Computer Science and Application 计算机科学与应用, 2026, 16(3), 558-571
Published Online March 2026 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2026.163084

Hans汉斯

引入模态自适应融合机制的多模态知识图谱推荐方法

李若雯*，徐晓婧

河北地质大学信息工程学院，河北 石家庄

收稿日期：2026年1月27日；录用日期：2026年2月26日；发布日期：2026年3月5日

摘　要

知识图谱(KG)已广泛用于提升推荐系统性能，特别是在缓解数据稀疏问题方面。然而，现有多模态知识图谱推荐方法多采用固定或简单的融合策略(如拼接或线性加权)，难以充分建模不同模态对推荐决策的动态贡献，限制了异构信息的有效利用。为此，本文提出一种新颖的多模态自适应融合知识图谱注意力网络(MAF-KGAT)。该模型引入模态自适应门控机制，通过门控网络动态学习图像特征与结构嵌入的融合权重，实现更细粒度的模态融合。我们还结合预训练的CLIP图像编码器提取判别性视觉特征，并在KGAT架构基础上融合多模态信息以增强图谱传播能力。在AmazonBooks和MovieLens数据集上，MAF-KGAT在Recall@20和NDCG@20上分别提升了4.2%和3.5% (AmazonBooks)以及3.6%和2.9% (MovieLens)，尤其在冷启动场景下表现出更强的推荐能力，验证了其在个性化推荐中的有效性与鲁棒性。该方法已在电商和娱乐平台中验证，可有效缓解数据稀疏问题。

关键词

推荐系统，知识图谱，图神经网络，图像特征，多模态融合

Modality-Adaptive Fusion with Knowledge Graph Attention Network for Multi-Modal Recommendation

Ruowen Li*, Xiaojing Xu

School of Information Engineering, Hebei GEO University, Shijiazhuang Hebei

Received: January 27, 2026; accepted: February 26, 2026; published: March 5, 2026

*通讯作者。

文章引用: 李若雯, 徐晓婧. 引入模态自适应融合机制的多模态知识图谱推荐方法[J]. 计算机科学与应用, 2026, 16(3): 558-571. DOI: 10.12677/csa.2026.163084

Abstract

Knowledge Graph (KG) has been widely used to improve the performance of recommendation systems, especially in alleviating the problem of data sparsity. However, most of the existing multimodal knowledge graph recommendation methods adopt fixed or simple fusion strategies (such as concatenation or linear weighting), making it difficult to fully model the dynamic contributions of different modalities to recommendation decisions and limiting the effective utilization of heterogeneous information. To this end, this paper proposes a novel multimodal adaptive fusion knowledge graph attention network (MAF-KGAT). This model introduces a modal adaptive gating mechanism. Through the gated network, it dynamically learns the fusion weights of image features and structural embeddings to achieve more fine-grained modal fusion. We also combine the pre-trained CLIP image encoder to extract discriminative visual features, and fuse multimodal information based on the KGAT architecture to enhance the graph propagation ability. Experiments on two real datasets, Amazonbook and MovieLens, show that MAF-KGAT significantly outperforms existing multimodal and knowledge graph recommendation methods in terms of metrics such as Recall and NDCG, especially demonstrating stronger recommendation capabilities in cold start scenarios. Its effectiveness and robustness in personalized recommendation have been verified.

Keywords

Recommendation System, Knowledge Graph, Graph Neural Network, Image Feature, Multimodal Fusion

1. 引言

推荐系统是现代在线平台中缓解信息过载、提升用户体验的关键技术[1]，其核心在于对用户与物品之间的交互关系进行建模，从而实现个性化推荐。然而，传统协同过滤方法高度依赖历史交互数据，普遍面临数据稀疏性和冷启动问题，在真实应用场景中性能受限。

为缓解上述问题，研究者将知识图谱(Knowledge Graphs, KGs)引入推荐系统，通过显式建模实体及其关系来补充用户－物品交互信息。代表性方法如 KGAT [2]和 KGCN [3]利用图神经网络(Graph Neural Networks, GNNs) [4]对知识图谱进行建模，实现高阶语义关系的传播与聚合，在一定程度上提升了推荐性能。

尽管基于知识图谱的推荐方法在结构化语义建模方面取得了显著进展，但其大多侧重于关系结构传播，忽视了现实场景中广泛存在的非结构化多模态信息，如图像和文本内容。在电子商务和内容推荐场景中，这类信息对用户决策具有重要影响。例如，在 Amazon-Book 数据集中，大多数商品包含图像信息，其视觉风格和内容呈现方式直接影响用户的点击与购买行为。因此，仅依赖结构化知识难以全面刻画物品特征，如何有效融合多模态信息成为知识图谱推荐的重要研究问题。

近年来，一些研究尝试将外部知识图谱与多模态数据相结合，以增强推荐模型的表达能力和泛化性能[5] [6]。典型方法如 MMGCN [7]和 MKGAT [8]通过多通道建模或联合传播机制融合不同模态信息，取得了一定效果。然而，这些方法通常采用固定权重或线性融合策略，缺乏对不同物品和推荐情境的动态适应能力；在模态质量不均或模态缺失的情况下，静态融合方式容易引入噪声，尤其在冷启动和长尾场

景中表现受限。

针对上述不足，本文提出一种新的多模态知识图谱推荐模型 MAF-KGAT (Modality-Adaptive Fusion with Knowledge Graph Attention Network)。该模型在 KGAT 框架中引入轻量级的模态自适应门控机制，在项目级别动态调节结构化知识表示与视觉表示的贡献。具体而言，本文利用预训练的 CLIP [9]图像编码器提取高层语义视觉特征，并通过自适应融合模块将其集成至知识图谱注意力传播过程中，从而在视觉信息有效时充分利用其优势，在噪声较大或缺失时自动依赖结构化知识表示，提升模型的鲁棒性。

在多个基准数据集上的实验结果表明，所提出的 MAF-KGAT 在多种评估设置下均优于现有先进方法，尤其在冷启动、模态缺失和长尾物品场景中表现出良好的稳定性和泛化能力。

本文的主要贡献包括：1) 提出了一种基于知识图谱的多模态推荐框架 MAF-KGAT，实现结构化知识与视觉信息的自适应融合；2) 设计了轻量级模态门控模块，将预训练 CLIP 视觉特征有效集成至 KGAT 传播过程；3) 通过广泛实验验证了所提方法在多种挑战性推荐场景中的有效性。

2. 相关工作

本章回顾与本文研究相关的主要技术方向，包括基于知识图谱的推荐方法、多模态特征与知识图谱的融合策略，以及对比学习与多模态预训练模型在推荐系统中的应用。通过分析代表性方法的研究思路与局限性，明确当前研究在模态融合灵活性、动态建模能力和复杂场景泛化性方面仍存在不足，为本文提出的 MAF-KGAT 模型奠定理论基础。

2.1. 基于知识图谱的推荐方法

知识图谱因其在实体建模和语义推理方面的优势，被广泛应用于推荐系统中，尤其在缓解数据稀疏性和冷启动问题方面表现突出。现有方法主要可分为嵌入融合方法和基于图神经网络的方法。

嵌入融合方法通过联合建模用户–物品交互数据与知识图谱中的实体关系信息，实现协同过滤与知识表示的协同优化。代表性工作如 CFKG [10]、MKR [11]和 DKN [12]分别通过联合训练、多任务学习或内容建模引入知识语义信息。然而，此类方法多采用线性拼接等静态融合策略，难以捕捉图结构中的高阶语义依赖关系，对图像、文本等非结构化模态的协同建模能力有限。

为增强对结构信息的建模能力，研究者进一步引入图神经网络(Graph Neural Networks, GNNs)进行知识图谱推荐。KGCN [3]、KGAT [2]和 KGIN [13]通过邻居聚合、注意力机制或语义路径建模，实现高阶语义传播；部分工作如 KGTN [14]、LGMRec [15]结合 Transformer 结构对图结构与多模态信息进行统一建模。尽管这些方法在结构建模方面取得了显著进展，但其大多侧重结构语义传播，对非结构化模态的建模能力有限，且缺乏对不同模态异构性与动态重要性的显式刻画。

2.2. 多模态特征与知识图谱的融合策略

多模态知识图谱通过引入图像、文本等非结构化模态，显著扩展了传统知识图谱的表达能力。MMKG [16]数据集的提出为多模态推荐研究提供了重要数据基础。

在模型设计方面，MMGCN [7]和 MKGAT [8]通过多通道建模或模态注意力机制融合结构与视觉信息，在一定程度上提升了推荐性能。然而，这类方法多采用节点级静态融合方式，难以根据不同物品或推荐上下文动态调整模态权重，容易引入信息冗余或模态冲突，在冷启动或模态缺失场景中表现受限。

为提升模态融合的鲁棒性，部分研究引入模态对齐与模态感知机制。MMKGR [17]利用对比学习增强跨模态一致性，VBKGC [18]基于预训练多模态模型联合建模图像与文本信息。然而，大多数方法仍未支持节点粒度的动态模态偏好建模，难以适应真实推荐场景中模态质量和重要性的变化。

2.3. 对比学习与多模态预训练方法

近年来，对比学习逐渐成为提升推荐表示鲁棒性的重要手段。SGL [19]、SimGCL [20]和 LightGCL [21]通过多视图或结构扰动策略增强模型稳定性。在多模态场景中，MR-MKG [22]通过跨模态对比学习提升图像与结构表示的一致性，并在多模态推荐任务中取得良好效果。

与此同时，CLIP [9]等视觉预训练模型被引入推荐系统，用于提取高质量图像语义表示。尽管其在视觉理解方面表现优异，但直接应用于推荐任务仍面临模型规模大、推理开销高以及视觉特征与结构嵌入语义不对齐等问题，若融合策略设计不当，可能导致表示退化。

综上，现有基于知识图谱的推荐方法在结构语义建模方面取得了显著进展，但在多模态融合策略上仍存在不足，主要体现在融合方式静态、模态异构性建模不足以及复杂场景下泛化能力有限。针对上述问题，本文提出的 MAF-KGAT 模型通过引入模态自适应门控机制，在节点层面动态调节结构与视觉模态的重要性，并结合 CLIP 视觉表示与图注意力传播，实现更灵活且鲁棒的多模态知识图谱推荐。

3. 引入模态自适应融合机制的多模态知识图谱推荐方法

本节介绍用于多模态推荐的 MAF-KGAT 模型，包括总体架构、模块设计及关键创新。与传统 KGAT [2]相比，MAF-KGAT 在多模态信息融合、基于注意力的图传播以及训练策略方面进行了系统扩展。

3.1. 模型整体框架概述

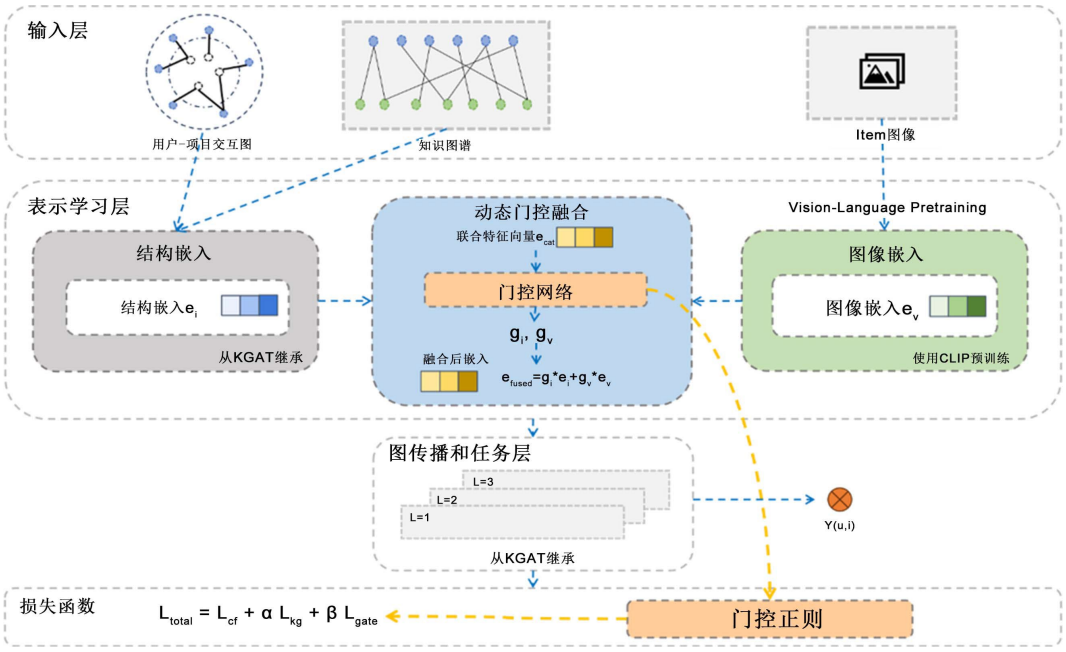

Figure 1. The overall framework of MAF-KGAT
图 1. MAF-KGAT 模型整体框架图

如图 1 所示，MAF-KGAT 由四个主要部分组成：输入层、表示学习层、图传播和任务层以及损失函数部分。下面将对各模块的设计与功能依次进行说明。

3.2. 输入层

输入层提供训练所需的多模态信息，包括：

用户－项目交互图：捕捉用户历史行为，提供协同过滤信号；

知识图谱：以实体－关系－实体三元组表示项目间高阶语义关联在数据稀疏时补充语义信息；

Item 图片：为每个物品提供视觉特征基础。

这些输入为表示学习模块提供基础，实现结构化与非结构化信息的统一建模。

3.3. 表示学习层

本节聚焦于 MAF-KGAT 的核心组件——模态自适应门控机制。该机制在图神经网络传播前动态判断每个实体应更多依赖结构信息还是视觉信息，从而提升多模态融合的灵活性与鲁棒性。

3.3.1. 结构嵌入处理

MAF-KGAT 基于协同知识图谱(CKG) [23]构建统一表示空间，整合用户－物品交互图与知识图谱信息，实现高阶关系建模。通过项目－实体对齐，将用户－项目交互与知识图谱三元组融合，捕捉高阶连接关系，为多模态融合提供结构基础。

具体地，推荐场景下的交互数据可建模为一个用户－物品双向图 $G_1=\{(u,y_{ui},i)|u\in U,i\in I\}$，其中 U 和 I 分别表示用户集与物品集，$y_{ui}=1$ 表示用户 u 与物品 i 之间存在交互行为(如点击、评分等)。

与此同时，知识图谱 G_2 以三元组 (h,r,t) 的形式组织辅助信息，描述物品及其属性之间的语义关系。例如，三元组(Hugh Jackman, ActorOf, Logan)表示演员与电影之间的关联。知识图中包含正向和反向关系(如 ActorOf 与 ActedBy)，实体集合记作 $\Im$，关系集合为 $\Re$。

为了将交互图与知识图整合，模型引入项目－实体对齐集 $A=\{(i,e)|i\in I,e\in\Im\}$，用于建立项目与知识图中实体之间的映射。进一步地，我们将每条用户－项目交互 (u,i) 转换为三元组形式 $(u,Interact,I)$，并将其与 KG 三元组融合，最终构建协同知识图谱：

$$G=\{(h,r,t)|h,t\in\Im'=\Im'\cup U,r\in\Re'=\Re\cup\{Interact\}\} \tag{1}$$

3.3.2. 图像嵌入处理

为了引入多模态视觉信息，本文使用预训练的 CLIP [9]模型(ViT-B/32 配置)提取每个项目图像的特征表示 e_v，采用的是 OpenAI 提供的 CLIP 模型中的 ViT-B/32 视觉编码器结构，其视觉输出维度为 $d_v=512$，即 clip.visual.output_dim = 512。由于图像特征与结构嵌入维度不一致，本文设计了一个图像特征投影模块将其映射到结构嵌入空间：

$$\hat{e}_v=\text{Dropout}\left(\text{LeakyReLU}\left(W\cdot e_v+b\right)\right) \tag{2}$$

其中，$\hat{e}_v$ 为经过投影后的图像嵌入，W 是一个大小为 $(d\times d_v)$ 的投影矩阵，d 表示结构嵌入的目标维度，即为 64，b 是一个大小为 d 的偏置项，与结构嵌入 e_i 处于同一低维空间。LeakyReLU 是非线性激活函数，Dropout 的比例为 mess_dropout[0]。该投影模块在训练过程中与主模型的其余参数一起更新，从而实现图像模态在低维空间中的对齐与压缩。

将结构嵌入和映射后的图像嵌入进行拼接，形成一个联合特征向量，作为后续门控的输入：

$$e_{cat}=[e_i;e_v]\in\mathbb{R}^{2d}$$

其中 $[\,;\,]$ 表示向量拼接操作。该拼接向量保留了它们之间的语义差异性。

3.3.3. 门控融合机制

我们设计了一个轻量级的门控网络，用于生成每个项目节点的门控向量 g_i，其核心结构为一个两层感知机(MLP)，包括如下步骤：

非线性映射与门控权重生成：

$$h=\mathrm{LeakyReLU}\left(W_1\times e_{cat}+b_1\right)$$

其中第一层线性变换的权重矩阵 $W_1=\mathbb{R}^{2d\times d}$，偏置项 $b_1=\mathbb{R}^d$。LeakyReLU 是激活函数，缓解特征稀疏性问题。

随后，通过第二层线性映射得到门控原始得分：

$$g_{logits}=W_2\times h+b_2$$

其中，$W_2=\mathbb{R}^{d\times 2}$ 表示第二层权重矩阵，$b_2=\mathbb{R}^d$ 是对应的偏置项。$g_{logits}=\mathbb{R}^2$ 包含结构模态和图像模态的未归一化得分。

Softmax 归一化：

为了将两个模态的得分转换为权重概率，引入 Softmax 函数进行归一化，并加入温度系数调控其分布平滑性：

$$g=\mathrm{softmax}\left(g_{logits}/T\right)$$

其中，温度参数 T(默认为 0.5)用于控制权重分布的平滑程度，我们在结构嵌入 e_i 与图像嵌入 e_v 上定义模态门控向量 g_i 和 g_v，$g=\left[g_i,g_v\right]$ 分别表示结构模态与图像模态的权重系数，满足 $g_i+g_v=1$。

最后，融合后的嵌入计算为：

$$e_{fused}=g_i*e_i+g_v*e_v$$

其中 e_{fused} 为融合之后的嵌入，这种融合形式可看作一维的残差控制机制，可证明具有凸组合性质，在优化中稳定、收敛性强。

3.4. 图传播和任务层

融合后的节点表示输入 KGAT 的知识感知注意力网络，递归进行高阶邻居聚合。

在第 l 跳传播中，对于当前实体节点 h，我们构建以 h 为头实体的 ego-network：

$$N_h=\left\{(h,r,t)\middle|(h,r,t)\in G\right\}$$

其中，r 代表关系类型，t 代表尾实体，G 是知识图谱中所有三元组的集合。

对邻居节点嵌入进行加权聚合，得到邻居传播信息：

$$e_{N_h}^{(l)}=\sum_{(h,r,t)\in N_h}\pi\left(h,r,t\right)\cdot e_t^{(l-1)}$$

其中 $e_{N_h}^{(l)}$ 代表第 l 层中，来自邻居节点对 h 的聚合信息；$e_t^{(l-1)}$ 表示节点 t 在第 $l-1$ 层的嵌入表示；$\pi\left(h,r,t\right)$ 为边 $\left(h,r,t\right)$ 的注意力权重，表示节点 t 对 h 的信息贡献度。

该注意力权重通过关系感知注意力机制计算：

$$\pi\left(h,r,t\right)=\frac{\exp\left(\left(W_re_t^{(l-1)}\right)^{\mathrm{T}}\tanh\left(W_re_t^{(l-1)}+e_r\right)\right)}{\sum_{(h,r',t')\in N_h}\exp\left(\left(W_{r'}e_{t'}^{(l-1)}\right)^{\mathrm{T}}\mathrm{anh}\left(W_{r'}e_h^{(l-1)}+e_{r'}\right)\right)}$$

其中，$e_h^{(l-1)}$ 为节点 h 在第 $l-1$ 层的嵌入表示；e_r 是关系 r 的嵌入向量；$W_r\in\mathbb{R}^{d'\times d}$ 对应关系 r 的线性变换矩阵，将节点嵌入从维度 d 映射到 d'；$\tanh(\bullet)$ 为双曲正切激活函数，增加非线性建模能力；$\exp(\bullet)$ 为指数函数，用于将打分映射为正数；分母为 softmax 归一化项，确保所有 $\pi\left(h,r,t\right)$ 加起来为 1。

随后，节点 h 的第 l 跳最终嵌入表示为：

$$e_h^{(l)} = f\left(e_h^{(l-1)}, e_{N_h}^{(l)}\right)$$

其中 $f(\cdot)$ 为聚合函数，用于综合当前节点的自身表示和邻居传播信息。我们采用 Bi-Interaction 聚合器建模两类交互：

$$f\left(e_h, e_{N_h}\right) = \text{LeakyReLU}\left(W_1\left(e_h + e_{N_h}\right) + W_2\left(e_h \odot e_{N_h}\right)\right)$$

其中，$\odot$ 表示元素级乘法；$W_1, W_2 \in \mathbb{R}^{d' \times d}$ 为加性项与乘积项分别设计的可学习变换矩阵；$\text{LeakyReLU}(\cdot)$ 为激活函数，用于增加非线性与保持负值信息；e_h 表示节点 h 在当前层的输入嵌入。

堆叠 L 层传播后，获得融合多模态语义与高阶结构特性的实体表示 $e_h^{(L)}$，用于推荐预测。

3.5. 损失函数

联合训练框架包含三部分损失。

3.5.1. 协同过滤损失(BPR)

基于排序优化用户－项目偏好。对于每个用户 u，样本项目 i 与负样本项目 j，损失函数为：

$$L_{cf} = \sum_{(u,i,j)\in O} -\log\sigma\left(\hat{y}_{ui} - \hat{y}_{uj}\right)$$

其中 $O = \left\{(u,i,j) \middle| (u,i) \in \Re^+, (u,j) \in \Re^-\right\}$ 表示训练集，$\Re^+$ 表示观察到的用户 u 与物品 i 之间的(正)交互，$\Re^-$ 是采样到的未观察到的(负)交互集；σ 为 sigmoid 函数，$\hat{y}_{ui} = e_u^{\mathrm{T}} e_i$ 表示预测评分，用户 u 对项目 i 的预测偏好，λ 为正则化系数。

3.5.2. 知识图谱损失(TransR)

提升关系建模区分性。对每个三元组 (h,r,t)，最小化正负样本的评分差异进行嵌入优化，损失函数定义如下：

$$L_{kg} = \sum_{(h,r,t,t')\in\Gamma} -\ln\sigma\left[f\left(h,r,t'\right) - f\left(h,r,t\right)\right]$$

其中 $\Gamma = \left\{(h,r,t,t') \middle| (h,r,t) \in G, (h,r,t') \notin G\right\}$，$(h,r,t')$ 是通过随机替换有效三元组中的一个实体而构建的破碎三元组；$f(h,r,t) = \left\|W_r e_h + e_r - W_r e_t\right\|_2^2$，$W_r \in \mathbb{R}^{k \times d}$ 是关系 r 的变换矩阵，将 d 维实体空间中的实体投射到 k 维关系空间中。$f(h,r,t)$ 的得分越低，表明三元组更有可能为真，反之亦然。

3.5.3. 门控正则损失

引入门控正则项约束门控值分布，提升融合稳定性。第 i 个项目的门控权重 $g_i \in [0,1]$，表示该项目偏向协同过滤(ID embedding)表示的程度。

$$L_{gate} = \frac{1}{|I|}\sum_{i\in I}\left(g_i - \mu\right)^2$$

其中 $\mu \in (0,1)$ 为先验的理想门控均值(通常设为 0.5)，I 为项目集合。

3.5.4. 总损失

交替优化 L_{kg} 和 L_{cf}，其中采用小批量 Adam 来优化嵌入损失和预测损失。最终训练目标是最小化包含三部分的联合损失函数：

$$L_{total} = L_{cf} + \alpha \cdot L_{kg} + \beta \cdot L_{gate}$$

其中，α 和 β 为可调权重超参数，分别控制知识图谱损失与门控正则项在总损失中的贡献。

3.6. 模型复杂度分析

MAF-KGAT 的主要计算开销来自：

模态门控融合模块：两层全连接，复杂度 $O(M \cdot D \cdot 2D) = O(M \cdot D^2)$，其中 M 表示项目数量，D 为嵌入维度。

图神经传播模块：基于注意力的消息传递，第 l 层的传播复杂度为：$O(|G| \cdot d_l \cdot d_{l-1})$

其中 $|G|$ 表示知识图谱中的边数，d_l 和 d_{l-1} 分别为当前层与上一层的变换维度。L 层传播的总复杂度为：

$$O\left(\sum_{l=1}^{L} |G| \cdot d_l \cdot d_{l-1}\right)$$

最后，预测阶段仅通过用户和项目嵌入的内积进行偏好估计，计算复杂度为 $O(|G| \cdot d_L)$，其中 d_L 是最终一层的嵌入维度。

综上，MAF-KGAT 的总训练时间复杂度可表示为：

$$O\left(M \cdot D^2 + \sum_{l=1}^{L} |G| \cdot d_l \cdot d_{l-1}\right)$$

4. 实验

4.1. 实验设置

为验证所提出模型的有效性与泛化能力，本文在两个广泛使用的推荐数据集上进行实验：Amazon Books 和 MovieLens-10M。

Amazon Books 来源于 Amazon Review 数据集的图书子集，包含用户对图书的评分记录。我们通过既有映射收集了对应的封面图像。在预处理阶段，剔除低分辨率的低质量图像；用零向量填充缺失图像，由预训练 CLIP 模型提取图像特征。最终，Amazon Books 数据集包含 70,679 名用户、24,915 个项目及 2,557,746 条知识图谱三元组。

MovieLens-10M 包含约 100 万条电影评分记录，电影海报通过 TMDb API 获取，使用 CLIP 提取视觉表示。其知识图谱来源于 Wang 等人[24]基于 Microsoft Satori 构建的子图，筛选置信度大于 0.9 的三元组，并保留与有效电影实体相关的部分，最终得到 145,406 条三元组。

对于两个数据集，均采用 80%/10%/10%的训练、验证与测试划分，并对每个正样本随机采样一个未交互项目作为负样本。数据集统计信息如表 1 所示，两个数据集均呈现高度稀疏性。

Table 1. Statistics of the datasets
表 1. 数据集的统计信息

Dataset		Amazon-book	MovieLens
User-Item Interaction	# Users	70,679	41,849
	# Items	24,915	4828
	# Interactions	847,733	1,813,382
Knowledge Graph	# Entities	88,572	65,801
	# Relations	39	19
	# Triplets	2,557,746	145,406

本文采用 Top-K 推荐中常用的 Recall@K 与 NDCG@K 评估模型性能。为了进行稳健的比较，本实验将每个方法运行 10 次后取平均值作为最终结果，K 设置为 20。

在两个数据集上采用统一参数设置。用户、实体及关系嵌入维度均设为 64；图神经网络采用 bi-interaction 聚合器，卷积层维度为[64, 32, 16]，邻接矩阵使用随机游走归一化以稳定消息传播。模型采用 Adam 优化器，学习率为 1e−3，L2 正则系数设为 1e−5，并在验证集指标连续 10 轮无提升时提前停止训练。

4.2. 不同方法的性能比较

本文将所提出的 MAF-KGAT 模型与多种代表性基线方法进行了比较，涵盖协同过滤、知识图谱增强以及多模态推荐等不同研究方向。具体包括：1) NFM [25]，一种结合因子分解与神经网络的协同过滤方法；2) CKE [26]，联合建模协同过滤信号与多模态知识嵌入的推荐模型；3) CFKG [10]，通过知识图谱路径建模用户与物品的多跳语义关系；4) RippleNet [27]，模拟用户兴趣在知识图谱中的逐层传播过程；5) KGAT [2]，基于图注意力机制的知识图谱推荐模型；6) MKGAT [7]，在 KGAT 框架中引入视觉特征进行多模态融合；7) MMGCN [8]，基于图卷积的多模态推荐方法。通过与上述方法的系统对比，验证了 MAF-KGAT 在多模态融合与推荐性能方面的优势。

表 2 汇总了各模型在 Amazon Books 和 MovieLens 上的性能结果。

Table 2. Performance comparison on two datasets
表 2. 两个数据集上的性能对比

Model	AmazonBooks		MovieLens	
	Recall@20	NDCG@20	Recall@20	NDCG@20
NFM	0.1323	0.0847	0.3520	0.4632
CKE	0.1311	0.0852	0.3585	0.4661
CFKG	0.1109	0.0755	0.3447	0.4520
RippleNet	0.1270	0.0883	0.3627	0.4718
KGAT	0.1443	0.0971	0.3750	0.4793
MMGCN	0.1500	0.1011	0.3900	0.4977
MKGAT	0.1539	0.1036	0.3978	0.5081
MAF-KGAT	0.1603	0.1072	0.4121	0.5230
%Improv.	**+4.2%**	**+3.5%**	**+3.6%**	**+2.9%**

可以看出，MAF-KGAT 在两个数据集的 Recall@20 与 NDCG@20 指标上均取得最优表现。在 Amazon Books 上，MAF-KGAT 相较 MKGAT 在 Recall@20 与 NDCG@20 上分别提升 4.2%与 3.5%；在 MovieLens 上分别提升 3.6%与 2.9%。结果表明，引入模态自适应门控机制后，模型能够更有效地融合视觉与结构信息。

性能提升主要源于两方面：一是模态自适应门控机制能够根据项目语义动态调整视觉与结构特征的重要性；二是 KGAT 中的图注意力机制增强了高阶语义传播能力。相较采用静态融合策略的 MKGAT，MAF-KGAT 在多模态信息利用上更具灵活性与鲁棒性。

4.3. 消融实验

为验证各融合策略的作用，在 Amazon Books 上进行了消融实验，结果如表 3 所示。

Table 3. Results of ablation experiments
表 3. 消融实验结果

模型变体	Amazonbook	
	Recall@20	**NDCG@20**
Base (原始模型)	0.1462	0.0968
base + image (简单拼接)	0.1504	0.1007
base + image (线性融合)	0.1556	0.1031
MAF-KGAT (完整模型 − 门控融合)	0.1603	0.1072

在不使用图像信息的基础模型(KGAT)上，推荐性能较为有限；引入图像特征后，无论是简单拼接还是线性融合，均能带来稳定提升，说明视觉模态具有有效补充作用。然而，采用模态自适应门控机制的完整模型取得了最优性能，Recall@20 和 NDCG@20 分别提升 9.6%和 10.7%。

该结果表明，动态、细粒度的模态权重分配机制相比静态融合策略，更有助于提升多模态表示的选择性与语义一致性。

4.4. 超参数敏感性分析

4.4.1. 嵌入维度分析

图 2 给出了不同嵌入维度下的性能表现。当维度为 64 时模型性能最佳；较小维度限制表达能力，而过大维度可能引入冗余并导致轻微性能退化。

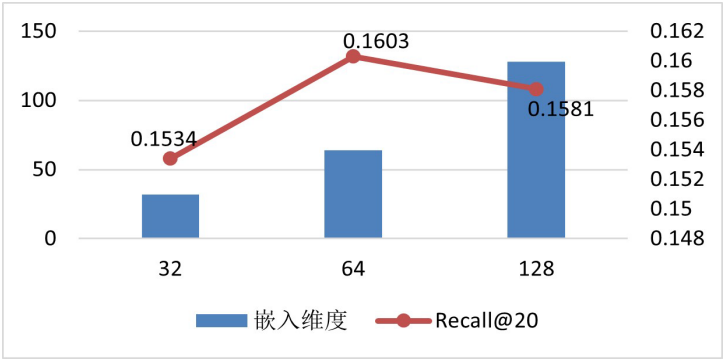

Figure 2. Effect of embedding dimension
图 2. 嵌入维度对算法的影响

4.4.2. 模态门控偏置项初始化敏感性分析

表 4 展示了不同门控偏置初始化对性能的影响，我们对门控网络中的偏置项进行初始化设置：

$$g = \{0.1, 0.3, 0.5, 0.7, 0.9\}$$

其中，较大的偏置值会使模型在训练初期更倾向于依赖结构信息，较小或 0 值则使模型在初始阶段对结构和视觉信息更加均衡。

Table 4. Effect of the initialization sensitivity of modal gating bias terms
表 4. 模态门控偏置项初始化敏感性实验结果

偏置项 g 初始化	Recall@20	NDCG@20
0.1	0.1567	0.1043
0.3	0.1556	0.1023
0.5	0.1590	0.1065
0.7	0.1564	0.1045
0.9	**0.1603**	**0.1072**

结果表明，在训练初期适当增强结构信息权重有助于模型学习稳定地表示，但整体性能随偏置变化较为平滑，说明 MAF-KGAT 对该超参数具有良好鲁棒性。

4.4.3. 模型深度的影响

表 5 给出了不同 GNN 层数下的性能比较。

Table 5. Effect of model depth
表 5. 模型深度影响实验结果

Layer	AmazonBooks		MovieLens	
	Recall	NDCG	Recall	NDCG
One	0.1560	0.1042	0.4052	0.5160
Two	0.1603	0.1072	0.4121	0.5230
Three	0.1585	0.1061	0.4094	0.5201

两层模型在两个数据集上均取得最优结果，而三层模型出现轻微退化，验证了过深模型可能带来的过平滑问题。

4.5. 多模态表示分析与可视化

为分析视觉特征与结构特征的表示差异，本文采用 t-SNE 对线性投影后的 CLIP 视觉嵌入与 KGAT 结构嵌入进行可视化。

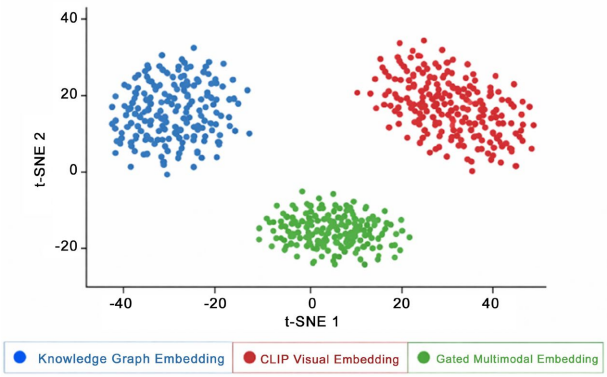

Figure 3. t-SNE visualization results
图 3. t-SNE 可视化结果

如图 3 所示，两种模态在低维空间中呈现出部分重叠但整体分布差异明显的特征，表明视觉信息与知识图谱结构信息在语义表达上具有互补性。该结果验证了本文在融合前通过向量拼接保留模态差异，并在高层引入模态自适应门控机制进行动态加权的设计合理性。

4.6. 冷启动场景实验分析

在 Amazon Books 上构建冷启动子集(用户交互次数 ＜5)，结果如表 6 所示。

Table 6. Cold start scenario experiment results
表 6. 冷启动场景实验结果

Model	Recall (@20)	NDCG (@20)
NFM	0.0924	0.0468
CKE	0.1013	0.0556
CFKG	0.1079	0.0591
RippleNet	0.1128	0.0619
KGAT	0.1164	0.0655
MMGCN	0.1182	0.0668
MKGAT	0.1241	0.0720
MAF-KGAT	**0.1256**	**0.0732**

相比仅依赖结构信息的模型，引入图像模态的 MKGAT 与 MAF-KGAT 均显著提升性能，其中 MAF-KGAT 在 Recall@20 与 NDCG@20 上取得最高结果，表明其在冷启动场景下具有更强的鲁棒性。

4.7. 图像缺失场景实验

为评估模型在模态不完整场景下的表现，我们随机移除不同比例的图像特征，结果如表 7 所示。

Table 7. Comparison of recommendation performance under different proportions of missing images
表 7. 不同图像缺失比例下推荐性能对比

图像缺失比例	MAF-KGAT	KGAT	相对提升
0%	0.1603	0.1452	+10.4%
20%	0.1589	0.1451	+9.5%
50%	0.1552	0.1452	+7.0%
80%	0.1501	0.1450	+3.6%

随着图像缺失比例增加，MAF-KGAT 性能逐渐下降，但在 80%图像缺失时仍显著优于 KGAT。该结果验证了模态门控机制能够在图像缺失情况下自适应增强结构信号，从而保证模型稳定性。

5. 总结

本文提出了一种模态自适应融合的知识图谱注意力多模态推荐网络 MAF-KGAT。该模型通过引入轻量级门控机制，在项目层面动态平衡视觉信息与结构信息，实现了在统一知识图谱框架下的灵活且鲁棒的表示学习。在多个基准数据集上的实验结果表明，MAF-KGAT 在各类评估指标下均优于现有最先进方

法。尤其是在冷启动、模态缺失及长尾物品场景中，所提出的模型表现出较强的鲁棒性，验证了模态自适应融合策略的有效性和实用性。

针对未来工作，本文计划进一步扩展该框架以纳入更多模态信息，如文本描述和用户生成内容。同时，将大型语言模型(LLMs)用于深度语义对齐，以及探索不确定性或上下文感知的融合机制，将有助于提升基于知识图谱的多模态推荐系统的适应性、泛化能力及可解释性。

参考文献

[1] Koren, Y., Bell, R. and Volinsky, C. (2009) Matrix Factorization Techniques for Recommender Systems. *Computer*, **42**, 30-37. https://doi.org/10.1109/mc.2009.263

[2] Wang, X., He, X., Cao, Y., Liu, M. and Chua, T. (2019) KGAT: Knowledge Graph Attention Network for Recommendation. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Anchorage, 4-8 August 2019, 950-958. https://doi.org/10.1145/3292500.3330989

[3] Wang, H., Zhao, M., Xie, X., Li, W. and Guo, M. (2019) Knowledge Graph Convolutional Networks for Recommender Systems. *The World Wide Web Conference*, San Francisco, 13-17 May 2019, 3307-3313. https://doi.org/10.1145/3308558.3313417

[4] Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., *et al.* (2020) Graph Neural Networks: A Review of Methods and Applications. *AI Open*, **1**, 57-81. https://doi.org/10.1016/j.aiopen.2021.01.001

[5] Hu, H., Guo, W., Liu, Y. and Kan, M. (2023) Adaptive Multi-Modalities Fusion in Sequential Recommendation Systems. *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, Birmingham, 21-25 October 2023, 843-853. https://doi.org/10.1145/3583780.3614775

[6] Yi, Z. and Ounis, I. (2024) A Unified Graph Transformer for Overcoming Isolations in Multi-Modal Recommendation. *18th ACM Conference on Recommender Systems*, Bari, 14-18 October 2024, 518-527. https://doi.org/10.1145/3640457.3688096

[7] 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 27th ACM International Conference on Multimedia*, Nice, 21-25 October 2019, 1437-1445. https://doi.org/10.1145/3343031.3351034

[8] Sun, R., Cao, X., Zhao, Y., Wan, J., Zhou, K., Zhang, F., *et al.* (2020) Multi-Modal Knowledge Graphs for Recommender Systems. *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, Lyon, 23-27 April 2018, 19-23 October 2020, 1405-1414. https://doi.org/10.1145/3340531.3411947

[9] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., *et al.* (2021) Learning Transferable Visual Models from Natural Language Supervision. *Proceedings of the 38th International Conference on Machine Learning*, **139**, 8748-8763. https://proceedings.mlr.press/v139/radford21a.html

[10] Ai, Q., Azizi, V., Chen, X. and Zhang, Y. (2018) Learning Heterogeneous Knowledge Base Embeddings for Explainable Recommendation. *Algorithms*, **11**, Article 137. https://doi.org/10.3390/a11090137

[11] Wang, H., Zhang, F., Zhao, M., Li, W., Xie, X. and Guo, M. (2019) Multi-Task Feature Learning for Knowledge Graph Enhanced Recommendation. *The World Wide Web Conference*, San Francisco, 13-17 May 2019, 2000-2010. https://doi.org/10.1145/3308558.3313411

[12] Wang, H., Zhang, F., Xie, X. and Guo, M. (2018) DKN: Deep Knowledge-Aware Network for News Recommendation. *Proceedings of the* 2018 *World Wide Web Conference on World Wide Web—WWW'*18, Lyon, 23-27 April 2018, 1835-1844. https://doi.org/10.1145/3178876.3186175

[13] Wang, X., Huang, T., Wang, D., Yuan, Y., Liu, Z., He, X., *et al.* (2021) Learning Intents behind Interactions with Knowledge Graph for Recommendation. *Proceedings of the Web Conference* 2021, Ljubljana, 19-23 April 2021, 878-887. https://doi.org/10.1145/3442381.3450133

[14] Chang, C., Zhou, J., Weng, Y., Zeng, X., Wu, Z., Wang, C., *et al.* (2023) KGTN: Knowledge Graph Transformer Network for Explainable Multi-Category Item Recommendation. *Knowledge-Based Systems*, **278**, Article ID: 110854. https://doi.org/10.1016/j.knosys.2023.110854

[15] Guo, Z., Li, J., Li, G., Wang, C., Shi, S. and Ruan, B. (2024) LGMRec: Local and Global Graph Learning for Multimodal Recommendation. *Proceedings of the AAAI Conference on Artificial Intelligence*, **38**, 8454-8462. https://doi.org/10.1609/aaai.v38i8.28688

[16] Liu, Y., Li, H., Garcia-Duran, A., Niepert, M., Onoro-Rubio, D. and Rosenblum, D.S. (2019) MMKG: Multi-Modal Knowledge Graphs. In: Hitzler, P., *et al.*, Eds., *The Semantic Web*, Springer, 459-474. https://doi.org/10.1007/978-3-030-21348-0_30

[17] Zheng, S., Wang, W., Qu, J., Yin, H., Chen, W. and Zhao, L. (2023) MMKGR: Multi-Hop Multi-Modal Knowledge Graph Reasoning. 2023 *IEEE* 39*th International Conference on Data Engineering* (*ICDE*), Anaheim, 3-7 April 2023, 96-109. https://doi.org/10.1109/icde55515.2023.00015

[18] Zhang, Y. and Zhang, W. (2022) Knowledge Graph Completion with Pre-Trained Multimodal Transformer and Twins Negative Sampling. arXiv: 2209.07084.

[19] Wu, J., Wang, X., Feng, F., He, X., Chen, L., Lian, J., *et al.* (2021) Self-Supervised Graph Learning for Recommendation. *Proceedings of the* 44*th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 11-15 July 2021, 726-735. https://doi.org/10.1145/3404835.3462862

[20] Liu, C., Yu, C., Gui, N., Yu, Z. and Deng, S. (2023) SimGCL: Graph Contrastive Learning by Finding Homophily in Heterophily. *Knowledge and Information Systems*, **66**, 2089-2114. https://doi.org/10.1007/s10115-023-02022-1

[21] Cai, X., Huang, C., Xia, L., *et al.* (2023) LightGCL: Simple Yet Effective Graph Contrastive Learning for Recommendation. arXiv: 2302.08191. https://openreview.net/forum?id=FKXVK9dyMM

[22] Lee, J., Wang, Y., Li, J. and Zhang, M. (2024) Multimodal Reasoning with Multimodal Knowledge Graph. *Proceedings of the* 62*nd Annual Meeting of the Association for Computational Linguistics* (*Volume* 1: *Long Papers*), Bangkok, 11-16 August 2024, 10767-10782. https://doi.org/10.18653/v1/2024.acl-long.579

[23] Tang, X., Sun, T., Zhu, R. and Wang, S. (2021) CKG: Dynamic Representation Based on Context and Knowledge Graph. 2020 25*th International Conference on Pattern Recognition* (*ICPR*), Milan, 10-15 January 2021, 2889-2895. https://doi.org/10.1109/icpr48806.2021.9412877

[24] Wang, H., Zhang, F., Zhang, M., Leskovec, J., Zhao, M., Li, W., *et al.* (2019) Knowledge-Aware Graph Neural Networks with Label Smoothness Regularization for Recommender Systems. *Proceedings of the* 25*th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Anchorage, 4-8 August 2019, 968-977. https://doi.org/10.1145/3292500.3330836

[25] He, X. and Chua, T. (2017) Neural Factorization Machines for Sparse Predictive Analytics. *Proceedings of the* 40*th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Shinjuku, 7-11 August 2017, 355-364. https://doi.org/10.1145/3077136.3080777

[26] Zhang, F., Yuan, N.J., Lian, D., Xie, X. and Ma, W. (2016) Collaborative Knowledge Base Embedding for Recommender Systems. *Proceedings of the* 22*nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, 13-17 August 2016, 353-362. https://doi.org/10.1145/2939672.2939673

[27] Wang, H., Zhang, F., Wang, J., Zhao, M., Li, W., Xie, X., *et al.* (2018) RippleNet: Propagating User Preferences on the Knowledge Graph for Recommender Systems. *Proceedings of the* 27*th ACM International Conference on Information and Knowledge Management*, Torino, 22-26 October 2018, 417-426. https://doi.org/10.1145/3269206.3271739