NDCG*@20的比较 (d) *Recall*@20的比较

Fig. 7 Effect of the balancing hyper-parameter λ_{align}

图 7 平衡超参数λ_{align}的影响



(a) 不同超边数量N下的模型性能



(b) 不同可调节因子α下的模型性能

Fig. 8 Performance of models on Baby dataset with different N and α settings

图 8 Baby 数据集上的不同N和α设置下模型的性能

散度计算，其中第 3 级别的复杂度为前 2 级别的 3 倍，3 级对齐的总耗时占单个训练 epoch 的 5%，约为 1.8 s 且内存占用稳定在 3.5 GB 以内。

由此可见，当不将离线数据处理的时间计算在内时，整体的开销比较可观。即这种通过超图结构的多级消息传递与跨模态分布对齐的协同设计，能够在不牺牲模型表达能力的前提下有效解决多模态推荐系统中复杂关系建模的计算效率瓶颈。

4 结 论

本文提出了一种融合记忆与扩展的图卷积神经网络推荐模型 DPRec。具体来说，首先利用 LLM 进行特征表示操作，然后将所得的多模态嵌入投入到经过特别设计的双视角信息编码层中。记忆视角编码通过 LLM 提取的实体词嵌入来丰富用户行为信息，从而获得更细粒度的记忆能力。同时，为有效解决模态空间分布不一致问题，提出了利用超图学习到的模态嵌入与双视角编码层学习到的模态嵌入进行多级跨模态精准对齐，将不同模态信息映射至同一空间中。这有助于在模态融合层获得丰富且准确的用户与项目嵌入，从而提升推荐性能。

作者贡献声明：李佚名、于之晟提出了总体思路和实验方案；于亚新提出指导意见并修改论文；司一廷完成实验并撰写论文；叶育松完成数据采集和预处理。

参 考 文 献

[1] Zhou Jun, Dong Xiaolei, Cao Zhenfu. Research advances on privacy preserving in recommender systems[J]. Journal of Computer Research and Development, 2019, 56(10): 2033−2048 (in Chinese)
(周俊, 董晓蕾, 曹珍富, 等. 推荐系统的隐私保护研究进展[J]. 计算机研究与发展, 2019, 56(10): 2033−2048)

[2] He Ruining, McAuley J. VBPR: Visual Bayesian personalized ranking from implicit feedback[C] //Proc of the 30th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2016: 144−150

[3] Zhang Fuzheng, Lian Defu, Xie Xing, et al. Collaborative knowledge base embedding for recommender systems[C] //Proc of the 22nd ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2016: 353−362

[4] Koren Y, Bell R, Volinsky. Matrix factorization techniques for recommender systems[J]. IEEE Computer, 2009, 42(8): 30−37

[5] Chen Lei, Wu Le, Hong Richang, et al. Revisiting graph based collaborative filtering: A linear residual graph convolutional network approach[C] //Proc of the 34th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2020: 27−34

[6] Mao Kelong, Zhu Jieming, Xiao Xi, et al. UltraGCN: Ultra simplification of graph convolutional networks for recommendation[C] //Proc of the 30th ACM Int Conf on Information & Knowledge Management. New York: ACM, 2021: 1253−1262

[7] Wu Shiwen, Sun Fei, Zhang Wentao, et al. Graph neural networks in recommender systems: A surery[J]. ACM Computing Surveys, 2022, 55(5): 1−37

[8] Wei Yinwei, Wang Xiang, He Xiangnan, 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

[9] Wei Yinwei, Wang Xiang, He Xiangnan, 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

[10] Wang Qifan, Wei Yinwei, Yin Jianhua, et al. DualGNN: Dual graph neural network for multimedia recommendation[J]. IEEE Transactions on Multimedia, 2021, 25(25): 1074−1084

[11] Pagliardini M, Gupta P, Jaggi M. Unsupervised learning of sentence embeddings using compositional n-gram features [C] //Proc of the 2018 Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: ACL, 2018: 528−540

[12] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[C/OL] //Proc of the 3rd Int Conf on Learning Representations. 2014[2024-05-17]. https://www.semanticscholar.org/reader/eb42cf88027de515750f230b23b1a057dc782108

[13] Lin Guojiao, Zhen Meng, Wang Dongjie, et al. GUME: Graphs and user modalities enhancement for long-tail multimodal recommendation[C] //Proc of the 33rd ACM Int Conf on Information and Knowledge Management. New York: ACM, 2024: 1400−1409

[14] Shu Wentao, Li Ruixiao, Sun Tianxiang, et al. Large language models: Principles, implementation, and progress[J]. Journal of Computer Research and Development, 2024, 61(2): 351−361(in Chinese)
(舒文韬, 李睿潇, 孙天祥, 等. 大型语言模型: 原理、实现与发展[J]. 计算机研究与发展, 2024, 61(2): 351−361)

[15] Devlin J, Chang Mingwei, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C] //Proc of the 2019 Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: ACL, 2019: 4171−4186

[16] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C] //Proc of the 31st Int Conf on Neural Information Processing Systems. Long Beach, CA: NeurIPS, 2017: 6000−6010

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

[18] Tan Juntao, Xu Shuyuan, Hua Yingqiang, et al. IDGenRec: LLM-RecSys alignment with textual id learning[C] //Proc of the 47th Int ACM Conf on Research and Development in Information Retrieval. New York: ACM, 2024: 355−364

[19] Bao Keqin, Zhang Jizhi, Zhang Yang, et al. TALLRec: An effectiveand 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

[20] Wang Suhang, Wang Yilin, Tang Jiliang, et al. What your images reveal: Exploiting visual contents for point-of-interest recommendation[C] //Proc of the 26th Int Conf on World Wide Web. New York: ACM, 2017: 391−400

[21] Chen Jingyuan, Zhang Hanwang, He Xiangnan, et al. Attentive collaborative filtering: Multimedia recommendation with item-and component-level attention[C] //Proc of the 40th Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2017: 335−344

[22] 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

[23] Zhou Xin, Shen Zhiqi. A tale of two graphs: Freezing and denoising graph structures for multimodal recommendation[C] //Proc of the 31st ACM Int Conf on Multimedia. New York: ACM, 2023: 935−943

[24] Sun Rui, Cao Xuezhi, Zhao Yan, et al. Multi-modal knowledge graphs for recommender systems[C] //Proc of the 29th ACM Int Conf on Information & Knowledge Management. New York: ACM, 2020: 1405−1414

[25] Lin Jingjing, Ye Zhonglin, Zhao Haixing, et al. Survey on hypergraph nerual networks[J]. Journal of Computer Research and Development, 2024, 61(2): 362−384(in Chinese)
(林晶晶, 冶忠林, 赵海兴, 等. 超图神经网络综述[J]. 计算机研究与发展, 2024, 61(2): 362−384)

[26] Feng Yifan, You Haoxuan, Zhang Zizhao, et al. Hypergraph neural networks[C] //Proc of the 33rd AAAI Conf on Artificial Intelligence and 31st Innovative Applications of Artificial Intelligence Conf and 9th AAAI Symp on Educational Advances in Artificial Intelligence. Palo Alto, CA: AAAI, 2019: 3558−3565

[27] Jang E, Gu Shixiang, Poole B. Categorical reparameterization with Gumbel-Softmax[C/OL] //Proc of the 5th Int Conf Learning Representations. 2016[2023-03-19]. https://www.researchgate.net/publication/309663606_Categorical_Reparameterization_with_Gumbel-Softmax

[28] Xia Lianghao, Huang Chao, Zhang Chuxu. Self-supervised hypergraph transformer for recommender systems[C] //Proc of the 28th ACM SIGKDD Conf on Knowledge Discovery and Data Mining. New York: ACM, 2022: 2100−2109

[29] 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 and 36th Conf on Innovative Applications of Artificial Intelligence and 14th Symp on Educational Advances in Artificial Intelligence. Palo Alto, CA: AAAI, 2024: 8454−8462

[30] Xu Jinfeng, Chen Zheyu, Yang Shuo, et al. MENTOR: Multi-level self-supervised learning for multimodal recommendation[C] //Proc of the 39th Annual AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2025: 12908−12917

[31] McAuley J, Targett C, Shi Qinfeng, et al. Image-based recommendations on styles and substitutes[C] //Proc of the 38th Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2015: 43−52

[32] Rendle S, Freudenthaler C, Gantner Z, et al. BPR: Bayesian personalized ranking from implicit feedback[C] //Proc of the 25th Conf on Uncertainty in Artificial Intelligence. Palo Alto, CA: AUAI, 2009: 452−461

[33] 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

[34] Lin Zihan, Tian Changxin, Hou Yupeng, et al. Improving graph collaborative filtering with neighborhood-enriched contrastive learning[C] //Proc of the ACM Web Conf. New York: ACM, 2022: 2320−2329

[35] Yi Zixuan, Wang Xi, Ounis I, et al. Multi-modal graph contrastive learning for micro-video recommendation[C] //Proc of the 45th Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2022: 1807−1811

[36] Tao Zhulin, Liu Xiaohao, Xia Yewei, et al. Self-supervised learning formultimedia recommendation[J]. IEEE Transactions on Multimedia, 2022, 25(25): 5107−5116

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

[38] Jiang Yangqin, Xia Lianghao, Wei Wei, et al. DiffMM: Multi-modal diffusion model for recommendation[C] //Proc of the 32nd ACM Conf on Multimedia. New York: ACM, 2024: 7591−7599

[39] Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks[C] //Proc of the 13th Int Conf on Artificial Intelligence and Statistics. Brookline, MA: JMLR. org, 2010: 249−256

[40] Kingma D, Ba J. Adam: A method for stochastic optimization[C/OL] //Proc of the 3rd Int Conf on Learning Representations. 2014[2024-04-12]. https://www.semanticscholar.org/reader/a6cb366736791bcccc5c8639de5a8f9636bf87e8



Li Yiming, born in 2000. Master candidate. His main research interests include knowledge graph and recommendation system.

李佚名, 2000 年生。硕士研究生。主要研究方向为知识图谱、推荐系统。



Yu Yaxin, born in 1971. PhD, associate professor, master supervisor. Member of IEEE, ACM and CCF. Her main research interests include data mining and social network.

于亚新, 1971 年生。博士，副教授，硕士生导师。IEEE, ACM, CCF 会员。主要研究方向为数据挖掘、社交网络。



Yu Zhisheng, born in 2001. Master candidate. His main research interests include recommendation system and causal relationship inference.

于之晟, 2001 年生。硕士研究生。主要研究方向为推荐系统、因果关系推理。



Si Yiting, born in 2001. Master candidate. His main research interests include recommendation system and large language model.

司一廷, 2001 年生。硕士研究生。主要研究方向为推荐系统、大语言模型。



Ye Yusong, born in 2001. Master candidate. His main research interests include data mining and recommendation system.

叶育松, 2001 年生。硕士研究生。主要研究方向为数据挖掘、推荐系统。