Software Engineering and Applications 软件工程与应用, 2026, 15(2), 154-167
Published Online April 2026 in Hans. https://www.hanspub.org/journal/sea
https://doi.org/10.12677/sea.2026.152016

Hans汉斯

融合检索增强与动静态兴趣建模的多模态序列推荐方法研究

张若妍

上海理工大学光电信息与计算机工程学院，上海

收稿日期：2026年2月22日；录用日期：2026年4月8日；发布日期：2026年4月21日

摘　要

多媒体内容的快速增长给推荐系统带来新挑战。传统序列推荐过度依赖文本信息，难以充分利用视觉语义；仅依赖时序建模，易忽略用户稳定偏好与物品结构关系；现有可解释推荐缺乏事实支撑，可信度较低。为此，本文在RACL-KAL基础上提出多模态推荐模型MM-RACL-KAL。模型融合文本与图像信息增强物品语义表示，通过检索增强扩展用户行为序列；采用Transformer与GNN实现动静态偏好融合建模，并结合多模态对比学习提升跨模态表示一致性；引入知识锚定大模型，生成有事实依据的可解释推荐。在Amazon-Fashion和MovieLens-Poster数据集上的实验表明，该模型在推荐性能与解释质量上均优于现有方法，验证了其有效性与可扩展性。

关键词

多模态推荐，跨模态对齐，序列推荐，对比学习，图神经网络，可解释性推荐

Research on Multimodal Sequential Recommendation Method Fusing Retrieval Augmentation and Dynamic-Static Interest Modeling

Ruoyan Zhang

School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai

Received: February 22, 2026; accepted: April 8, 2026; published: April 21, 2026

文章引用: 张若妍. 融合检索增强与动静态兴趣建模的多模态序列推荐方法研究[J]. 软件工程与应用, 2026, 15(2): 154-167. DOI: 10.12677/sea.2026.152016

Abstract

The rapid growth of multimedia content poses new challenges to recommender systems. Traditional sequential recommendation relies excessively on textual information, making it difficult to fully utilize visual semantics. It only depends on temporal modeling, which tends to ignore users' stable preferences and item structural relationships. Existing explainable recommendation methods lack factual support and thus have low credibility. To address these issues, this paper proposes a multimodal recommendation model MM-RACL-KAL based on RACL-KAL. The model fuses textual and visual information to enhance item semantic representation and extends user behavior sequences via retrieval augmentation. It adopts Transformer and GNN to achieve dynamic-static preference fusion modeling, combined with multimodal contrastive learning to improve cross-modal representation consistency. A knowledge-anchored large model is introduced to generate explainable recommendations with factual basis. Experiments on Amazon-Fashion and MovieLens-Poster datasets demonstrate that the proposed model outperforms state-of-the-art methods in both recommendation performance and explanation quality, verifying its effectiveness and scalability.

Keywords

Multimodal Recommendation, Cross-Modal Alignment, Sequential Recommendation, Contrastive Learning, Graph Neural Network, Explainable Recommendation

1. 引言

随着在线内容与商品规模不断扩大，推荐系统在信息过滤与用户体验提升中愈发重要[1]。近年来，序列推荐借助时序行为建模取得显著进展，GRU4Rec [2]将 RNN 引入会话推荐，SASRec [3]与 BERT4Rec [4]等基于自注意力的方法进一步增强了长序列建模能力。但在多模态环境下，现有方法仍存在明显不足：一是模态单一，多数模型仅依赖文本与隐式反馈，未充分利用图像等视觉语义信息；二是兴趣建模片面，侧重动态时序依赖，忽略物品结构关系与用户长期稳定偏好[5]；三是可解释性薄弱，模板化与注意力权重解释缺乏事实依据，降低系统透明度与可信度[6]。

针对上述挑战，本文在第三章所提出的文本模态推荐框架 RACL-KAL (融合检索增强、对比学习与知识锚定大语言模型)的基础上进行系统性扩展。RACL-KAL 虽在长序列语义建模、判别性表示学习与可信解释生成方面取得了良好效果，但其输入仍局限于文本模态，未能充分挖掘多模态场景下的数据潜力。为此，本文提出一种面向多模态场景的序列推荐模型——MM-RACL-KAL。为验证所提模型的有效性，本文在 Amazon-Fashion 与 MovieLens-Poster 两个真实多模态数据集上进行了全面的实验评估。结果表明，MM-RACL-KAL 在推荐准确性与解释质量上均显著优于现有基线模型，验证了其在多模态推荐场景下的优越性与实用性。

本文贡献主要体现在以下三个层面：第一，通过融合文本与图像双模态信息，构建更丰富的物品语义表示，以克服单一文本的语义表达瓶颈；第二，设计了一种结合 Transformer 与图神经网络(GNN)的动静态融合用户兴趣编码器，旨在协同建模用户动态兴趣与静态结构偏好；第三，引入多模态对比学习机

制与知识锚定的大语言模型解释生成，旨在提升跨模态语义的一致性，并确保推荐理由兼具真实性与个性化。

2. 相关工作

2.1. 多模态推荐研究

多模态推荐[7]旨在融合文本、图像、音频等多种模态信息以更全面地理解用户偏好。早期多模态推荐工作如 VBPR [8]通过引入视觉特征增强传统协同过滤模型。通常将不同模态的特征向量进行拼接或加权求和，这类方法虽简单直接，但未能深入挖掘模态间的复杂关联与互补语义。随着表示学习技术的发展，研究重点逐渐转向跨模态语义对齐。近年来，CLIP [9]模型通过对比学习实现跨模态语义对齐，为多模态表示学习提供了重要基础。然而，此类方法多侧重于静态的物品表征，未能与用户动态行为序列进行深度融合。近期，一些研究如 MV-RNN [10]尝试在序列建模框架中融合多视图特征，但其模态融合机制仍较为粗糙，未充分考虑不同模态在序列不同阶段的重要性差异。此外，对比学习在多模态推荐中的应用也逐渐受到关注，通过构建跨模态对比损失来拉近相关语义、推开无关内容，从而增强模型的判别能力。然而，现有方法往往将对比学习与序列建模视为两个独立的优化目标，未能实现端到端的协同训练[11] [12]。与上述工作不同，本文提出的多模态检索增强机制(MM-SRA)不仅实现了图文特征的深度融合，更将多模态语义相似性作为序列重构的依据，从而在行为序列层面实现了跨模态的语义扩展。同时，本文将多模态对比学习(MM-EF-CL)有机融入整体训练框架，从物品表示到用户序列进行全链路的一致性约束。

2.2. 序列推荐与动静态用户建模

序列推荐的核心在于捕捉用户兴趣随时间演变的动态规律。基于循环神经网络(RNN)或门控循环单元(GRU)的早期模型在处理长序列时易出现梯度消失或遗忘问题。随后，基于自注意力机制(Transformer)的模型(如 SASRec、BERT4Rec)凭借其强大的长程依赖捕捉能力，成为序列推荐的主流架构。然而，这类方法主要关注用户动态兴趣，其建模完全依赖于显式的历史交互序列，难以刻画用户潜在的、长期稳定的静态偏好[13]。另一方面，基于图神经网络(GNN)的推荐模型(如 LightGCN [14]、NGCF [15] [16])通过挖掘用户–物品交互图中的高阶连通性，能够有效捕获用户与物品间的静态结构关系，但因其无时序的聚合机制，无法对行为发生的先后顺序进行建模[17] [18]。近年来，动静态兴趣融合成为序列推荐的一个重要方向，一些研究尝试结合各种神经网络，或设计记忆网络来同时建模长短期兴趣。然而，现有方法大多采用简单的向量拼接或静态权重相加，未能根据具体上下文自适应地平衡动态与静态兴趣的贡献[19]-[21]。本文设计的用户动静态融合编码器，一方面利用 Transformer 精确刻画短期行为序列的时序模式，另一方面借助轻量级 GNN 从全局交互图中提取用户静态偏好，并创新性地引入可学习的门控融合单元，实现了两种兴趣信号的自适应加权整合，从而更全面、灵活地建模用户兴趣。

2.3. 可解释推荐与大语言模型的应用

可解释推荐旨在揭示推荐决策的内在逻辑，以增强用户信任与系统透明度。传统方法主要包括基于模板的文本生成[22]、基于注意力权重的可视化[23]、以及基于知识图谱的路径推理等。这些方法或因生成文本僵硬、可读性差，或因依赖预设的图谱结构、泛化能力弱，在实际应用中存在诸多限制[24]。近年来，大语言模型(LLM)凭借其强大的自然语言理解和生成能力，为可解释推荐带来了新的范式[25]。研究者尝试将 LLM 作为推荐主体(生成式推荐)或解释生成器。然而，当 LLM 缺乏外部知识约束时，极易产生与用户真实偏好或物品客观属性不符的“幻觉”解释，严重损害可信度[26]。为此，知识锚定

(Knowledge-Anchoring)思想被引入，即利用知识图谱、用户行为日志等结构化信息对 LLM 的生成过程进行约束。现有研究多侧重于利用文本模态的知识(如商品属性、用户评论)进行锚定。在多模态场景下，如何同时利用文本和视觉信息来共同支撑并约束解释生成，仍是一个探索不足的课题。本文提出的知识锚定大模型解释生成机制，不仅融合了用户的多模态行为偏好，还将物品的图文描述作为关键证据注入提示工程，从而引导 LLM 生成同时符合用户视觉与文本偏好的、事实依据充分的个性化解释，有效缓解了多模态场景下的解释幻觉问题。

3. 方法概述

在多模态推荐场景下，用户的兴趣表达不仅来源于历史交互行为本身，还受到物品文本语义、视觉外观以及长期偏好结构等多方面因素的共同影响。给定某个用户的历史交互序列，以及每个物品所对应的文本描述与图像信息[27] [28]，本文旨在学习统一的用户兴趣表示，并预测用户在下一时刻最可能交互的物品，同时生成具有事实依据和个性化特征的推荐解释。

在前序工作 RACL-KAL 中，模型已通过检索增强序列建模与对比学习机制提升了文本模态下的推荐性能与解释可信度。然而，该模型仍局限于单一文本模态，难以充分挖掘多模态信息在语义补充与兴趣刻画方面的潜力。为此，本文在 RACL-KAL 框架基础上提出多模态序列推荐模型 MM-RACL-KAL，通过将文本与图像信息系统性地融入推荐流程，实现对用户动静态兴趣的协同建模。

如图 1 所示，MM-RACL-KAL 由五个核心模块组成：(1) 物品多模态表示模块，用于统一文本与图像语义；(2) 多模态检索增强序列生成模块，用于扩展用户行为序列的语义上下文；(3) 用户动静态兴趣融合编码器，用于同时刻画用户短期动态兴趣与长期稳定偏好；(4) 多模态对比学习模块，用于提升表示判别性与跨模态一致性；(5) 知识锚定的大语言模型解释生成模块，用于生成可信、可理解的推荐理由。上述模块协同工作，最终输出推荐结果及其对应解释。

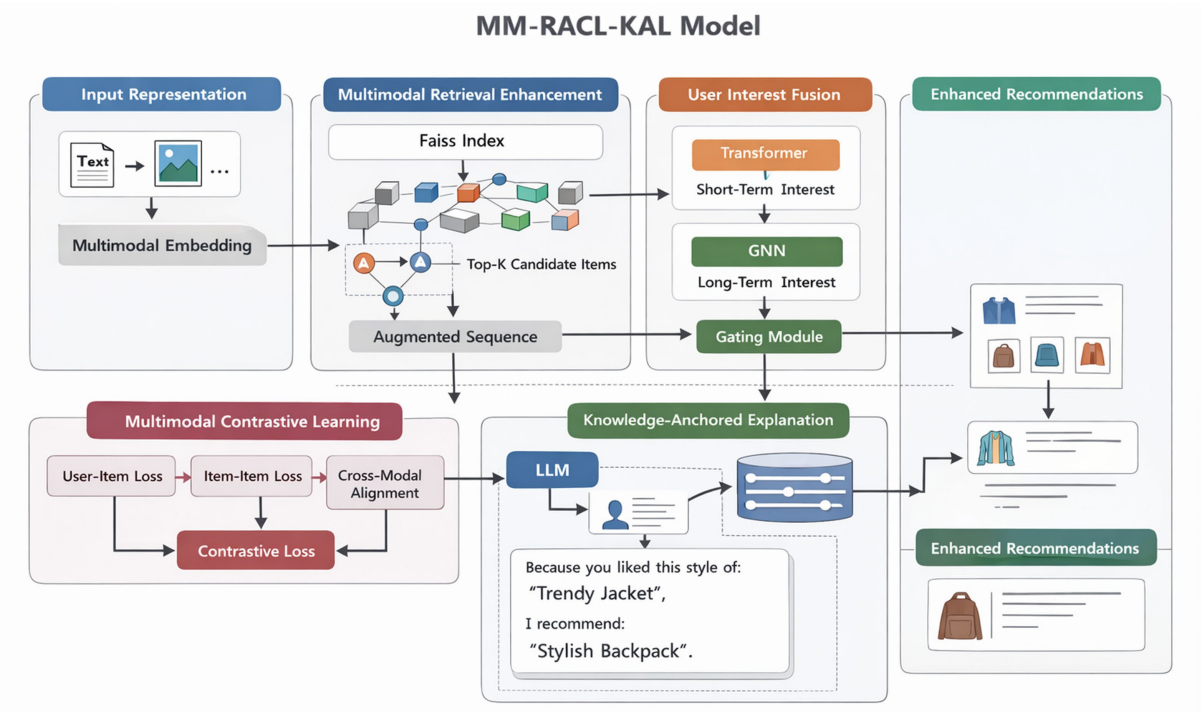

Figure 1. Overall framework of the MM-RACL-KAL model
图 1. MM-RACL-KAL 模型整体框架

3.1. 物品多模态表示

为了充分挖掘物品的语义信息，本模块旨在为每个物品构建一个融合文本与视觉特征的多模态表示向量。具体而言，对于物品 i，我们分别提取其文本描述 x_i^{text} 和商品图像 x_i^{img}。

首先，利用预训练的多模态编码器(如 CLIP)分别获取高质量的初始特征向量：

$$v_i^t = f_{text}\left(x_i^{text}\right), v_i^m = f_{img}\left(x_i^{img}\right) \tag{1}$$

其中 $v_i^t \in \mathrm{R}^d$，$v_i^m \in \mathrm{R}^d$，$f_{text}(\cdot)$ 和 $f_{img}(\cdot)$ 分别为文本编码器和图像编码器。

考虑到不同模态特征可能存在分布差异，我们引入可学习的模态对齐投影矩阵，将其映射到同一语义空间：

$$\hat{v}_i^t = W_t v_i^t, \hat{v}_i^m = W_m v_i^m \tag{2}$$

其中 $W_t, W_m \in \mathrm{R}^{d\times d}$ 为可训练参数。

最终物品 i 的融合表示 e_i 通过自适应加权得到：

$$e_i = \alpha \hat{v}_i^t + (1-\alpha)\hat{v}_i^m \tag{3}$$

其中 α 为可学习的标量权重参数，或在图像特征缺失时退化为 $\alpha = 1$。该表示将作为后续所有模块的基础输入。

3.2. 多模态检索增强序列生成

为解决传统序列建模受限于局部时间窗口、难以捕获长程语义关联的问题，本文提出多模态检索增强序列生成模块。该模块的核心思想是利用物品的多模态表示，从全局物品池中检索语义相似的物品，以增强原始行为序列的语义密度和多样性。

首先基于 3.1 节得到的物品融合表示 $\{e_i\}$，构建全局向量库：

$$\mathcal{D} = \{e_i \mid i \in \text{Items}\} \tag{4}$$

对于序列 $S = [i_1, i_2, \cdots, i_n]$ 中的每个物品 i_t，使用高效的近似最近邻搜索库(如 Faiss)在 $\mathcal{D}$ 中检索其 Top-K 个语义近邻，构成邻域集合 $\mathcal{N}(i_t)$：对序列中每一物品 i 执行相似检索，检索依据为向量间的余弦相似度：

$$\mathcal{N}(i) = \text{FAISS}(e_i, k) \tag{5}$$

对原始序列中的每个位置，以一定概率 p 将其物品替换为来自 $\mathcal{N}(i_t)$ 的语义相似物品，或以 $1-p$ 的概率保留原物品。此过程生成增强后的序列 S'：

$$S' = \{\tilde{i}_1, \tilde{i}_2, \cdots, \tilde{i}_n \mid \tilde{i} \in \mathcal{N}(i) \cup \{i\}\} \tag{6}$$

该机制不仅引入了跨时间点的语义关联，还通过视觉相似性扩展了用户的潜在兴趣范围，为后续建模提供了更丰富的上下文。

3.3. 用户动静态兴趣融合编码器

用户兴趣是动态演变与静态偏好的综合体。为全面建模此双重特性，本模块设计了一个由动态编码器和静态编码器组成的融合架构。

3.3.1. 动态兴趣编码

动态编码器以增强序列 S' 对应的物品向量序列 $\left[e_{\tilde{i}_1}, \cdots, e_{\tilde{i}_n}\right]$ 作为输入，采用多层 Transformer 编码器捕

捉序列内的复杂时序依赖与动态兴趣演变：

$$U^d = \text{Transformer}(S') \tag{7}$$

取最后一层最后一个位置的隐藏状态作为用户的动态兴趣表示：

$$u^d = U^d[-1] \tag{8}$$

3.3.2. 静态兴趣编码

静态编码器旨在从全局用户－物品交互图中挖掘用户长期稳定的结构偏好。首先构建无向交互图 $G=(V,E)$，其中节点 V 包含所有用户和物品，边 E 表示交互关系。采用 LightGCN 进行信息传播：

$$E^{(k)} = \tilde{A}E^{(k-1)} \tag{9}$$

其中 $\tilde{A}$ 为归一化的邻接矩阵，$E^{(k)}$ 为第 k 层的节点嵌入。经过 K 层传播后，对用户节点各层嵌入进行平均池化，得到静态兴趣表示：

$$u^s = \sum_{k=0}^{K} E_u^{(k)} \tag{10}$$

3.3.3. 门控融合机制

为自适应地整合动态与静态兴趣，设计一个门控融合单元：采用门控机制自动平衡动静态贡献：

$$g = \sigma\left(W_g\left[u^d; u^s\right]\right) \tag{11}$$

$$u = g \odot u^d + (1-g) \odot u^s \tag{12}$$

其中 W_g, b_g 为可学习参数，σ 为 Sigmoid 函数，$\odot$ 表示逐元素乘法。最终得到的 u 即为用户的统一兴趣表示。

3.4. 多模态对比学习

在多模态推荐场景中，不同用户与物品之间的语义关系往往呈现出高度复杂且稀疏的分布特征。仅依赖监督信号或单一模态表示，容易导致模型在高维语义空间中判别能力不足，进而影响推荐性能与泛化能力。为此，本文在原有 EF-CL 框架基础上，进一步引入多模态对比学习机制，从多个层次约束表示空间结构，以提升模型在多模态场景下的表示判别性与语义一致性。

具体而言，本文设计了由用户－物品对比、物品－物品对比以及跨模态对齐组成的多层次对比学习目标，分别从交互关系建模、物品语义区分以及跨模态一致性三个角度对表示空间进行联合约束。

3.4.1. 用户－物品对比损失 $\mathcal{L}_{ui}$

用户－物品对比学习旨在从交互行为层面增强用户表示与其真实兴趣物品之间的关联性。该损失通过拉近用户表示与其正样本物品表示之间的距离，同时推远与未交互物品(负样本)的距离，从而强化用户兴趣表示的判别能力。该机制有助于缓解隐式反馈数据中正负样本不平衡带来的训练偏差，使模型能够更准确地区分用户偏好与非偏好物品。其形式定义如下：

$$\mathcal{L}_{ui} = -\log \frac{\exp\left(\frac{\langle u, e_{i^+} \rangle}{\tau}\right)}{\sum_{i^-} \exp\left(\frac{\langle u, e_{i^-} \rangle}{\tau}\right)} \tag{13}$$

3.4.2. 物品–物品对比损失 $\mathcal{L}_{ii}$

在多模态推荐任务中，仅依赖随机负采样往往难以充分刻画物品之间细粒度的语义差异。为此，本文引入基于 MM-SRA 模块检索得到的语义相似物品作为困难负样本，构建物品–物品对比学习目标。该损失通过拉近语义相关物品之间的表示距离，并显式区分语义相近但用户未选择的负样本，从而提升物品表示的判别性与鲁棒性。该机制有助于避免模型仅学习粗粒度相似性，而忽略细微但关键的兴趣差异，其定义如下：

$$\mathcal{L}_{ii} = -\log \frac{\exp\left(\frac{\langle e_i, e_{i^+} \rangle}{\tau}\right)}{\sum_{i^-} \exp\left(\frac{\langle e_i, e_{i^-} \rangle}{\tau}\right)} \tag{14}$$

3.4.3. 跨模态对齐损失 $\mathcal{L}_{mm}$

由于文本与图像模态在特征分布与语义表达方式上存在显著差异，直接融合不同模态特征可能引入语义偏移。为缓解这一问题，本文借鉴 CLIP 的跨模态对齐思想，引入跨模态对比损失，以约束同一物品的文本表示与图像表示在共享语义空间中的一致性。该损失鼓励同一物品的跨模态表示相互靠近，同时拉开不同物品之间的跨模态表示距离，从而提升多模态融合表示的语义稳定性与可解释性，其定义如下：

$$\mathcal{L}_{mm} = -\log \frac{\exp\left(\frac{\langle \hat{v}_i^t, \hat{v}_i^m \rangle}{\tau}\right)}{\sum_{j \neq i} \exp\left(\frac{\langle \hat{v}_i^t, \hat{v}_j^m \rangle}{\tau}\right)} \tag{15}$$

上述三类对比损失从不同层面对表示空间进行约束，具有显著的互补性。用户–物品对比关注兴趣匹配，物品–物品对比强化语义区分，而跨模态对齐则保证多模态融合的一致性。本文将三者加权组合，构成最终的多模态对比学习目标：

$$\mathcal{L} = \lambda_{ui}\mathcal{L}_{ui} + \lambda_{ii}\mathcal{L}_{ii} + \lambda_{mm}\mathcal{L}_{mm} \tag{16}$$

3.5. 知识锚定的多模态可解释生成

在获得推荐结果后，本文进一步利用大语言模型生成自然语言解释。为避免生成内容与用户真实偏好或物品属性不符，引入知识锚定机制。具体而言，将用户历史行为、多模态兴趣特征以及候选物品的图文语义信息组织为结构化提示输入 LLM，引导其生成初始解释文本。随后，通过知识验证与语义一致性约束对生成结果进行校验与修正。需要指出的是，该解释模块作为推荐结果的后处理步骤，不参与推荐打分过程，也不影响模型参数的反向更新，从而在保证解释可信度的同时避免对推荐性能造成干扰。

3.5.1. 多模态知识注入提示工程

构建结构化提示模板，将用户的历史行为序列 S、推荐候选物品 c_j、以及从物品多模态表示中提取的关键语义特征(如文本关键词、视觉主题词)整合输入 LLM。

$$\text{Prompt} = \text{TaskDesc} + \text{History}(S) + \text{Candidate}(c_j) + \text{MultimodalFeatures}(c_j) \tag{17}$$

3.5.2. 生成与约束

LLM 根据提示生成初始解释 γ_{raw}。为确保生成内容不偏离事实，引入一个知识验证层：将生成解释

中的关键实体与知识图谱中该物品的真实属性进行匹配验证，并计算其与用户多模态兴趣表示 u 的语义相似度。若通过验证，则输出解释；否则触发修正机制。

最终，模型的训练总目标由推荐任务的主损失(如交叉熵)与对比学习损失共同构成：

$$\mathcal{L} = \mathcal{L}_{Rec} + \beta \mathcal{L}_{CL} \tag{18}$$

4. 实验

本章围绕所提出的多模态序列推荐模型 MM-RACL-KAL，在真实多模态数据集上开展系统而全面的实验评估，旨在从整体性能、模型结构合理性以及关键模块贡献等多个角度，验证模型设计的有效性与先进性。具体而言，本章实验主要回答以下四个问题：(1) 所提出模型在多模态序列推荐任务中是否整体优于现有代表性方法；(2) 文本与图像模态在不同建模结构下分别发挥何种作用；(3) 模型中各核心模块对性能提升的具体贡献如何；(4) 所提出方法对关键超参数是否具有良好的鲁棒性。围绕上述问题，本文依次介绍数据集与实验设置，对主实验结果进行深入分析，并通过消融实验与参数敏感性分析进一步论证模型设计的合理性，最后讨论模型的效率与工程可行性。

4.1. 数据集与预处理方法

为评估模型在多模态推荐场景下的适用性与稳定性，本文选取了两个公开且被广泛使用的多模态推荐数据集：Amazon-Fashion [29]与 MovieLens-Poster [30]。Amazon-Fashion 数据集来源于亚马逊电商平台，包含用户对时尚类商品的交互记录、文本评论及商品图像信息；MovieLens-Poster 数据集在经典 MovieLens-25M 的基础上，补充了电影的文本简介与海报图像[31]，从而构建了具备图文双模态信息的推荐数据集。

在数据预处理阶段，本文遵循序列推荐领域的通用设置以保证实验结果的可靠性与可复现性。首先，对原始数据进行核心交互过滤，仅保留至少具有 5 次交互行为的用户以及至少被交互 20 次的物品[32]，以缓解极端稀疏问题。其次，为确保多模态信息的完整性，仅保留同时具备有效文本描述与可用图像的物品条目。随后，按照时间戳对每个用户的交互行为进行排序，并采用时间感知的数据划分方式：前 80%的交互用于模型训练，随后 10%用作验证集，最后 10%作为测试集。需要指出的是，MovieLens-Poster 数据集中所报告的用户与物品规模均为上述过滤后的结果。

经过上述预处理后，各数据集的统计信息如表 1 所示。可以看出，两个数据集在用户规模、物品规模及交互密度方面存在显著差异，有助于验证模型在不同数据分布与应用场景下的泛化能力。

Table 1. Data statistics of the dataset
表 1. 数据集数据统计

数据集	用户数	物品数	交互数
Amazon-Fashion	80,123	40,210	705,432
MovieLens-25M	150,321	45,123	20,123,456

4.2. 对比模型与评价指标

为全面验证 MM-RACL-KAL 的性能优势，本文从建模结构与输入模态两个维度选取了多类具有代表性的对比模型，具体包括：(1) 经典序列推荐模型 GRU4Rec 与 SASRec，用于验证仅依赖时序建模的性能上限；(2) 图推荐模型 LightGCN，用于刻画基于交互图的静态协同过滤信号；(3) 单模态深度推荐模型，在统一建模结构的前提下分别仅使用文本或图像特征作为物品表示，并结合 Transformer 或 GNN 进

行建模，包括 Text-SASRec、Image-SASRec、Text-LightGCN 与 Image-LightGCN；(4) 多模态推荐模型 VBPR 与 MMGCN [33]，通过多模态特征融合提升推荐性能；(5) 本文前序模型 RACL-KAL，用于评估多模态扩展带来的增益。

本文采用推荐系统中广泛使用的排序评价指标，包括 HR@K (Hit Ratio)、NDCG@K (Normalized Discounted Cumulative Gain)以及 MRR (Mean Reciprocal Rank)，其中 K 取 10 与 20。所有实验结果均在测试集上计算，并报告三次独立实验运行的平均值。为验证性能提升的统计显著性，采用配对 t 检验进行评估 [34]，显著性水平设为 $p < 0.05$。

4.3. 主实验结果比较与分析

4.3.1. 整体性能对比

表 2 与表 3 分别给出了 Amazon-Fashion 与 MovieLens-Poster 数据集上的推荐性能对比结果。从结果可以看出，MM-RACL-KAL 在两个数据集上的 HR、NDCG 以及 MRR 指标均显著优于所有对比模型，验证了所提出方法在多模态序列推荐任务中的整体有效性。

与经典序列推荐模型(如 GRU4Rec 与 SASRec)相比，MM-RACL-KAL 在两个数据集上均取得了显著性能提升，表明仅依赖用户历史行为序列难以充分刻画复杂兴趣模式；与图推荐模型 LightGCN 相比，结果进一步说明在序列推荐任务中显式建模用户行为的时间顺序与动态演化具有不可替代的重要性。与此同时，相较于现有多模态推荐方法(如 VBPR 与 MMGCN)，MM-RACL-KAL 仍保持明显优势，说明简单的多模态特征融合不足以充分释放多模态信息的潜力。

Table 2. Comparison of recommendation performance on the Amazon-Fashion dataset
表 2. Amazon-Fashion 数据集上的推荐性能对比

模型	HR@10	NDCG@10	MRR	HR@20	NDCG@20
GRU4Rec	0.3124	0.1867	0.1501	0.4321	0.2265
SASRec	0.4012	0.2543	0.2015	0.5210	0.2987
LightGCN	0.3856	0.2411	0.1898	0.4987	0.2854
VBPR	0.4233	0.2689	0.2132	0.5433	0.3122
MMGCN	0.4415	0.2814	0.2256	0.5621	0.3278
RACL-KAL	0.4588	0.2956	0.2389	0.5789	0.3412
MM-RACL-KAL	**0.5021**	**0.3317**	**0.2614**	**0.6223**	**0.3798**

Table 3. Comparison of recommendation performance on the MovieLens-Poster dataset
表 3. MovieLens-Poster 数据集上推荐性能对比

模型	HR@10	NDCG@10	MRR	HR@20	NDCG@20
GRU4Rec	0.2876	0.1721	0.1389	0.3987	0.2078
SASRec	0.3789	0.2387	0.1894	0.4876	0.2789
LightGCN	0.3654	0.2254	0.1787	0.4765	0.2678
VBPR	0.4012	0.2532	0.2012	0.5123	0.2965
MMGCN	0.4189	0.2678	0.2145	0.5321	0.3123
RACL-KAL	0.4365	0.2833	0.2289	0.5543	0.3287
MM-RACL-KAL	**0.4789**	**0.3245**	**0.2576**	**0.5987**	**0.3621**

4.3.2. 单模态与结构对比分析

为进一步分析不同模态信息及建模结构对推荐性能的影响，本文在主实验中引入了基于单一模态输入的 Transformer 与 GNN 对比模型。具体的实验结果如表 4 所示。

Table 4. Comparison of single-modal model performance on the Amazon-Fashion dataset
表 4. Amazon-Fashion 数据集上单模态模型性能对比

模型	输入模态	结构	HR@10	NDCG@10	MRR
Text-SASRec	文本	Transformer	0.3897	0.2468	0.1962
Image-SASRec	图像	Transformer	0.3615	0.2284	0.1813
Text-LightGCN	文本	GNN	0.3726	0.2359	0.1865
Image-LightGCN	图像	GNN	0.3489	0.2196	0.1748
SASRec	ID	Transformer	0.4012	0.2543	0.2015
LightGCN	ID	GNN	0.3856	0.2411	0.1898
RACL-KAL	文本	Transformer	0.4588	0.2956	0.2389
MM-RACL-KAL	图像 + 文本	Trans + GNN	**0.5021**	**0.3317**	**0.2614**

实验结果表明，仅使用文本特征的模型性能优于仅使用图像特征的模型，说明在当前场景下，文本对物品语义与用户偏好的表达更直接稳定，视觉信息虽可提供补充语义，但单独使用难以完整支撑用户决策。在相同模态下，基于 Transformer 的序列模型性能略优于 GNN 图模型，说明时序建模对捕捉动态兴趣演化至关重要；而 GNN 在单模态下仍具竞争力，体现其在捕获长期稳定偏好与高阶协同关系上的优势，验证了本文动静态融合机制的合理性。多模态融合模型性能显著优于单模态模型，且在两个数据集上表现一致，证明文本与视觉模态具有强互补性，多模态融合是提升推荐效果的关键。

综上，单一模态或单一结构难以完整建模复杂用户兴趣。本文提出的 MM-RACL-KAL 通过多模态融合与 Transformer-GNN 动静态协同建模，有效弥补了单模型缺陷，实现最优性能，充分验证了模型设计的合理性与有效性。

4.4. 核心机制消融实验分析

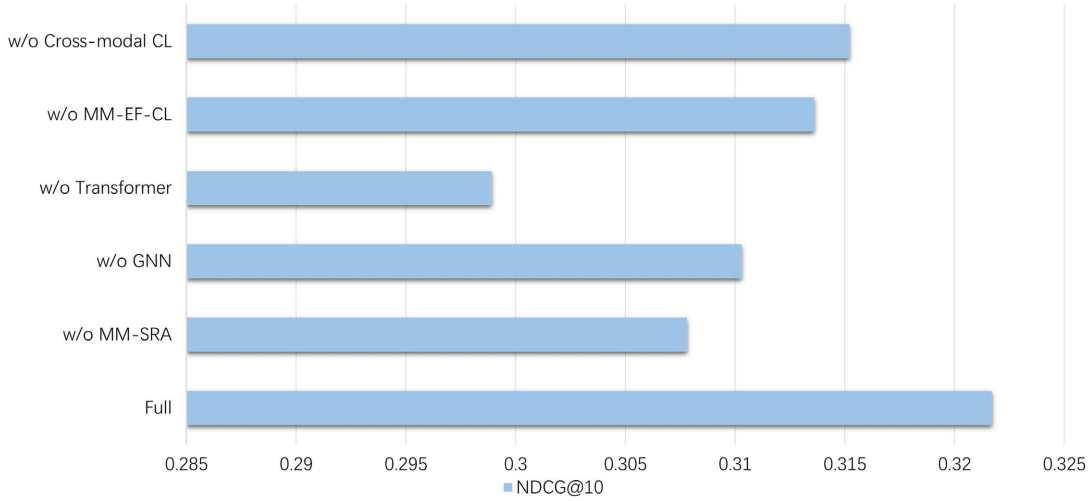

Figure 2. Ablation study results of different core modules (Amazon-Fashion, NDCG@10)
图 2. 不同核心模块的消融实验结果(Amazon-Fashion, NDCG@10)

为验证模型中各核心模块的有效性，本文在 Amazon-Fashion 数据集上进行了系统的消融实验，结果如图 2 所示。实验分别移除图像模态、多模态检索增强模块(MM-SRA)、静态兴趣编码器(GNN)以及多模态对比学习模块(MM-EF-CL)。

图 2 给出了 MM-RACL-KAL 不同核心机制的消融实验结果。从整体结果可以观察到，移除任一关键模块均会导致推荐性能下降，说明各组件在整体框架中均对性能提升起到重要作用。

为验证模型中各核心模块的独立贡献及逐步融合的增益效果，设置基础模型与逐步加模块两个实验组，同时补充对比学习的细分实验，分析权重分配与样本构造的影响。结果如表 5 所示。设置基础模型(Base)：以 RACL-KAL 为基础，移除所有多模态相关机制，仅保留文本模态、Transformer 时序建模与基础推荐损失，作为消融实验的基准。在此基础上逐步添加核心模块，依次为：多模态表示融合(M)、多模态检索增强(MM-SRA)、动静态兴趣融合(GNN + 门控，S)、多模态对比学习(MM-EF-CL, C)，最终得到完整模型 MM-RACL-KAL (Base + M + S + RA + C)。

Table 5. Ablation study results
表 5. 消融实验结果

模型配置	核心模块	NDCG@10	相对基础模型提升
Base	文本模态 + Transformer + 基础推荐损失	0.2813	—
Base + M	基础模型 + 多模态表示融合	0.2956	5.08%
Base + M + RA	Base + M + 多模态检索增强	0.3128	11.20%
Base + M + RA + S	Base + M + RA + 动静态兴趣融合	0.3245	15.36%
MM-RACL-KAL	完整模型 + 多模态对比学习	0.3317	17.92%

从阶梯式消融结果可以看出，每个核心模块的添加均能带来稳定且显著的性能提升，无模块叠加导致的边际效益骤降，证明各模块均具有独立的贡献价值：多模态表示融合实现了图文语义的互补，是多模态扩展的基础；多模态检索增强通过语义扩展缓解了序列稀疏问题，带来了最大的单模块增益；动静态兴趣融合弥补了单一时序建模的不足，进一步挖掘了用户长期偏好；多模态对比学习则通过表示空间约束，提升了跨模态一致性与特征判别性，实现了性能的最终优化。

为分析对比学习的权重分配与样本构造对性能的具体影响，针对 3.4 节的多模态对比学习模块，设计两组细分实验：权重分配实验和样本构造实验。具体实验结果如表 6 所示。

Table 6. Detailed experimental results of multimodal contrastive learning (Amazon-Fashion, NDCG@10)
表 6. 多模态对比学习细分实验结果(Amazon-Fashion, NDCG@10)

实验类型	设置	NDCG@10	相对基础配置提升
基础配置	无对比学习(Base + M + RA + S)	0.3245	—
权重分配	仅($\mathcal{L}_{ui}$)	0.3278	1.02%
权重分配	仅($\mathcal{L}_{ii}$)	0.3289	1.36%
权重分配	仅($\mathcal{L}_{mm}$)	0.3267	0.68%
样本构造	$\mathcal{L}_{ii}$ 采用随机负样本(最优权重)	0.3291	1.42%
样本构造	$\mathcal{L}_{ii}$ 采用检索困难负样本(最优权重)	0.3317	2.22%

从结果可以得出两个关键结论：权重分配：三类对比损失项均能带来性能提升，且联合使用的效果显著优于单一损失项，证明用户－物品对比、物品－物品对比与跨模态对齐的互补性；验证了多模态对

比学习框架的合理性。样本构造：采用检索困难负样本的物品–物品对比损失，性能显著优于随机负样本，说明困难负样本能够更精准地刻画物品间的细粒度语义差异，提升特征判别性，避免模型学习粗粒度相似性，这也是对比学习模块的核心增益来源。

4.5. 参数敏感性分析

为分析多模态检索增强模块中关键超参数对模型性能的影响，本文在 Amazon-Fashion 数据集上考察了语义检索近邻规模 Top-K 以及序列替换概率 p 对 NDCG@10 的影响，实验结果如图 3 与图 4 所示。

Figure 3. Recommendation performance under different Top-K settings
图 3. 不同 Top-K 设置下的推荐性能

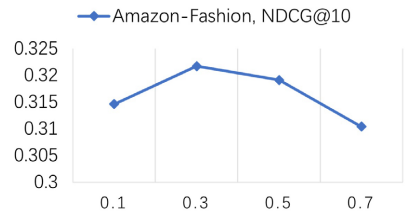

Figure 4. Recommendation performance under different replacement probabilities p
图 4. 不同替换概率 p 下的推荐性能

实验结果表明，当 Top-K 取适中值时，模型性能达到最优，过小难以充分发挥语义扩展作用，过大则可能引入噪声；替换概率 p 在适中范围内同样能够有效提升性能，而过高的替换比例会破坏原始行为序列的时序一致性。整体来看，所提出方法在较宽的参数区间内均能保持稳定性能，体现了良好的鲁棒性。

4.6. 效率与开销分析

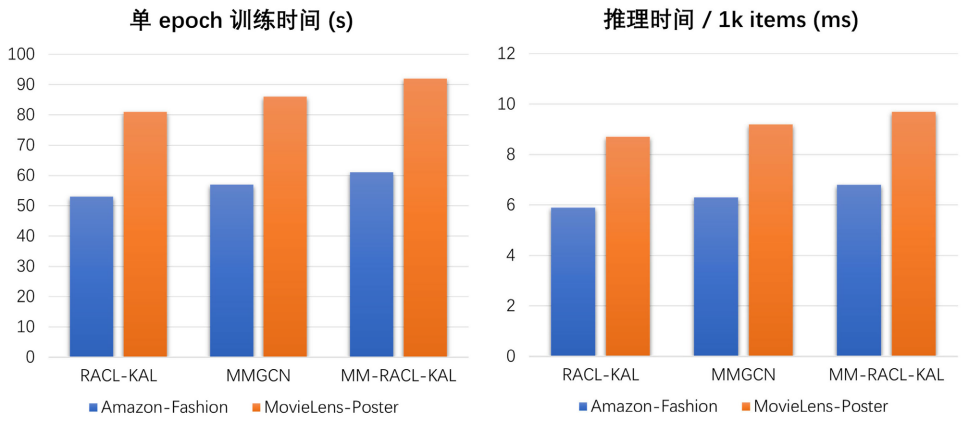

Figure 5. Comparison of computational efficiency of different models
图 5. 不同模型的计算效率比较

在效率与计算开销方面，图 5 的结果可以表明 MM-RACL-KAL 在引入多模态建模与检索增强机制的同时，仍保持了良好的工程可行性。模型中的静态 GNN 用户与物品嵌入可在离线阶段预先计算，避免了对在线推理过程的额外负担；多模态相似检索所依赖的向量索引亦通过离线构建完成，在线阶段仅执行高效的 Top-K 近邻搜索，从而有效控制了时延开销。尽管模型引入了基于 CLIP 的多模态特征表示，但相关计算主要集中于训练阶段，整体训练规模仍处于可接受范围。实验结果表明，在显著提升推荐性能与解释质量的同时，所提出模型在在线推理效率与系统扩展性方面均表现稳定，具备实际部署潜力。

5. 结论

本文针对多模态推荐中语义表达不足、用户兴趣建模不全面、推荐解释可信度有限等问题，提出多模态序列推荐与可解释框架 MM-RACL-KAL。该框架在 RACL-KAL 基础上引入文本与图像双模态信息，通过多模态检索增强扩展用户行为序列，结合 Transformer 与图神经网络协同建模用户动静态兴趣；同时设计多模态对比学习提升跨模态语义一致性，并借助知识锚定的大语言模型生成可信推荐解释。

在 Amazon-Fashion、MovieLens-Poster 数据集上的实验表明，MM-RACL-KAL 在 HR、NDCG、MRR 等指标上优于主流序列推荐与多模态推荐方法，消融实验验证了关键模块的有效性，且模型可通过离线计算与高效检索控制在线推理开销，具备工程可行性。

该方法仍存在局限：仅聚焦文本与图像模态，未充分融合视频、音频等复杂模态；大语言模型解释生成依赖固定提示与校验机制，效率与灵活性有待提升。未来可研究轻量化多模态表示与对齐、引入更丰富外部知识，并探索交互式、可控的推荐解释机制。

参考文献

[1] 于蒙, 何文涛, 周绪川, 等. 推荐系统综述[J]. 计算机应用, 2022, 42(6): 1898-1913.

[2] Hidasi, B., Karatzoglou, A., Baltrunas, L., *et al*. (2016) Session-Based Recommendations with Recurrent Neural Networks. *International Conference on Learning Representations*, San Juan, 2-4 May 2016, 1-14.

[3] Kang, W. and McAuley, J. (2018) Self-Attentive Sequential Recommendation. 2018 *IEEE International Conference on Data Mining* (*ICDM*), Singapore, 17-20 November 2018, 197-206. https://doi.org/10.1109/icdm.2018.00035

[4] Sun, F., Liu, J., Wu, J., Pei, C., Lin, X., Ou, W., *et al*. (2019) BERT4Rec. *Proceedings of the* 28*th ACM International Conference on Information and Knowledge Management*, Beijing, 3-7 November 2019, 1441-1450. https://doi.org/10.1145/3357384.3357895

[5] 吴正洋, 汤庸, 刘海. 个性化学习推荐研究综述[J]. 计算机科学与探索, 2022, 16(1): 21-40.

[6] Zhang, Y., Lai, G., Zhang, M., Zhang, Y., Liu, Y. and Ma, S. (2014) Explicit Factor Models for Explainable Recommendation Based on Phrase-Level Sentiment Analysis. *Proceedings of the* 37*th International ACM SIGIR Conference on Research & Development in Information Retrieval*, Queensland, 6-11 July 2014, 83-92. https://doi.org/10.1145/2600428.2609579

[7] 陈烨, 周刚, 卢记仓. 多模态知识图谱构建与应用研究综述[J]. 计算机应用研究, 2021, 38(12): 3535-3543.

[8] Liu, B., Liu, X., Luo, Q., *et al*. (2025) Variational Bayesian Personalized Ranking. arXiv:2503.11067.

[9] Radford, A., Kim, J.W., Hallacy, C., *et al*. (2021) Learning Transferable Visual Models from Natural Language Supervision. 38*th International Conference on Machine Learning* (*ICML*), Virtual Event, 18-24 July 2021, 8748-8763.

[10] Cui, Q., Wu, S., Liu, Q., Zhong, W. and Wang, L. (2020) MV-RNN: A Multi-View Recurrent Neural Network for Sequential Recommendation. *IEEE Transactions on Knowledge and Data Engineering*, **32**, 317-331. https://doi.org/10.1109/tkde.2018.2881260

[11] Lu, J. and Yamashita, H. (2025) MORE: Modality-Embracing Contrastive Learning for Multimodal Recommendation. *International Conference on Multimodal Interaction* (*ICMI*), Canberra, 13-17 October 2025, 1-9.

[12] Chen, H., Li, J., Zhang, X., *et al*. (2023) Multi-Modal Self-Supervised Learning for Recommendation. *ACM International Conference on Multimedia Retrieval* (*ICMR*), Thessaloniki, 12-15 June 2023, 215-223.

[13] 吕学强, 王夏雨, 马登豪. 面向推荐系统的用户兴趣建模综述[J]. 计算机工程与应用, 2025, 61(21): 15-29.

[14] He, X., Deng, K., Wang, X., Li, Y., Zhang, Y. and Wang, M. (2020) LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation. *Proceedings of the* 43*rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, Virtual Event, 25-30 July 2020, 639-648. https://doi.org/10.1145/3397271.3401063

[15] Wang, X., He, X., Wang, M., Feng, F. and Chua, T. (2019) Neural Graph Collaborative Filtering. *Proceedings of the* 42*nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, Paris, July 21-25, 2019, 165-174. https://doi.org/10.1145/3331184.3331267

[16] Kipf, T.N. and Welling, M. (2017) Semi-Supervised Classification with Graph Convolutional Networks. 5*th International Conference on Learning Representations* (*ICLR*), Toulon, 24-26 April 2017.

[17] Chen, Y., Liu, Z., Li, J., McAuley, J. and Xiong, C. (2022) Intent Contrastive Learning for Sequential Recommendation. *Proceedings of the ACM Web Conference* 2022, Lyon, 25-29 April 2022, 2172-2182. https://doi.org/10.1145/3485447.3512090

[18] 吴静, 谢辉, 姜火文. 图神经网络推荐系统综述[J]. 计算机科学与探索, 2022, 16(10): 2249-2263.

[19] 孙文彬, 林伟, 方滨兴. 基于门控融合的长短期兴趣序列推荐方法[J]. 计算机学报, 2024, 47(9): 1892-1908.

[20] Zhang, C., Yao, L. and Sun, A. (2020) FISSA: Fusing Item Similarity Models with Self-Attention Networks for Sequential Recommendation. *ACM International Conference on Information and Knowledge Management*, Virtual Event, 22-26 September 2020, 3412-3421.

[21] Li, J., Wang, X. and Hu, X. (2023) Adaptive Gating Fusion for Dynamic-Static Interest in Sequential Recommendation. *Knowledge-Based Systems*, **275**, Article 110789.

[22] Li, L., Zhang, Y. and Chen, L. (2020) Generate Neural Template Explanations for Recommendation. *Proceedings of the* 29*th ACM International Conference on Information & Knowledge Management*, Virtual Event, 19-23 October 2020, 1299-1308. https://doi.org/10.1145/3340531.3411992

[23] Jain, S. and Wallace, B.C. (2019) Attention Is Not Explanation. *Proceedings of the* 2019 *Conference of the North*, Minneapolis, 2 June-7 June 2019, 3543-3556. https://doi.org/10.18653/v1/n19-1357

[24] Wang, X., He, X., Cao, Y., *et al*. (2019) Reinforcement Knowledge Graph Reasoning for Explainable Recommendation. *ACM International Conference on Information and Knowledge Management*, Paris, 21-25 July 2019, 2070-2078.

[25] 高广尚. 可解释推荐模型中的可解释性方法研究综述[J]. 数据分析与知识发现, 2024, 8(8/9): 6-19.

[26] Gong, J., Cheng, M., Shen, H., Vandenbussche, P., Jenq, J. and Eldardiry, H. (2025) Visual Zero-Shot E-Commerce Product Attribute Value Extraction. *Proceedings of the* 2025 *Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics*: *Human Language Technologies* (*Volume* 3: *Industry Track*), Albuquerque, 30 April 2025, 460-469. https://doi.org/10.18653/v1/2025.naacl-industry.38

[27] 张瑞, 卞志鹏. 面向推荐系统的多模态生成研究综述[J]. 计算机科学与探索, 2025, 19(12): 3224-3242.

[28] 吴晔, 陆俊霖. 大模型驱动的多模态信息生成与信息推荐[J]. 河南师范大学学报(自然科学版), 2025, 53(5): 145-151+181.

[29] Hou, M., Wu, L., Chen, E., Li, Z., Zheng, V.W. and Liu, Q. (2019) Explainable Fashion Recommendation: A Semantic Attribute Region Guided Approach. *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, Macao, 10-16 August 2019, 4681-4688. https://doi.org/10.24963/ijcai.2019/650

[30] Harper, F.M. and Konstan, J.A. (2015) The Movielens Datasets. *ACM Transactions on Interactive Intelligent Systems*, **5**, 1-19. https://doi.org/10.1145/2827872

[31] Xia, L., Yang, Y., Chen, Z., *et al*. (2024) Movie Recommendation with Poster Attention via Multi-Modal Transformer Feature Fusion. arXiv:2407.09157.

[32] Mancino, A.C.M., Attimonelli, M., Di Fazio, A., Malitesta, D. and Di Noia, T. (2025) Standard Practices for Data Processing and Multimodal Feature Extraction in Recommendation with Datarec and Ducho (d&d4rec). *Proceedings of the Nineteenth ACM Conference on Recommender Systems*, Prague Czech, 22-26 September 2025, 1432-1434. https://doi.org/10.1145/3705328.3748009

[33] Wei, Y., Wang, X., Nie, L., He, X., Hong, R. and Chua, T. (2019) MMGCN: Multi-Modal Graph Convolution Network for Personalized Recommendation of Micro-Video. *Proceedings of the* 27*th ACM International Conference on Multimedia*, Nice, 21-25 October 2019, 1437-1445. https://doi.org/10.1145/3343031.3351034

[34] Anelli, V.W., Bellogin, A., Ferrara, A., Malitesta, D., Merra, F.A., Pomo, C., *et al*. (2021) Elliot: A Comprehensive and Rigorous Framework for Reproducible Recommender Systems Evaluation. *Proceedings of the* 44*th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Virtual Event, 11-15 July 2021, 2405-2414. https://doi.org/10.1145/3404835.3463245