基于 MENTOR 模型的多模态推荐算法研究

万欣然 2410103017

摘要

随着多媒体信息的不断增加，多模态推荐逐渐引起广泛关注，它能缓解推荐系统中的数据稀疏问题，提升推荐准确性，然而这类方法对标记数据的依赖制约了模型的性能。尽管自监督学习已被用于多模态推荐以减轻标签稀疏问题，但现有的方法在对齐多模态信息时难以避免噪声的干扰。为此，MENTOR [17] 论文中提出了一种用于多模态推荐的多层次自监督学习方法，以解决标签稀疏和模态对齐问题。MENTOR 先利用图卷积网络（GCN）增强各模态的特定特征并融合视觉和文本模态，接着通过项目语义图增强模态的项目表示，然后引入多层次跨模态对齐任务和通用特征增强任务。本文对基于 MENTOR 模型的多模态推荐算法进行研究，并在此基础上进行了一些创新性调整和对比实验。实验结果表明，本文提出的加入多头自注意力机制进行模态融合的方法在部分数据集上的部分指标取得了一定的提升。

关键词： 推荐系统；多模态推荐；图神经网络

1 引言

随着互联网的快速发展，电子商务、社交媒体、视频平台等应用在全球范围内蓬勃发展，产生了大量的多模态数据。这些数据涵盖了文本、图像、音频等多种形式，反映了用户的行为、情感、偏好等信息。在这种信息爆炸的环境中，如何从海量的、多模态的数据中提取有价值的信息，并为用户提供个性化推荐，成为了人工智能领域的重要研究课题。推荐系统（Recommender Systems, RS）近年来在商业、金融、电商等领域的应用逐渐普及。然而，传统的推荐系统在面对数据稀疏性问题时，表现出了明显的局限性。为了克服这些挑战，研究者们开始探索基于多模态信息的推荐方法，即多模态推荐系统（Multimodal Recommendation Systems, MRS）。

单模态推荐系统通常仅依赖于用户与物品之间的交互信息或物品本身的描述进行推荐。这些方法往往忽视了信息的多维性和模态之间的互补性。例如，在电子商务平台中，商品的推荐不仅仅基于文字描述，还需要考虑商品图像、用户评价以及其他信息。因此，基于多模态数据的推荐系统能够综合多种数据源，全面挖掘用户的兴趣和需求，从而提供更加精准和个性化的推荐。近年来，研究人员不断探索如何利用多模态数据来提升推荐效果，但这一过程面临着许多挑战，包括模态间特征融合、数据稀疏性以及模型的复杂度等问题。

多模态推荐系统也面临着诸多问题。首先，不同模态的数据通常具有不同的结构和性质。例如，文本数据是非结构化的，而图像数据则需要通过深度学习模型提取出有效的特征。此外，模态之间的数据融合问题也至关重要。如何有效地对齐不同模态之间的特征，并将这些信息进行融合，从而提高推荐系统的性能，是当前多模态推荐系统研究的一个热点问题。

数据稀疏性是推荐系统的核心问题之一。数据稀疏性问题指的是用户与物品之间的交互数据非常稀疏，导致系统无法准确地捕捉到用户的兴趣和偏好。在传统的推荐系统中，尤其是在冷启动问题，例如新物品的推荐中，这一问题尤为突出。在多模态推荐系统中，虽然通过引入更多的模态数据能够缓解这一问题，但数据稀疏性依然是一个不可忽视的挑战。

此外，标签稀疏性问题也限制了多模态推荐系统的性能。推荐系统通常需要大量的标注数据，例如用户的评分、评论等来训练模型。但是，在实际应用中，标注数据往往是稀缺的，且获取这些标注数据的成本较高。为了克服这一问题，近年来，研究者们提出了一些无监督或自监督学习（Self-supervised Learning, SSL）方法，通过从未标注的数据中学习有效的特征表示，减少对标注数据的依赖。例如，在图像和文本的多模态推荐任务中，图像的自监督学习任务可以通过重建图像的部分信息来引导模型学习到图像的潜在特征，而文本的自监督学习任务则可以通过语言模型来学习文本的上下文信息。通过将自监督学习应用于不同模态，推荐系统能够更好地捕捉用户的兴趣和需求。然而，现有的自监督学习方法在多模态推荐系统中的应用仍然存在一些局限性。大多数现有的方法仅仅关注单一模态的特征学习，忽略了不同模态之间的相互关系。而在实际应用中，模态之间的相互作用和信息融合对于提高推荐效果至关重要。因此，如何有效地将自监督学习与多模态数据的特征融合结合起来，是当前研究的一个重要方向。

针对以上问题，《MENTOR: Multi-level Self-supervised Learning for Multimodal Recommendation》提出了一种新型的多层次自监督学习方法（MENTOR）[17]，旨在解决标签稀疏性和模态对齐问题。MENTOR 方法通过引入多层次的自监督学习任务，有效地提升了模型的鲁棒性和推荐准确性。

具体来说，MENTOR 方法首先利用图卷积网络（Graph Convolutional Network, GCN）增强每个模态的特征表示，并通过融合视觉和文本模态来构建融合模态。随后，MENTOR 通过构建项目语义图，进一步增强不同模态的项目表示。此外，MENTOR 引入了两种多层次的自监督任务：多层次跨模态对齐任务和通用特征增强任务。多层次跨模态对齐任务通过多级对齐策略有效地将不同模态的特征进行对齐，同时保持历史交互信息。而通用特征增强任务则通过图和特征的扰动来增强模型的鲁棒性，从而提高推荐系统在复杂场景中的表现。

通过这些创新，MENTOR 方法不仅有效地缓解了数据稀疏性和标签稀疏性问题，还通过多模态信息的深度融合，提高了推荐的精度和鲁棒性。大量的实验结果表明，MENTOR 方法在多个公开数据集上都取得了显著的性能提升，证明了其在多模态推荐领域的有效性。

总而言之，随着多模态数据的不断增加和推荐系统需求的日益复杂化，如何高效利用这些多模态信息并提升推荐系统的性能，已成为当前推荐系统研究的一个重要方向。本文对于 MENTOR 模型进行复现研究，并在此基础上进行思考，做出相关模块的创新实验调整以及对比试验。

2 相关工作

2.1 多模态推荐

随着多模态数据的快速增长，近年来多模态推荐系统逐渐成为提高推荐准确性和解决数据稀疏性问题的有效手段。多模态推荐系统通过结合来自不同模态例如文本、图像、音频等的信息，构建更加全面的用户兴趣和物品特征模型。多模态信息的引入能够有效缓解传统推

荐系统在稀疏数据下面临的困难，提高推荐的精准度和鲁棒性。

VBPR（Visual Bayesian Personalized Ranking）是最早采用视觉内容来缓解数据稀疏性问题的多模态推荐方法之一。VBPR 通过将视觉内容作为辅助手段，增强了物品表示，并基于矩阵分解技术提升了推荐性能 [2]。随后，更多的工作开始尝试融合视觉和文本模态，以进一步解决数据稀疏性问题。例如，ACF（Attention-based Collaborative Filtering）采用注意力机制，根据用户偏好自适应地加权各模态的特征，从而改进不同模态的融合质量 [1]。这些方法的目标是通过充分利用不同模态的数据来提高推荐质量。

为了提取更多的潜在信息并减小模态融合中的噪声，研究者们引入了图卷积网络（GCN）。MMGCN（Multimodal Graph Convolutional Network）通过构建用户-物品二分图，分别从每种模态中提取潜在信息，并聚合多模态预测结果以生成最终评分 [15]。GRCN（Graph Convolutional Recommendation Network）进一步通过剪除用户-物品图中的假阳性边，减少了噪声的干扰 [14]。除了 GCN 外，MKGAT（Multimodal Knowledge Graph Attention Network）通过图注意力机制来平衡不同模态在融合过程中的权重，从而提高融合效果 [10]。DualGNN 则通过构建一个额外的用户共现图，显式挖掘用户之间的共同偏好 [12]。

尽管这些方法在一定程度上提高了推荐精度，但仍存在一些问题。例如，现有的方法通常难以同时提取模态特定特征和模态共性特征，且在模态对齐和噪声控制方面仍有不足。因此，如何更好地提取模态特有的特征，同时有效融合各模态的共性信息，仍是当前多模态推荐系统中的一个关键挑战。

2.2 自监督学习在推荐中的应用

自监督学习（SSL）是一种无监督学习的技术，通过设计特定的任务或目标，使模型能够从未标注的数据中自我学习特征表示。在推荐系统中，传统的自监督学习方法主要用于提高模型的鲁棒性并缓解标签依赖问题。自监督学习的应用使得模型能够在缺少大量标注数据的情况下，仍然有效地进行训练。

在传统的推荐系统中，SelfCF（Self-supervised Collaborative Filtering）和 BUIR（Bipartite User-Item Representation）等方法利用自监督信号，通过生成不同的视图来学习用户和物品的表示 [23] [4]。此外，MixGCF（Mixing Graph Convolutional Filtering）通过设计负采样插件来探索负交互信息，进一步增强了基于图卷积网络的推荐系统的鲁棒性 [3]。

近年来，结合自监督学习和对比学习（Contrastive Learning, CL）的方法也在推荐系统中得到了广泛应用。例如，SSL4Rec 和 SEPT 提出了基于对比学习的推荐方法，通过对用户数据视图的增强，利用自监督信号从其他用户的行为中学习表示 [18] [20]。这些方法通过不断迭代地提升编码器，逐步提高模型的推荐性能。此外，MHCN（Multimodal Hypergraph Collaborative Network）和 SGL（Self-supervised Graph Learning）等方法也提出了基于图的增强自监督学习任务，通过构建语义邻居和随机游走图采样来生成自监督信号，从而提升模型的准确性和鲁棒性 [21] [16]。

在多模态推荐领域，自监督学习方法同样得到了广泛应用。SLMRec（Self-supervised Learning for Multimodal Recommendation）提出了两种自监督任务，通过特征扰动和多模态模式发现增强模型的鲁棒性 [11]。BM3（Bootstrap Multimodal Model）通过简化 SLMRec 中的自监督任务，提出了两种任务来对齐模态特征，分别从模态间和模态内的角度进行对齐 [24]。MMSSL（Multimodal Self-supervised Learning）则设计了一个跨模态对比学习任务，以保持

模态之间的语义共性，并增强用户偏好的多样性 [13]。

然而，这些方法在进行模态对齐时，往往会引入大量噪声，导致历史交互信息的丧失，进而影响推荐效果。为了克服这一问题，本文进行研究的 MENTOR 模型提出了一种新的多层次跨模态对齐任务，能够有效地对齐不同模态的特征，同时保留历史交互信息，从而增强模型的鲁棒性和推荐的精确性。

3 本文方法

3.1 方法概述

在本节中，将介绍复现的 MENTOR 模型的具体方法。MENTOR 模型的架构图如1所示：首先利用图卷积网络为每个模态提取特定的特征。其次融合视觉和文本模态，并基于 VT 融合、ID、视觉和文本四种模态表示，利用项目语义图探索潜在信息。我们利用对齐自监督任务（2）在不损失交互信息的情况下对齐每个模态。此外，我们利用自监督任务在特征掩码任务（1）和图扰动任务（3）中增强通用特征。

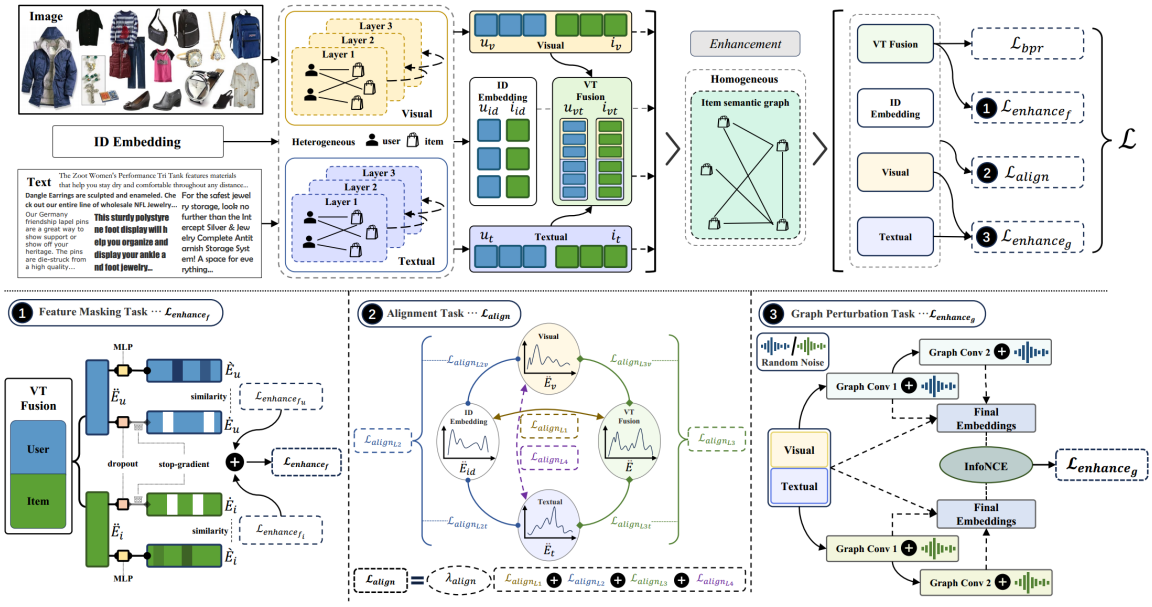

图 1. MENTOR 模型架构图

3.2 多层次跨模态对齐

MENTOR 模型中提出的多层次跨模态对齐任务是文章中最重要的任务之一。在本小节中将对多模态对齐任务进行详细介绍。不同模态的特征分布差异极大，这在模态融合过程中会产生大量噪声。然而，现有的模态对齐方法 [24] [13] 基本上会干扰历史交互信息。因此，MENTOR 模型提出了一个多层次跨模态对齐组件，利用自监督学习从数据分布的角度对齐模态。

具体而言，多层次跨模态对齐组件有四个层次，包括 ID 直接引导、ID 间接引导、模态直接对齐和模态间接对齐层次。ID 直接引导和 ID 间接引导层次充分利用 ID 模态中的历史

交互信息特征，以增强融合模态、视觉模态和文本模态中的历史交互特征。模态直接对齐和模态间接对齐层次使用两种不同层次的自监督信号来对齐视觉和文本模态。

3.2.1 ID 直接引导对齐

在 ID 直接引导层面，原文将融合模态 $\vec{E}$ 与 ID 模态 $\ddot{E}_{id}$ 进行对齐，以增强历史交互的重要性。受 PPMDR [5] 的启发，原文采用高斯分布来参数化 $\bar{E}$ 和 $\vec{E}_{id}$。然后，计算这两个分布之间的距离作为损失如下：

$$\ddot{E} \sim N(\mu_{vt}, \sigma_{vt}^2)\ \ddot{E}_{id} \sim N(\mu_{id}, \sigma_{id}^2)$$

$$\mathcal{L}_{align_{L1}} = |\mu_{id} - \mu_{vt}| + |\sigma_{id} - \sigma_{vt}|$$

其中 (μ_{id}, σ_{id}) 和 (μ_{vt}, σ_{vt}) 分别表征融合模态 $\ddot{E}$ 和 ID 模态 $\vec{E}_{id}$ 的高斯分布。

3.2.2 ID 间接引导对齐

在 ID 间接引导层面，原文分别将视觉模态 E_v 和文本模态 E_t 与 ID 模态 $\ddot{E}_{id}$ 进行对齐。与 ID 直接引导层面相同，最终损失通过以下方式计算：

$$\ddot{E}_v \sim N(\mu_v, \sigma_v^2)\ \ddot{E}_t \sim N(\mu_t, \sigma_t^2)$$

$$\mathcal{L}_{align_{L2v}} = |\mu_{id} - \mu_v| + |\sigma_{id} - \sigma_v|$$

$$\mathcal{L}_{align_{L2t}} = |\mu_{id} - \mu_t| + |\sigma_{id} - \sigma_t|$$

$$\mathcal{L}_{align_{L2}} = \mathcal{L}_{align_{L2v}} + \mathcal{L}_{align_{L2t}}$$

其中 (μ_v, σ_v) 和 (μ_t, σ_t) 分别表征视觉模态 $\vec{E}_v$ 和文本模态 $\vec{E}_t$ 的高斯分布。

3.2.3 模态直接对齐

在模态直接引导层面，原文分别将视觉模态 $\vec{E}_v$ 和文本模态 $\vec{E}_t$ 与融合模态 $\bar{E}$ 进行对齐。与上述两个层面不同，此层面旨在直接对齐模态分布，以减少融合模态中的模态噪声，公式如下：

$$\mathcal{L}_{align_{L3v}} = |\mu_{vt} - \mu_v| + |\sigma_{vt} - \sigma_v|$$

$$\mathcal{L}_{align_{L3t}} = |\mu_{vt} - \mu_t| + |\sigma_{vt} - \sigma_t|$$

$$\mathcal{L}_{align_{L3}} = \mathcal{L}_{align_{L3v}} + \mathcal{L}_{align_{L3t}}$$

3.2.4 模态间接对齐

在模态间接引导层面，原文中对视觉模态 $\ddot{E}_v$ 和文本模态 $\ddot{E}_t$ 进行对齐。其目的是间接对齐模态分布，对齐损失定义如下：

$$\mathcal{L}_{align_{L4}} = |\mu_v - \mu_t| + |\sigma_v - \sigma_t|$$

最终计算总体多层次跨模态对齐损失 L_{align} 如下，其中 λ_{align} 是平衡超参数。

$$\mathcal{L}_{align} = \lambda_{align}(\mathcal{L}_{align_{L1}} + \mathcal{L}_{align_{L2}} + \mathcal{L}_{align_{L3}} + \mathcal{L}_{align_{L4}})$$

3.3 通用特征增强

原 MENTOR 模型中提出的另一个重要任务是通用特征增强任务。在本小节中将对通用特征增强任务进行详细介绍。原文中提出了一个通用特征增强组件，基于表示生成多个视图，然后捕获这些视图之间的一致特征，以增强推荐的鲁棒性。这种自监督方法从原始数据中提取通用特征，从而缓解数据稀疏问题。通用特征增强组件可以分为两个任务：特征掩码和图扰动，分别如图1中的（1）和图1中的（3）所示。

3.3.1 特征掩码

原文中首先将 $\ddot{E}$ 拆分为 $\vec{E}_u$ 和 $\vec{E}_i$ 两部分。然后，对这些嵌入的一个子集进行掩码处理，以生成对比视图 $\dot{E}_u$ 和 $\dot{E}_i$。

$$\dot{E}_u = \ddot{E}_u \cdot Bernoulli(p)$$

$$\dot{E}_i = \ddot{E}_i \cdot Bernoulli(p)$$

受 BM3 [24] 的启发，原文中借助丢弃（dropout）机制 [7]，对表示 $\vec{E}_u$ 和 $\vec{E}_i$ 的一个子集进行随机掩码处理，其中 P 为丢弃率。

受 [24] 启发，原文对对比视图 $\dot{E}_i$ 和 $\dot{E}_u$ 应用停止梯度操作。然后通过多层感知机（MLP）对 $\vec{E}_u$ 和 $\vec{E}_i$ 进行变换。

$$\dot{E}_u = \ddot{E}_u W + b$$

$$\grave{E}_i = \grave{E}_i W + b$$

其中 $W \in \mathbb{R}^{d \times d}$、$b \in \mathbb{R}^d$ 分别表示线性变换矩阵和偏置。

最后，特征掩码损失定义如下，其中，Sim 是余弦相似度。

$$\mathcal{L}_{enhance_{fu}} = 1 - Sim(\dot{E}_u, \dot{E}_u)$$

$$\mathcal{L}_{enhance_{fi}} = 1 - Sim(\dot{E}_i, \grave{E}_i)$$

$$\mathcal{L}_{enhancef} = \mathcal{L}_{enhance_{fu}} + \mathcal{L}_{enhance_{fi}}$$

3.3.2 图扰动

原文中遵循图中基于丢弃（dropout）机制的最常用的增强方法 [16] [19]，为视觉模态和文本模态构建结构扰动的对比视图。然后提出了一种扰动后的用户-项目图：

$$\dagger E_m^{(l)} = \sum_{i \in \mathcal{N}_u} \frac{1}{\sqrt{|\mathcal{N}_u|}\sqrt{|\mathcal{N}_i|}} \dagger E_m^{(l-1)} + \Delta^{(l)}$$

并且 $\Delta^{(l)} \in \mathbb{R}^{d_m} \sim U(0,1)$ 是一个随机噪声向量。每个模态最终的扰动嵌入通过逐元素求和来计算。形式上：

$$\dagger \overline{E}_m = \sum_{l=0}^{L} \dagger E_m^{(l)}$$

对于视觉模态和文本模态而言，原文中为每个模态生成两个对比视图 $\dagger\bar{E}_m^1$ 和 $\dagger\bar{E}_m^2$，并采用 InfoNCE [8] 进行对比学习。形式上，每个模态的图扰动损失定义如下：

$$\mathcal{L}_{enhance} = \sum_{u \in \mathcal{U}} -\log \frac{\exp\left(e_{u,m}^1 \cdot e_{u,m}^2/\tau\right)}{\sum_{v \in \mathcal{U}} \exp\left(e_{v,m}^1 \cdot e_{v,m}^2/\tau\right)} + \sum_{i \in I} -\log \frac{\exp\left(e_{i,m}^1 \cdot e_{i,m}^2/\tau\right)}{\sum_{j \in I} \exp\left(e_{j,m}^1 \cdot e_{j,m}^2/\tau\right)}$$

其中，$e_{u/v,m}^1$ 和 $e_{u/v,m}^2$ 是对比视图 $\dagger\bar{E}_m^1$ 和 $\dagger\bar{E}_m^2$ 中用户 u/v 的模态 m 特征。此外，$e_{i/j,m}^1$ 和 $e_{i/j,m}^2$ 是对比视图 $\dagger\bar{E}_m^1$ 和 $\dagger\bar{E}_m^2$ 中项目 i/j 的模态 m 特征。τ 是 softmax 的温度超参数。

总的图扰动损失计算如下，这个自监督任务旨在提取用户-项目交互的通用结构意义。

$$\mathcal{L}_{enhance} = \mathcal{L}_{enhance_w} + \mathcal{L}_{enhance_t}$$

最后，总体的通用特征增强损失如下，其中 λ_g 和 λ_f 是平衡超参数。

$$\mathcal{L}_{enhance} = \lambda_g \mathcal{L}_{enhance} + \lambda_f \mathcal{L}_{enhance_f}$$

3.4 优化函数定义

原文中采用贝叶斯个性化排序（BPR）损失 [9] 作为基本优化函数。贝叶斯个性化排序旨在扩大训练集 D 中每个三元组 (u, p, n)（其中 D 表示训练集）里正、负项目之间预测偏好的差值。正项目 p 指的是用户 u 已经与之产生交互的项目，而负项目 n 是从用户 u 未与之交互的项目集合中随机选取的。贝叶斯个性化排序函数定义如下：

$$\mathcal{L}_{bpr} = \sum_{(u,p,n) \in \mathcal{D}} -\log(\sigma(y_{u,p} - y_{u,n}))$$

其中，$y_{u,p}$ 和 $y_{u,n}$ 分别是用户 u 对正项目 p 和负项目 n 的评分，它们分别通过 $\vec{E}_u^T \ddot{E}_p$ 以及 $\vec{E}_u^T \vec{E}_n$ 来计算。σ 为 Sigmoid 函数。通过结合贝叶斯个性化排序（BPR）损失、多层次跨模态对齐损失以及通用特征增强损失来更新用户和项目的表示。表示为：

$$\mathcal{L} = \mathcal{L}_{bpr} + \mathcal{L}_{align} + \mathcal{L}_{enhance} + \lambda_E(\|E_v\|_2^2 + \|E_t\|_2^2)$$

其中，E_v 和 E_t 是模型参数。λ_E 是一个超参数，用于控制 L_2 正则化的影响。

4 复现细节

4.1 与已有开源代码对比

本文在复现的过程中所参考的代码为原文提供的代码，代码地址在原文中给出。在对数据集进行预处理时使用 MMRec [22] 中的生成 u-u-matirx 的代码。

表 1. 实验数据集的统计信息

Dataset	Users	Items	Interaction	Sparsity
Baby	19445	7050	160792	99.88%
Sports	35598	18357	296337	99.95%
Clothing	39387	23033	278677	99.97%

表 2. 实验环境配置

实验环境	**具体配置**
硬件环境	
CPU	Intel Core i7-12700
GPU	NVIDIA A100-PCIE-40GB
软件环境	
深度学习框架	Pytorch 2.1.0
CUDA	11.8

4.2 实验数据集、实验环境、评价指标介绍

1. 数据集介绍。

本文使用的是公开数据集 Amazon [6] 数据集中的三大类别进行实验，分别是 Baby、Sports、Clothing 数据集。数据集的介绍如表1所示。

2. 实验环境介绍。本文复现的实验环境如表2所示。

3. 评价指标介绍。

本文采用了两种广泛使用的指标，分别是 Recall@K（R@K）和 NDCG@K（N@K）。最后给出测试数据集中所有用户在 $K = 10$ 和 $K = 20$ 两种情况下的平均指标。本文遵循流行的评估设置，将数据随机划分为 8:1:1，分别用于训练、验证和测试。

4.3 实验复现

本文首先对原文工作进行复现，在 Baby、Sports、Clothing 三个数据上进行实验。其次进行两个消融实验，探究不同模块的影响，分别是（1）多层次跨模态对齐消融实验和（2）通用特征增强消融实验，具体实验内容和结果见第 5 章节。

4.4 创新性探究

4.4.1 参数调整

在原 MENTOR 模型中，模型使用了图卷积网络为每个模态提取特定的特征，将视觉和文本模态融合，并且基于融合的 VT 模态、ID、视觉和文本四种模态表示，利用项目语义图寻找潜在的信息和联系。在对于视觉和文本模态的邻接语义图进行融合时，对于视觉和文本两种模态的权重参数设定初始值为 image_weight=0.1，本文对此参数设定进行分析和实验。

在多模态推荐系统中，不同模态的数据具有不同的特征和重要性，因此在进行模态融合时，如何合理设定每个模态的权重至关重要。在 MENTOR 模型中，初始时将视觉模态的权重设为较低的值，这是基于对不同模态在推荐任务中的作用差异的考虑。

视觉模态与文本模态相比，通常具有更高的维度和复杂性，且其特征提取通常需要更高的计算资源和较长的训练时间。因此，过高的视觉模态权重可能导致模型在训练过程中对视觉特征过度依赖，从而影响模型的训练效率和效果。另外，视觉和文本模态在实际推荐任务中的贡献可能有所不同，在某些特定领域中，文本信息的作用可能更为重要，而在一些领域中，视觉信息的作用会更加重要。

因此，对视觉模态参数的调整，尤其是权重参数的设定，可以进一步优化多模态数据融合的效果，确保视觉和文本信息的协同作用得到最大化。通过实验验证不同权重设定的影响，能够帮助我们更精确地理解视觉模态在具体推荐场景中的作用，并对模型进行更有效的调优。

4.4.2 使用多头自注意力机制的模态融合

在原 MENTOR 模型中，四种模态的融合和对齐是核心设计之一，其中最为重要的便是视觉和文本模态的信息融合。分析源代码可以发现，原始实现中采用了图卷积网络（GCN）来分别提取视觉和文本模态的特征，并通过将这两个特征信息的邻接语义图采用直接加权融合的方式进行融合。该方法的优势在于其高效性和简洁性，通过直接融合的方式，可以迅速将来自不同模态的信息整合在一起，减少了计算和模型复杂度，从而提高了模型的运行效率。

然而，尽管这种方式在一些场景下效果良好，它在面对更加复杂的多模态数据时，仍存在一定的局限性。由于视觉和文本模态的数据特性差异较大，它们之间的关系往往是非线性且难以直接捕捉的。这种方式可能无法充分利用这些模态间的潜在关联，导致信息融合不够细致和全面，特别是在模态间信息不对称的情况下，模型可能无法有效地提取到所有有价值的特征。此时，简单的加权融合操作可能会忽视视觉和文本之间深层次的交互和关联，从而影响推荐系统的精度和鲁棒性。

为了进一步提高融合效果，本文在原有设计的基础上，提出了改进——引入多头自注意力机制来捕捉视觉和文本模态之间的信息交互。多头自注意力机制能够根据不同模态的相关性动态调整其在融合过程中的权重，从而更精细地选择对推荐任务最有帮助的特征。具体来说，多头自注意力机制会根据视觉和文本模态在当前任务中的重要性，自动为它们分配不同的权重，以便在融合时强调模态间重要信息的交互，而弱化不相关或冗余的信息。这种方式能够更加灵活地适应不同任务中的需求，并有效地提升模型对复杂多模态数据的处理能力。

通过引入多头自注意力机制，我们不仅可以加强视觉和文本信息之间的协同作用，还能够更好地捕捉到模态间的细节特征，从而优化特征融合的质量。

5 实验结果分析

本节将对实验具体内容进行介绍和实验结果进行探究分析，分别包含了复现实验结果的对比分析、消融实验结果分析和创新探究实验对比分析。

5.1 复现实验结果分析

本文对于原文实验的复现结果如表3所示，其中包含了三个数据集的实验结果，其中本文复现的结果用 MENOTR(OURS) 表示，原文复现结果用 MENTOR 表示。在后文中进行消融实验对比和创新探究实验对比时，将原文复现的实验结果 MENOTR(OURS) 作为 Baseline 模型的结果进行对比。

对于实验结果进行分析可知，首先对于三个数据集上的表现，复现的结果与原文的结果存在微小差距，这种数值上的差距带来的影响可以忽略不计。每个数据集的 Recall 指标都存在着一定的差距，但是在 NDCG 指标上几乎无差别。甚至在 Sports 数据集上本文复现的结果比原论文的部分指标的结果更佳，R@10、N@10、N@20 分别提升了 0.5%、1.7%、0.1%。Clothing 数据集上本文复现的结果比原论文的部分指标的结果更佳，N@10 和 N@20 分别提升了 0.8%、0.9%。这种数值上的浮动会与实际硬件实验环境的差距有一定的联系。

表 3. 复现实验结果

DataSet	**Baby**				**Sports**				**Clothing**			
Model Source	R@10	R@20	N@10	N@20	R@10	R@20	N@10	N@20	R@10	R@20	N@10	N@20
MENTOR	0.6078	0.1048	0.0362	0.0450	0.0763	0.1139	0.0409	0.0511	0.0668	0.0989	0.0360	0.0441
Baseline	0.0652	0.1030	0.0350	0.0447	**0.0767**	0.1139	**0.0416**	**0.0512**	0.0663	0.0987	**0.0363**	**0.0445**

5.2 消融实验结果分析

5.2.1 多层次跨模态对齐模块消融实验

由原文可知，在多层次跨模态对齐部分一共有四个关键的对齐部分，本小节分别对其中的模块进行去除消融实验，共进行了以下的模型对比，其中 Baseline 模型指的是就是本文复现的完整的 MENTOR(OURS) 模型。

（1）MENTOR_M1，直接将所有多模态对齐去除。

（2）MENTOR_M2，保留多模态对齐的 id 直接引导。

（3）MENTOR_M3，保留多模态对齐的 id 直接、id 间接引导。

（4）MENTOR_M4，保留多模态对齐的 id 直接、id 间接引导、模态直接对齐。

消融实验的结果如表4所示，其中我们对 R@20 指标的结果进行对比展现，可视化结果如图2所示。

由实验结果可知，总体来看，依次增加模态对齐的模块会使得各指标数值有所提升，尽管其中会有些数值上的微小浮动，但初步表明多层次跨模态对齐组件的每一层都会使推荐性能得到提升，而且它们的效果可以相互叠加。

但是仔细分析可知，在不同的数据集上各层对齐组件带来的效果是不一致的，例如在 Sports 数据集上的结果表明，除了将多层次跨模态对齐组件全部删除外，其他的每一层的叠加带来的影响几乎可以忽略不计，这种结果与原文中展现的趋势存在一些不同之处，原因可能是对于不同的数据集而言，所需要的对齐模块会存在差异性，更需要探究的是各种模态融合带来的影响。

表 4. 消融实验 1 模型对比实验结果

DataSet	**Baby**	**Sports**	**Clothing**
Model Source	R@20	R@20	R@20
MENTOR_M1	0.1021	0.1133	0.0976
MENTOR_M2	0.1018	0.1137	0.0975
MENTOR_M3	0.1027	0.1137	0.0960
MENTOR_M4	0.1029	0.1138	0.0975
Baseline	**0.1030**	**0.1139**	**0.0987**

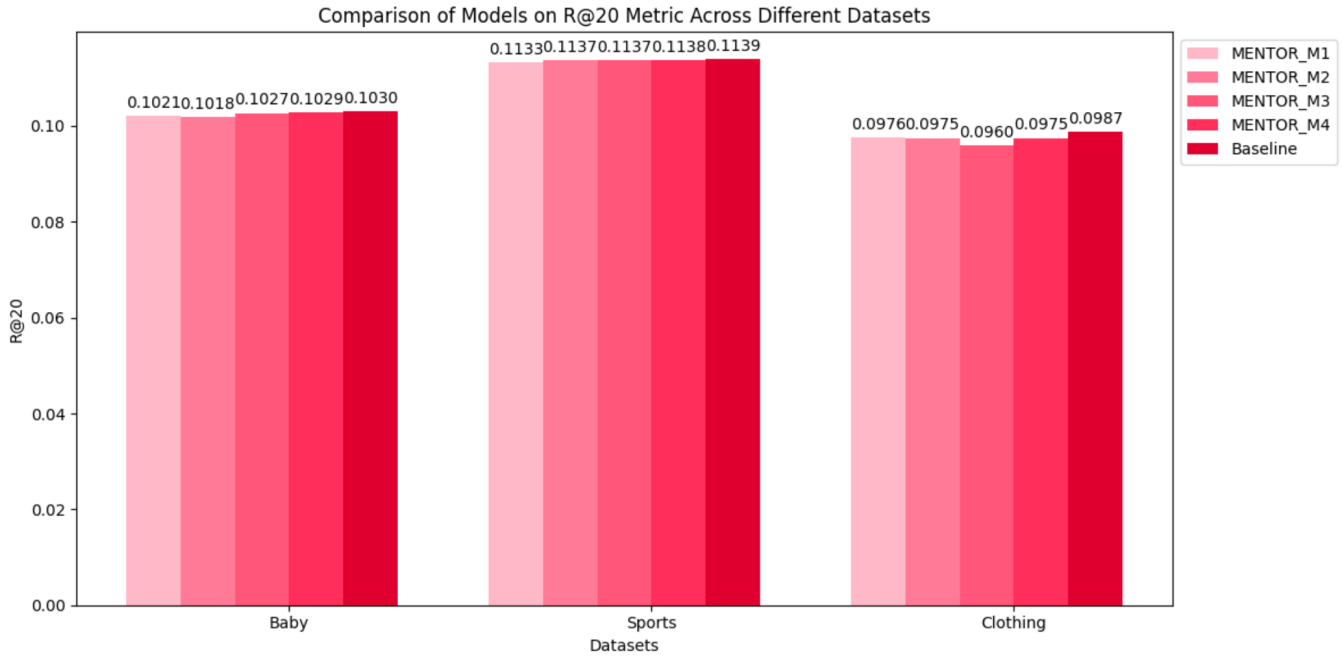

图 2. 消融实验 1 结果对比图

5.2.2 通用特征增强模块消融实验

由原文可知，在通用特征增强模块使用到的两个关键模块是特征掩码（Feature Masking）和图扰动（Graph perturbation）。本小节分别对这两个模块进行去除消融实验，共进行了以下的实验对比。

(1) MENTOR_G1，将所有特征增强模块去除。

(2) MENTOR_G2，将特征掩码的模块去除。

(3) MENTOR_G3，将图扰动的模块去除。

消融实验的结果如表5所示，其中我们对 R@20 指标的结果进行对比展现，可视化结果如图3所示。

由实验结果可知，当我们依次将特征增强模块去除进行对比时，发现实验的结果与原文存在一些不同之处，叠加通用特征增强模块在每个数据集上的表现并不一致。在 Baby 数据集上，我们可以发现，将图扰动的增强模块去掉后，反而取得了更好的效果。在 Sports 数据集上，将所有的特征增强模块去除后，取得的效果其实是最好的，从 R@20 这一指标的数值

上对比，消融模型的结果最后比 Baseline 模型还提升了 0.7%。据结果而言，我们还需要对特征通用模块的作用进行更详细的探究。

表 5. 消融实验 2 模型对比实验结果

DataSet	**Baby**	**Sports**	**Clothing**
Model Source	R@20	R@20	R@20
MENTOR_G1	0.1024	**0.1148**	0.0952
MENTOR_G2	0.1026	0.1136	0.0981
MENTOR_G3	**0.1035**	0.1142	0.0975
Baseline	0.1030	0.1139	**0.0987**

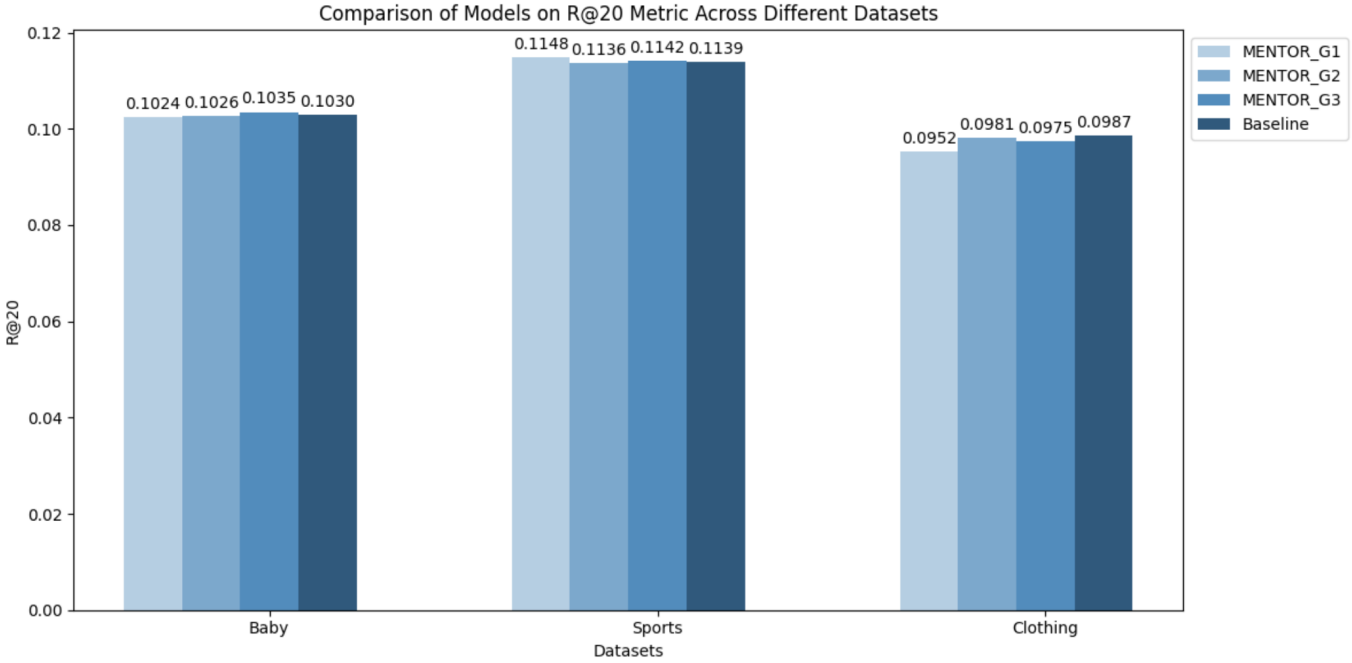

图 3. 消融实验 2 结果对比图

5.3 创新探究实验对比结果分析

5.3.1 参数调整实验

在第 4 章节中，本文介绍了基于 MENTOR 模型的创新性探究实验，本小节将依据其分析进行实验，并给出实验结果。首先在对于各类超参数的调整设置方面，本文对原文中的超参数设置进行分析，最后对以下的设置进行了保留：reg_wegiht=0.001,learning_wegiht=0.0001, dropout=0.5,mask_wegiht_f=1.5, mask_weight_g=[0.0001,0.001],align_weight=0.1, temp=[0.2,0.4,0.6]。

对于第四章节中提到的视觉和文本融合中 image_weight 的设置，再进行了相关实验后，选择了权重设置为 0.2，其中在 Baby 数据集上 R@20 上和 N@20 的指标分别提高了 1.1%、0.6%。与 Baseline 对比效果如图4所示，其中 MENTOR(I1) 指的是调整后的模型。

通过对实验结果的分析，我们发现，在调整视觉和文本模态邻接矩阵融合的权重参数后，模型的表现出现了部分提升。我们初步推测，视觉模态在 Baby 数据集中的作用可能比我们

最初预期的更加重要。这一发现表明，在不同模态信息的融合过程中，灵活调整各模态的融合权重是非常关键的。

因此，针对不同数据集和任务需求，合理调整各模态的融合权重显得尤为重要。单一的固定融合权重可能无法适应所有场景的需求，而根据实验结果动态调整模态权重，能够确保每种模态在特定任务中的优势得到充分发挥。

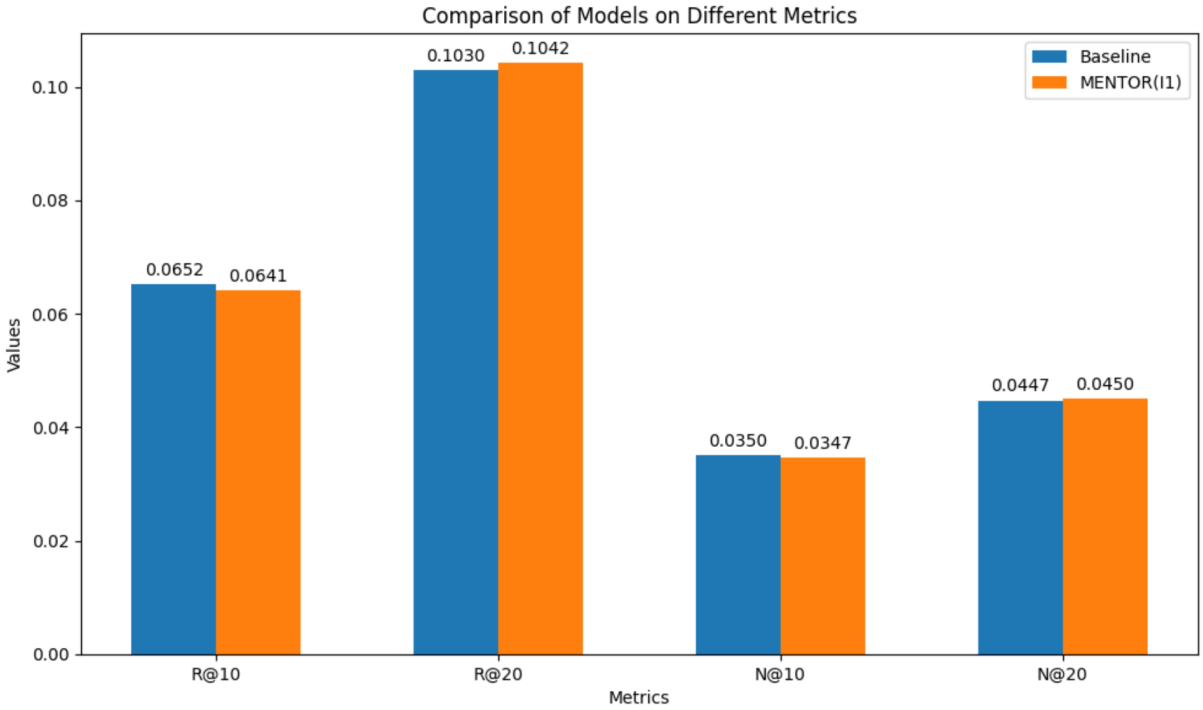

图 4. 参数调整实验对比图

5.3.2 使用多头自注意力机制的模态融合实验

在第 4 章节中，本文提出了可以使用多头自注意力机制进行模态的融合实验。分析原文的源代码可以发现，原始实现中采用了图卷积网络（GCN）来分别提取视觉和文本模态的特征，并通过将这些特征的邻接语义图进行加权的方式进行融合，这种方式虽然高效但是简单，可能会难以捕捉到各模态特征的重要信息。本文在此基础上进行探究实验，提出使用多头自注意力机制对于视觉和文本模态进行融合。

实验结果如表6所示，可视化结果如图5所示，展现的是在 Baby 数据集上的指标对比结果，其中 MENTOR(I2) 模型指的是调整后的模型。

分析实验结果可知，在 Baby 数据集上，使用了多头自注意力机制的 MENTOR(I2) 模型在 R@10 和 N@10 指标上分别取得了 1.2% 和 0.2% 的提升，但是在 R@20 和 N@20 两个指标上都存在下降的现象。这表明，在较小的推荐列表中，应用多头自注意力机制有助于更精确地捕捉用户与物品之间的关系，从而提高推荐的相关性和准确性。自注意力机制通过动态地调整不同特征之间的权重，有助于模型更好地理解和融合视觉与文本等多模态信息，进而提升了排名前 10 的推荐项的表现。

然而，在 R@20 和 N@20 这两个指标上，我们注意到模型的表现出现了下降。这一现象可能反映出多头自注意力机制在处理较长推荐列表时的局限性。在较短的推荐列表中，模型能够充分利用自注意力机制来精细调整不同模态之间的权重和关系，从而在前 10 项的推荐中

获得较大的提升。然而，随着推荐列表的扩展，模型可能会对较长列表中的部分较低排名项的处理不足，导致推荐准确度的下降。

总体而言，这一实验结果表明，虽然多头自注意力机制在提高推荐系统精度方面具有一定的优势，特别是在短列表推荐中，但其在处理长列表时仍然面临一定的挑战。

表 6. 加入多头自注意力机制的实验对比结果

Model Source	R@10	R@20	N@10	N@20
Baseline	0.0652	0.1030	0.0350	0.0447
MENTOR(I2)	**0.0660**	0.1005	**0.0351**	0.0440

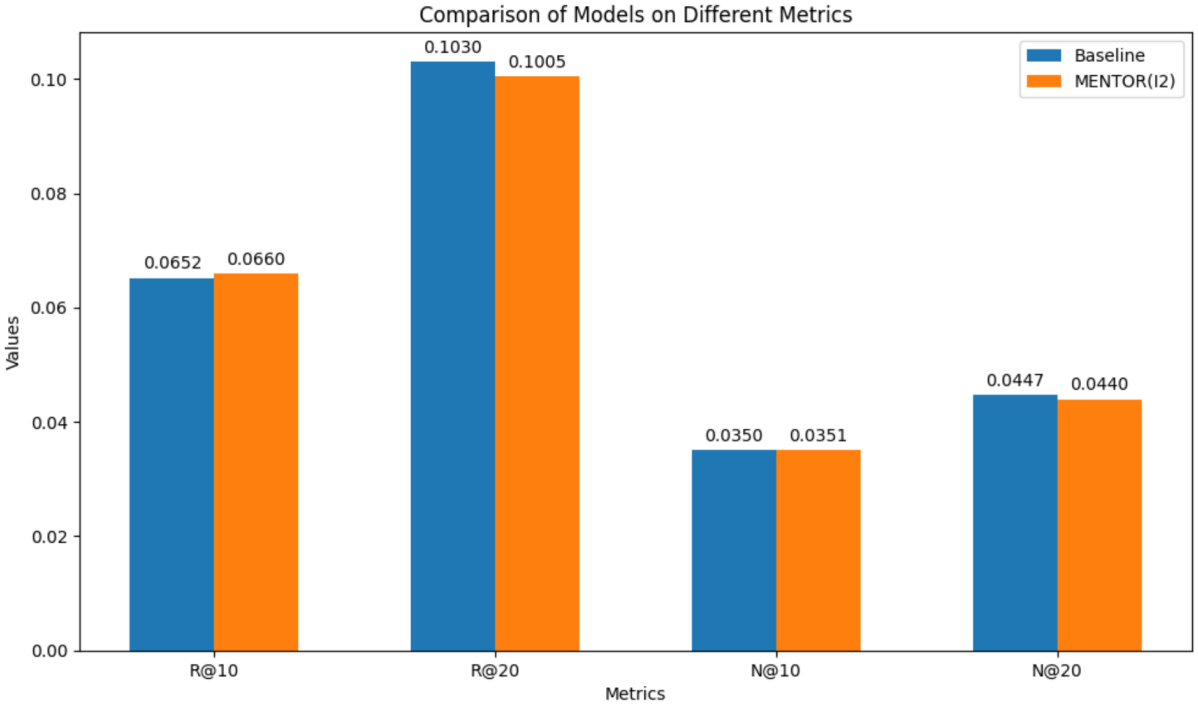

图 5. 加入多头自注意力机制的实验对比图

6 总结与展望

本文基于 MENTOR 模型进行了多模态推荐算法研究，对 MENTOR 这一方法进行复现，MENTOR 模型在多模态推荐领域提出了一种多层次的自监督学习方法，利用了视觉、文本、VT 融合、ID 这四种模态实现推荐任务，其中引入了两个关键模块，分别是多层次跨模态对齐和通用特征增强模块，这两个模块的提出，给模型带了优秀的性能提升表现。

本文对此模型进行复现和多项消融实验的对比，验证了模型关键模块的作用，但是对于部分数据集上的表现仍需要进一步分析，比如在 Sports 数据集上的实验结果表明，去掉原文提出的通用特征增强模块反而取得了超过原文结果的表现，这一结果值得后续进一步思考和实验分析。

其次，在原文的基础上，本文对视觉和文本模态的邻接矩阵的融合权重进行实验调整，在 Baby 数据集上进行实验。另外，本文对模态融合的方式进行思考调整，提出使用多头自注意

力机制进行融合的动态调整，实验结果表明在 Baby 数据集上的 R@10 和 N@10 指标上取得了提升，但是在面对前 20 个推荐列表时，指标却出现了下降的情况。

在未来，处理长推荐列表时，如何充分利用全局信息是提升推荐精度的关键。可以探讨如何改进自注意力机制，加入全局信息加权策略，探究通过更精细的全局特征融合来提升推荐系统对更大规模候选集的处理能力。同时分析原论文模型可知，引入多种不同的模态也许会给模型带来噪声，所以在未来的工作中，可以关注如何减少自监督任务中由不同模态信息所引发的噪声。

参考文献

[1] Jingyuan Chen, Hanwang Zhang, Xiangnan He, Liqiang Nie, Wei Liu, and Tat-Seng Chua. Attentive collaborative filtering: Multimedia recommendation with item-and component-level attention. In *Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 335–344, 2017.

[2] Ruining He and Julian McAuley. Vbpr: visual bayesian personalized ranking from implicit feedback. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

[3] Tinglin Huang, Yuxiao Dong, Ming Ding, Zhen Yang, Wenzheng Feng, Xinyu Wang, and Jie Tang. Mixgcf: An improved training method for graph neural network-based recommender systems. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 665–674, 2021.

[4] Dongha Lee, SeongKu Kang, Hyunjun Ju, Chanyoung Park, and Hwanjo Yu. Bootstrapping user and item representations for one-class collaborative filtering. In *Proceedings of the 44th international ACM SIGIR conference on Research and Development in information retrieval*, pages 317–326, 2021.

[5] Weiming Liu, Chaochao Chen, Xinting Liao, Mengling Hu, Jianwei Yin, Yanchao Tan, and Longfei Zheng. Federated probabilistic preference distribution modelling with compactness co-clustering for privacy-preserving multi-domain recommendation. In *IJCAI*, pages 2206–2214, 2023.

[6] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 43–52, 2015.

[7] Srivastava Nitish. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15:1, 2014.

[8] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[9] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618*, 2012.

[10] Rui Sun, Xuezhi Cao, Yan Zhao, Junchen Wan, Kun Zhou, Fuzheng Zhang, Zhongyuan Wang, and Kai Zheng. Multi-modal knowledge graphs for recommender systems. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 1405–1414, 2020.

[11] Zhulin Tao, Xiaohao Liu, Yewei Xia, Xiang Wang, Lifang Yang, Xianglin Huang, and Tat-Seng Chua. Self-supervised learning for multimedia recommendation. *IEEE Transactions on Multimedia*, 25:5107–5116, 2022.

[12] Qifan Wang, Yinwei Wei, Jianhua Yin, Jianlong Wu, Xuemeng Song, and Liqiang Nie. Dualgnn: Dual graph neural network for multimedia recommendation. *IEEE Transactions on Multimedia*, 25:1074–1084, 2021.

[13] Wei Wei, Chao Huang, Lianghao Xia, and Chuxu Zhang. Multi-modal self-supervised learning for recommendation. In *Proceedings of the ACM Web Conference 2023*, pages 790–800, 2023.

[14] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, and Tat-Seng Chua. Graph-refined convolutional network for multimedia recommendation with implicit feedback. In *Proceedings of the 28th ACM international conference on multimedia*, pages 3541–3549, 2020.

[15] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. Mmgcn: Multi-modal graph convolution network for personalized recommendation of micro-video. In *Proceedings of the 27th ACM international conference on multimedia*, pages 1437–1445, 2019.

[16] Jiancan Wu, Xiang Wang, Fuli Feng, Xiangnan He, Liang Chen, Jianxun Lian, and Xing Xie. Self-supervised graph learning for recommendation. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 726–735, 2021.

[17] Jinfeng Xu, Zheyu Chen, Shuo Yang, Jinze Li, Hewei Wang, and Edith C-H Ngai. Mentor: Multi-level self-supervised learning for multimodal recommendation. *arXiv preprint arXiv:2402.19407*, 2024.

[18] Tiansheng Yao, Xinyang Yi, Derek Zhiyuan Cheng, Felix Yu, Ting Chen, Aditya Menon, Lichan Hong, Ed H Chi, Steve Tjoa, Jieqi Kang, et al. Self-supervised learning for large-scale item recommendations. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pages 4321–4330, 2021.

[19] Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations. *Advances in neural information processing systems*, 33:5812–5823, 2020.

[20] Junliang Yu, Hongzhi Yin, Min Gao, Xin Xia, Xiangliang Zhang, and Nguyen Quoc Viet Hung. Socially-aware self-supervised tri-training for recommendation. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pages 2084–2092, 2021.

[21] Junliang Yu, Hongzhi Yin, Jundong Li, Qinyong Wang, Nguyen Quoc Viet Hung, and Xiangliang Zhang. Self-supervised multi-channel hypergraph convolutional network for social recommendation. In *Proceedings of the web conference 2021*, pages 413–424, 2021.

[22] Xin Zhou. Mmrec: Simplifying multimodal recommendation. In *Proceedings of the 5th ACM International Conference on Multimedia in Asia Workshops*, pages 1–2, 2023.

[23] Xin Zhou, Aixin Sun, Yong Liu, Jie Zhang, and Chunyan Miao. Selfcf: A simple framework for self-supervised collaborative filtering. *ACM Transactions on Recommender Systems*, 1(2):1–25, 2023.

[24] Xin Zhou, Hongyu Zhou, Yong Liu, Zhiwei Zeng, Chunyan Miao, Pengwei Wang, Yuan You, and Feijun Jiang. Bootstrap latent representations for multi-modal recommendation. In *Proceedings of the ACM Web Conference 2023*, pages 845–854, 2023.