多模态推荐系统综述：分类体系、评估与未来方向

HONGYU ZHOU,南洋理工大学，新加坡XIN ZHOU,南洋理工大学，新加坡ZHIWEI ZENG,南洋理工大学，新加坡LINGZI ZHANG,南洋理工大学，新加坡ZHIQI SHEN,南洋理工大学，新加坡

arXiv:2302.04473v1 [cs.IR] 9 Feb 2023

推荐系统通过基于隐式交互（例如购买和点击）对用户偏好和物品属性进行建模，已成为帮助用户发现感兴趣物品的热门有效工具。人类通过处理模态信号（例如音频、文本和图像）来感知世界，这启发研究者构建能够理解和解释来自不同模态数据的推荐系统。这些模型能够捕捉不同模态之间的潜在关系，并有望恢复单模态方法和隐式交互无法捕捉的互补信息。本综述旨在全面回顾近期多模态推荐研究。具体而言，本文展示了清晰的流程，介绍了各个步骤中常用的技术，并根据所采用的方法对模型进行分类。此外，我们还设计了一个代码框架，帮助该领域的新研究者理解其原理和技术，并轻松运行最先进模型。我们的框架位于：https://github.com/enoche/MMRec。

CCS 概念：• **信息系统** → **推荐系统**；**多媒体与多模态推荐**。

附加关键词与短语：多模态推荐、多模态、多媒体、推荐系统、基于内容的推荐、推荐系统综述、深度学习、神经网络、机器学习、分类体系

ACM 参考格式：
Hongyu Zhou、Xin Zhou、Zhiwei Zeng、Lingzi Zhang 和 Zhiqi Shen。2023年。《多模态推荐系统综述：分类体系、评估与未来方向》。1，1（2023年2月），33页。https://doi.org/XXXXXX

1 引言

随着社会的发展，网络上呈现出数量庞大的物品和信息，用户很难判断哪些有用，并从中挑选出自己偏好的物品。推荐系统出现后，信息过载问题得到了解决[34]。它根据历史交互数据和公开信息，预测用户对所有物品的评分或偏好，并推荐用户可能会选择的最相关、最有可能的物品

作者地址：Hongyu Zhou，hongyu.zhou@ntu.edu.sg，新加坡南洋理工大学；Xin Zhou，xin.zhou@ntu.edu.sg，新加坡南洋理工大学；Zhiwei Zeng，zhiwei.zeng@ntu.edu.sg，新加坡南洋理工大学；Lingzi Zhang，lingzi001@e.ntu.edu.sg，新加坡南洋理工大学；ZhiqiShen，ZQShen@ntu.edu.sg，新加坡南洋理工大学。

允许免费制作本作品全部或部分内容的数字副本或纸质副本，用于个人或课堂使用，前提是副本不得用于营利或商业利益，也不得为此进行复制或传播，且副本须附有本声明及首页的完整引文。对于本作品中由 ACM 以外的其他方拥有版权的部分，必须尊重其版权。注明出处后，可以摘录。若要以其他方式复制、再版、发布到服务器或向列表重新分发，须事先获得明确许可和/或支付费用。请通过 permissions@acm.org 申请许可。© 2023年计算机协会。XXXX-XXXX/2023/2-ART $15.00 https://doi.org/XXXXXX

物品信息（如点击和销量）。推荐算法可分为协同过滤（CF）、基于内容的过滤和混合推荐系统 [43]。协同过滤基于对用户历史行为数据的分析与收集，其中包括历史交互（例如点击、浏览、购买）和用户偏好（例如评分）。基于内容的过滤根据用户画像和用户的物品画像推荐产品。物品以关键词描述，而用户画像则体现用户喜欢的物品类型。这种方法的核心理念是，用户很可能会选择以前喜欢过的相似物品。混合方法结合协同过滤和基于内容的过滤等不同技术，以获得更好的结果。

正如我们所介绍的，这些传统推荐方法存在不足：它们需要大量用户与物品之间的交互，才能做出更准确的推荐。交互很少甚至没有交互的用户和物品会影响推荐的准确率。为缓解数据稀疏问题和冷启动问题，多模态信息被引入推荐系统。多模态推荐利用辅助多模态信息补充历史用户-物品交互，从而提升推荐性能。多模态模型能够表示并发现不同模态之间的潜在关系，还可能恢复单模态方法和隐式交互无法捕捉的互补信息。这些能力对于自然语言处理也十分必要，能帮助其在各类人工智能任务中实现人类水平的理解。多媒体数据不仅展示用户与物品之间的关系，也反映用户在不同模态下的偏好。对于同一模态，数据可能反映不同物品之间的相似性和语义含义。为了将多模态信息融入推荐系统，当前方法会从不同模态中提取特征，然后将模态融合结果用作侧信息或物品表示。VBPR是首个考虑将视觉特征引入推荐系统的模型，它通过拼接视觉嵌入与ID嵌入来生成物品表示。基于GCN的方法会生成每种模态的表示，再将它们融合为最终表示。除了融合模态表示外，基于知识图谱的模态侧信息也被引入多模态推荐中。多种技术已被应用于多模态推荐，旨在为用户和物品寻找更优的表示，从而实现更准确的推荐结果。

如今，人们普遍使用各类在线分享平台，例如时尚、新闻、短视频和音乐推荐平台。这些热门在线平台包含大量多模态信息，会影响用户的选择。由于人们经常使用这些包含海量多模态信息的平台，因此必须应用多模态推荐，以更准确地学习用户偏好。与传统推荐不同，这些应用利用了物品的多模态内容信息，例如视频帧、音轨和物品描述。MMGCN [59], MGAT [49], DualGNN [53] 和 SLMRec [48] 是微视频推荐模型，它们利用视频中的描述、字幕、音频和帧来建模用户对微视频的多模态偏好。时尚推荐面临着构建高效推荐系统的困难，因为其中涉及的特征复杂且具有主观性。 [7, 19] 旨在利用多模态信息克服这一问题。以往的新闻推荐主要关注标题和文本等单模态信息，却忽略了图像等视觉信息。 [61, 63] 尝试利用多模态来提升推荐性能。

由于具备上述优势且有其必要性，多模态推荐系统得到了广泛应用。然而，目前尚无综述探讨用于

多模态推荐的技术。为了为研究人员提供指导，我们概述了多模态推荐系统，并根据所采用的方法对所综述的论文进行分类。本综述旨在通过展示清晰的流程，并介绍每个步骤所用的技术，说明多模态推荐系统的运作方式。我们希望帮助新研究人员理解多模态推荐系统中常用的原理和技术。此外，我们还提供了一个完善的代码框架，其中包含已实现的最先进模型以及预先提取的多模态特征，便于研究人员运行这些最先进模型，或基于该框架开发自己的模型。

1.1 文献检索策略

由于本综述侧重于回顾多模态推荐领域的研究，我们检索了大多数与推荐系统相关的顶级会议，例如 WWW、SIGIR、KDD、ICLR、AAAI、WSDM 和 RecSys，以及 TKDD、TMM 等顶级期刊。同时，我们也利用谷歌学术搜索相关论文。我们使用多模态推荐、多模态推荐、多媒体推荐、内容+RS、上下文+RS 等关键词检索相关论文。获取论文列表后，我们还查阅了这些论文的相关工作，以确保涵盖了大多数关于多模态推荐的论文。

1.2 与其他综述的差异

鉴于推荐系统的重要性和普及程度，已有一些综述探讨了推荐领域 [23, 39, 44, 54]。据我们所知，这些综述并未考虑多模态信息，只有一篇综述 [8] 讨论了多媒体推荐并提及多模态信息。尽管对多模态的讨论总是涉及模态和媒介，但这两个术语并非同义词，只是有所重叠 [22]。媒介是借助技术来存储、移动和传输信息的物理通道。它是用于生成和传播信息的工具和物质资源。模态则是人类的感官通道，在与系统交互时充当用户输入模态。它可以被视为一种信息表示方式。对于一种媒介，其中可能包含若干种模态。例如，书籍被视为媒介，而文本的不同语言则被视为不同模态。此外，本综述旨在将推荐系统与多媒体相结合，并仅根据媒介类型和应用对论文进行分类。它考察了这些模态可用于哪些应用（如音乐推荐、电影推荐等），但忽略了推荐系统中用于利用多模态信息的隐藏算法和方法，也忽略了不同模态的融合方法。

不同于现有的综述，本文综述按照多模态推荐系统中通常采用的不同方法进行组织。上述多媒体综述仅希望根据媒介类型对模型进行划分。我们希望实现的是，让初学者能够了解多模态推荐系统的主要步骤、可用于实现其目标的模型，以及每个步骤中主要采用的技术，例如从不同模态中提取特征并融合不同模态的表示。本文综述的另一项不同之处在于，我们不仅提供了这些多模态推荐模型的总结，还构建了一个框架，其中包含采用不同方法的最先进模型以及若干常用数据集。初学者阅读本文综述后即可轻松运行这些基线模型，也可以按照规定格式便捷地添加新数据集，开发自己的模型。

1.3 调研的组织与贡献

本文综述的主要内容是多模态推荐研究概览。我们的工作旨在总结该领域采用的主要方法，并强调多模态内容的重要性。此外，我们还将提供一个通用框架，其中包含常用模型的代码实现，可帮助初学者快速运行自己的数据或在该框架中实现这些模型。本文综述的组织结构如下：

- 引言
- 多模态推荐流程
- 特征提取
- 模型分类
- 模态融合
- 测量与优化
- 数据集与实验结果
- 挑战与未来研究方向
- 结论

在介绍了本综述的组织结构后，我们将主要贡献总结如下。

- 我们对多模态推荐系统（MMRec）进行了理论综述，概述了学习和利用模态信息的各种策略，并全面阐释了模型采用的技术以及各个 MMRec 模型的工作原理。

- 我们介绍了推荐系统利用多模态信息的主要步骤，并展示了 MMRec 的流程。我们还讨论了如何选择合适的数据集、特征提取方法和优化方法，为读者开发多模态推荐系统提供了清晰的指南。

- 我们在四个公开数据集上进行了大量实验，以验证当前最先进的多模态推荐模型的有效性和效率。此外，我们还讨论了影响模型性能的关键因素（例如，不同的模态信息、不同的数据划分方法）。

- 此外，我们创建了一个开源框架 MMRec [74][1]，，可用于实现本综述中提及的多模态推荐模型基线。读者也可以使用该代码框架开发自己的模型。我们希望该框架能帮助初学者更轻松地理解多模态推荐模型的工作原理，并便捷地开发自己的模型，而无需考虑其他步骤，只需专注于模型设计，其他步骤则由框架完成。此外，我们还创建了一个开源代码仓库，其中收录了所有综述论文和单模态模型[2]。

2 多模态推荐流程

多模态推荐系统通常遵循图1所示的步骤。根据设置的不同，模态融合方法会应用于不同步骤，这些步骤在图中以虚线框标出。以下将介绍各个步骤的详情。

多模态推荐系统旨在利用多模态特征，学习用户和物品的信息丰富的表示。第一步是从原始数据中提取模态特征。获得模态特征后，可以选择在将其输入

[1]https://github.com/enoche/MMRec [2]
https://github.com/hongyurain/Recommendation-with-modality-information

多模态模型之前进行融合，在模型的中间层进行融合，或者融合多模态模型的输出。系统会采用不同的技术来学习用户-物品关系并利用多模态信息，从而结合匹配的目标函数提高推荐准确率。

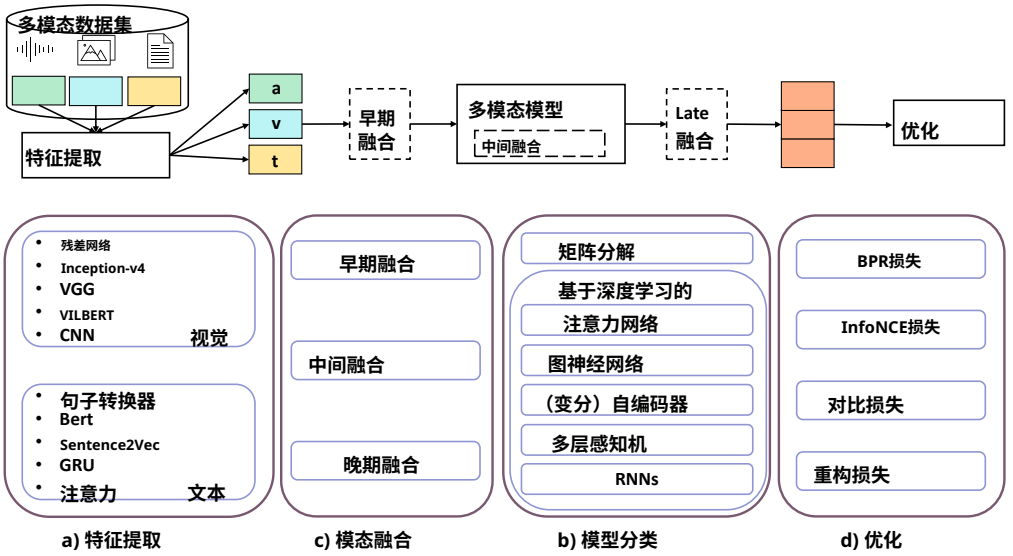


图 1. 多模态推荐流程

3 特征提取

常用数据集可能并不包含所有模态特征，因此也需要根据其包含的模态选择合适的数据集。表 1 展示了最常用的数据集，包括亚马逊数据集、MovieLens、TikTok 和快手。亚马逊数据集始终包含图像和文本信息，而且亚马逊旗下有许多类别，因此数据集的选择也取决于具体类别。婴儿、体育、美容和艺术类别常用于多模态推荐 [67, 71, 80]。对于时尚推荐等一些特殊领域，则通常使用亚马逊旗下的男装和女装类别。食品推荐使用包含图像和用户评论的 Yelp 数据集。新闻推荐使用 MIND 数据集。TikTok、快手和 MovieLens 数据集用于短视频推荐，其中包含文本、视觉和音频信息 [53, 59], , 但 TikTok 和快手并未公开。如果需要特定的模态信息，或者研究领域尚无相应数据集，也可以从网站抓取数据，或利用现有数据集来构建新的数据集。

特征提取的目的是利用嵌入，以低维且可解释的方式描述模态特征。模态数据可以通过两种方式加以利用。第一种是将预先提取的模态数据输入模型；第二种是通过应用端到端学习，将原始数据输入模型。根据大多数 MMRec 模型的设置，我们的框架采用第一种方法，使用预先提取的特征。不同的模态数据有各自特定的提取方法。表2总结了多模态特征提取方法。根据不同模态，数据资源和方法也各不相同。我们根据数据资源对任务进行了简单划分，并列出了主要使用的提取方法及示例模型。

数据集	模型
亚马逊	VBPR[15], VMCF[38], AMR[47],VECF[7],PAMD[14],PMGT[32], LATTICE[67], JRL[69], BM3[80], GraphCAR[62], DVBPR[19], MAML[27], DMRL[26], MVGAE[64],MMCPR[33],HCGCN[36], FREEDOM [73], DRAGON [71]
Kwai	MMGCN[59], SLMRec[48], GRCN[58], EliMRec[31], InvRL[10], A2BM2GL[3]
TikTok	MMGCN[59], DualGNN[53], MGAT[49], SLMRec[48], GRCN[58], MMGCL[65], EgoGCN[5],EliMRec[31],InvRL[10],A2BM2GL[3], LUDP[24]
MovieLens	MMGCN[59], DualGNN[53], MGAT[49], SLMRec[48], GRCN[58], MMGCL[65], MKGAT[46], EgoGCN[5], EliMRec[31], InvRL[10], A2BM2GL[3], LUDP[24]
Yelp	PAMD[14]
MIND	IMRec[63]
大众点评	MKGAT[46]
Alishop	MVGAE[64]
Pinterest	ACF[6], AMR[47]
爬取数据集	PAMD[14], MM-Rec[61],UVCAN[30],DVBPR[19], ACF[6]

表 1. 常用数据集

我们可以发现，对于视觉特征，常用的技术是卷积神经网络（CNN）及其变体。对于 [53, 59] 采用快手和 TikTok 的模型而言，这两个数据集并未公开，即使是拥有这些数据集的研究人员，也使用了预先提取的视觉特征和文本特征。对于文本特征，会根据模型的目的使用不同的预训练语言模型。

4 模型分类

如表3所示，我们根据模型所采用的方法对模型进行了分类。矩阵分解和基于多层感知机的方法是较早用于推荐的方法。过去，研究人员对多模态信息的利用较为简单，例如，VBPR [15] 将卷积神经网络提取的视觉特征与ID嵌入拼接，作为物品表示，然后输入矩阵分解模块。尽管这种方法很简单，但它可以将模态信息作为侧信息，帮助解决冷启动和数据稀疏问题。随后，许多深度学习技术被引入多模态推荐系统，并显著提升了对模态信息的利用效果。例如，自编码器和变分自编码器技术被引入推荐系统。此外，MMGCN [59] 开始使用图卷积网络学习每种模态的表示，然后将模态表示与ID嵌入融合，作为最终的物品表示。基于这一GCN结构，研究人员引入了知识图谱 [46], 注意力机制、边细化 [58] 以及其他融合方式，以

模态	内容	方法	模型
视觉	微视频	已发布特征	MMGCN[59], DualGNN[53], MGAT[49], SLMRec[48], GRCN[58], HHFAN[4], MMGCL[65], EgoGCN[5], InvRL[10], A2BM2GL[3], LUDP[24]
视觉	图像	VGG-19	VECF[7], PAMD[14]
		VILBERT	MM-Rec[61]
		卷积神经网络+多层感知机	IMRec[63]
		Inception-v4	PMGT[32]
		Inception-v3	UVCAN[30]
		深度卷积神经网络	VBPR[15], VMCF[38], LATTICE[67], FREEDOM [73], HCGCN[36], JRL[69], BM3[80], GraphCAR[62], DVBPR[19], DRAGON [71]
		Caffe	MAML[27],DMRL[26]
		残差网络	MKGAT[46], MVGAE[64], GraphCAR[62], ACF[6], AMR[47], EgoGCN[5], EliMRec[31], InvRL[10]
文本	视频标题	已发布特征	MMGCN[59], DualGNN[53], MGAT[49], SLMRec[48], GRCN[58], HHFAN[4], MMGCL[65], EgoGCN[5], A2BM2GL[3], LUDP[24]
文本	评论	GRU	VECF[7]
		Glove+多层感知机	PAMD[14]
		PV-DBOW	JRL[69]
		PV-DM	MAML[27]
		BERT	DMRL[26]
		Word2Vec+SIF	MKGAT[46]
文本	标题	VILBERT	MM-Rec[61]
		注意力	IMRec[63]
文本	物品描述	BERT	PMGT[32]
		词频-逆文档频率	ADDVAE[50]
		Word2Vec+SIF	MKGAT[46]
		句子-转换器	格架[67], FREEDOM [73], HCGCN[36], BM3[80], DRAGON [71]
		Sentence2Vec	MVGAE[64], EgoGCN[5], EliMRec[31], InvRL[10]

表2. 模态特征提取方法。

帮助学习更好的表示。近年来，随着预训练与微调在计算机视觉和自然语言处理领域取得良好效果，这种方法也被引入多模态推荐系统。该表仅展示了模型分类概况，后续内容将介绍各个模型，以及它们如何通过不同技术利用模态信息。由于某些模型采用了多种方法的组合，因此可能会在表中出现两次。

方法	模型
矩阵分解	VBPR[15], GraphCAR[62], VMCF[38], AMR[47], ACF[6], ConvMF[20], DeepCoNN[70], DVBPR[19]
多层感知机	JRL[69]
（变分）自编码器	CKE[66],ADDVAE[50], MVGAE[64]
注意力网络	UVCAN[30], ACF[6], MAML[27], MMRec[61], DMRL[26], IMRec[63]
图神经网络(GNN)	HHFAN[4], MKGAT[46], MMGCN[59], DualGNN[53], MGAT[49], MEGCF[28], GRCN[58], LATTICE[67], FREEDOM [73], HUIGN[57], HGCL[2], HCGCN[36], EgoGCN[5], EliMRec[31], InvRL[10], A2BM2GL[3],LUDP[24], DRAGON [71]
RNN+注意力	VECF[7]
矩阵分解+卷积神经网络	ConvMF[20],DeepCoNN[70], DVBPR[19]
图神经网络+预训练和微调	PMGT[32],PAMD[14], MMCPR[33]
图神经网络+自监督学习	SLMRec[48],BM3[80], MMGCL[65]

表3. 模型分类

4.1 矩阵分解模型

矩阵分解是一种协同过滤方法。通过矩阵分解方法，用户-物品交互矩阵被分解为两个维度较低的矩形矩阵的乘积。

评论和评分等侧信息已被用于辅助推荐。但在这项工作之前，物品的视觉外观一直被忽略。VBPR [15] 利用预训练的深度卷积神经网络提取视觉特征，然后将其线性变换到视觉空间。接着，将视觉特征与 ID 嵌入结合，作为物品表示，并将该表示输入矩阵分解模型，以进行偏好预测。AMR [47] 指出了 VBPR 等现有多媒体推荐系统的脆弱性，并利用对抗学习构建更稳健的推荐系统。VMCF [38] 构建了一个产品关联网络，其中节点表示产品，边表示“也被查看”关系。该方法不仅利用外观作为侧信息，还充分利用了关系中隐藏的信息。

ConvMF [20] 旨在利用卷积神经网络架构捕获文档信息，但现有的卷积神经网络方法用于解决分类问题，并不适合进行推荐等回归任务。随后，该模型将概率矩阵分解与卷积神经网络相结合，用于生成推荐。

DVBPR [19] 将视觉信号融入推荐模型，并将卷积神经网络模块整合到矩阵分解模块中，以训练图像表示和推荐

模型，从而共同提升性能。此外，使用该模型可以生成与训练语料库中的图像相似的图像。给定用户和类别，该模型结合生成对抗网络，可以生成符合用户偏好的新物品。

GraphCAR [62]将多媒体内容与传统的协同过滤方法相结合。它首先将用户-物品交互矩阵与用户属性矩阵或物品特征矩阵相乘，然后使用两层图卷积网络生成用户和物品的潜在表示。经过非线性特征变换后，再通过用户和物品潜在向量的内积计算偏好分数。

4.2 基于深度学习的模型

4.2.1 多层感知机。

丰富的多模态特征能够体现用户偏好的不同方面。在 JRL [69] 框架中，利用多层感知机将不同来源的信息投射到统一空间中，并学习将多模态特征转换为用户和物品嵌入的权重。同时，利用特征重构损失学习相应的用户和物品表示。

4.2.2 卷积神经网络。

卷积神经网络是一种具有卷积层和池化操作的前馈神经网络，用于处理多模态特征。大多数基于卷积神经网络的推荐系统利用卷积神经网络提取特征，或通过捕捉全局和局部特征来学习表示。

由于词袋模型固有的局限性，它难以充分捕捉文档信息。ConvMF [20] 利用卷积神经网络捕捉图像或文档的局部特征，并将概率矩阵分解与卷积神经网络结合起来，以同时利用交互信息和描述信息。

DeepCoNN [70] 利用预训练词嵌入技术，从评论文本中提取用户和物品的语义表示。然后，将用户和物品的语义嵌入输入两个并行的卷积神经网络模块，以学习物品属性和用户表示。顶层用于让用户和物品的隐因子相互作用，并共同学习物品属性和用户行为。

DVBPR [19] 利用孪生卷积神经网络框架，并将卷积神经网络的最后一层作为物品表示。这些具有时尚感知能力的图像表示与推荐系统一起以端到端方式进行训练。

4.2.3 注意力网络。

注意力机制的灵感来源于人类注意力。注意力机制模拟用户对不同模态或同一模态的不同方面给予不同程度的关注，从而能够捕捉用户偏好并丰富用户表示。

传统协同过滤系统并未针对多模态推荐进行良好设计，因而忽视了对多模态信息的利用。ACF [6] 设计了物品级和组件级注意力模块，并将其融入传统协同过滤模型中以进行推荐。用户的需求通常由固定的特征向量表示。然而，采用不同向量表示不同物品，应该能更好地刻画用户偏好。MAML [27] 特别关注多样化的物品特征，为不同物品建模用户偏好。检索到的文本特征和视觉特征首先进行拼接，然后输入多层神经网络。该方法引入了一种注意力神经网络，利用视觉特征与文本特征的融合，识别用户对各种产品属性的不同偏好。DMRL [26]也对用户的不同偏好进行建模，但关注的是各个模态的不同方面。DMRL 首先采用解耦表示方法，学习若干模态的独立因子表示

。基于这种解耦表示，它设计了一种权重共享的多模态注意力机制，能够将用户的注意力引向各个模态中与因子相关的不同方面。结合学习到的注意力以及用户和物品表示，或许能够预测用户对目标物品的偏好。DMR [55] 也学习解耦表示，以捕捉多种模态中的互补信息和共通信息。随后，它将这些解耦表示应用于推荐任务。

上述微视频推荐方法仅考虑用户的多模态信息，却忽略了微视频的多模态信息。UVCAN [30] 为用户和微视频学习多模态表示，以实现个性化微视频推荐。不同于仅关注视频注意力的方法，UVCAN 利用物品与微视频之间的协同注意力机制，以更好地联合执行注意力计算。研究使用堆叠注意力网络处理用户档案和微视频多模态特征，将多模态特征作为输入查询，并通过多步推理获得视频注意力。在学习视频表示后，将其作为输入查询，通过多步推理捕获用户注意力。

新闻表示对于新闻推荐至关重要，但现有的大多数新闻推荐模型忽略了新闻的视觉信息，只考虑标题等文本信息。MM-Rec [61] 融合这两种模态来学习新闻表示。首先，它执行目标检测，从图像中获取感兴趣区域（ROIs）。然后，使用视觉语言模型对文本和图像 ROIs 进行编码，并使用协同注意力 Transformer 学习其中固有的跨模态关系。此外，它还引入了跨模态候选感知注意力网络，学习已点击新闻与候选新闻之间的关系，从而更好地捕获用户对候选新闻的偏好。研究人员在设计多模态建模模块时，始终忽略了决策过程。IMRec [63] 基于这样一个事实：用户在搜索新闻时会更加关注视觉印象。它开发了局部印象建模模块，从印象图像中提取重要线索，并弥合语义与印象之间的差距，从而增强新闻标题的语义含义。全局印象建模模块用于融合不同线索，同时不丢失结构信息。实验数据集采用 MIND 数据集，并添加了印象图像的截图。

4.2.4 循环神经网络。

现有的时尚推荐研究主要利用多模态信息来构建新闻表示，但没有考虑到用户会关注图像中的不同区域来做出决策。VECF [7]首先使用 VGG 模型获取预先分割的图像区域，然后对人类感知进行建模，并利用图像区域上的注意力机制来捕捉用户偏好。此外，研究还将用户评论信息作为弱监督信号，通过使用标准 LSTM 将注意力图像嵌入融入词生成过程，从而更好地捕捉用户偏好。

4.2.5 自编码器。

协同过滤方法总是受到用户-物品交互数据匮乏的制约。现有方法已利用丰富的侧信息来改进推荐系统。CKE [66] 探究了如何利用知识库中收集的异构信息。该方法使用网络嵌入过程贝叶斯 TransR 提取结构知识嵌入，使用贝叶斯堆叠去噪自编码器提取文本的潜在表示，并使用贝叶斯堆叠卷积自编码器提取视觉方面的潜在表示。随后，提出了协同联合学习过程。

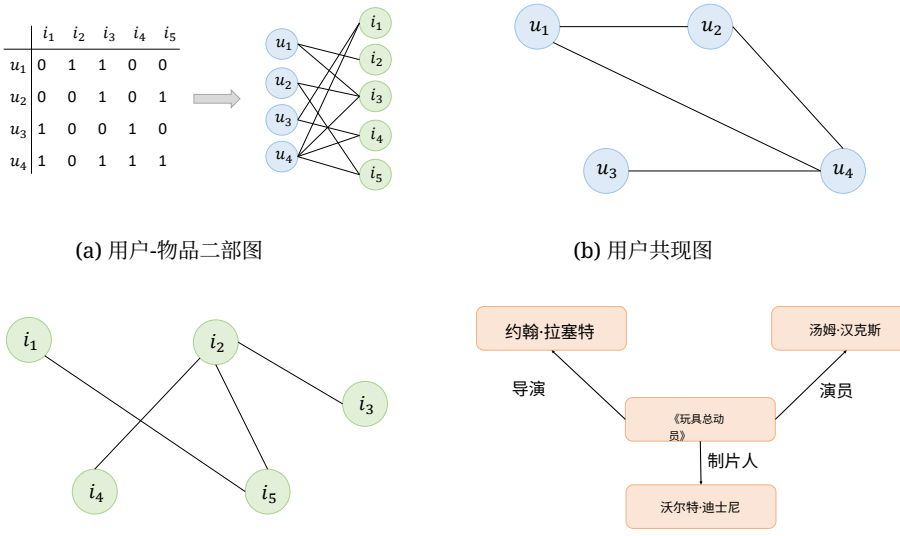
4.2.6 变分自编码器。

经典方法将用户偏好表示为潜在空间中的单个向量，无法表达其不同的方面。ADDVAE [50] 利用解耦表示，更好地捕捉受多个隐因子影响的用户偏好。它利用文本内容学习第二种解耦表示，通过互信息最大化耦合两个 MacridVAE 网络中的解耦因子，然后使用注意力机制使这些表示彼此对齐，从而解决用户-物品交互的稀疏问题，并将表示中不可解释的维度映射到词语。

用户-物品交互的稀疏度和特征噪声会影响用于推荐的图嵌入方法。MVGAE [64] 是一种多模态变分图自编码器模型，它利用模态特定变分编码器来学习节点表示。每个节点都使用了高斯变量，其中均值向量传达语义含义，而方差向量表示各模态的噪声水平。不同模态属性的权重通过专家乘积（PoE）进行组合，该组合取决于噪声水平。

4.2.7 图神经网络。

图神经网络已被证明能够有效学习不同领域图数据的表示 [9, 12, 29, 72, 79]。通过将用户-物品关系视为图关系，可以轻松地将其扩展到推荐系统中。推荐系统中的大多数数据都可以用图 2 所示的图结构表示。GNN 的核心思想是在传播过程中聚合邻居节点的信息并更新节点表示。我们根据融合步骤对这些基于图神经网络的模型进行了大致分类。



(a) 用户-物品二部图 (b) 用户共现图

(c) 物品-物品关系图 (d) 知识图谱

图2. 多模态推荐系统中的图结构

直接融合。 直接融合为每个节点提取的多模态嵌入。将多模态特征作为节点表示，然后应用图神经网络技术。其他模型会构建模态特定图来学习每种模态的表示，再融合 GNN 的输出作为最终表示。

PMGT [32]是一个预训练模型，利用融合后的多模态特征和交互。对于提取出的多模态嵌入，此处采用注意力机制来获得多模态表示。随后，将位置嵌入和基于角色的嵌入添加到多模态表示中，作为初始化的节点嵌入。接着，基于节点嵌入开始预训练和下游任务。

如图 2(d) 所示，基于知识图谱的推荐利用重要的外部知识来提升性能。然而，它们往往忽略了多模态特征的不同数据类型。MKGAT [46] 采用基于实体的方法构建多模态知识图谱。多模态知识图谱实体编码器采用不同的编码器对各种数据类型进行编码，并使用多模态知识图谱注意力层聚合邻居的信息。学习到每个实体的表示后，知识图谱嵌入将其作为输入，以学习知识推理关系。

异构图融合。 异构图是一种包含不同类型节点的信息网络。我们将用户和物品视为不同类型的节点，并将它们之间的交互视为边，如图 2(a) 所示。对于每种模态，如图 3 所示，直接在用户-物品二部图上使用 GCN（MMGCN [59], DualGNN [53]）或 GAT（MGAT [49]）进行特征聚合。

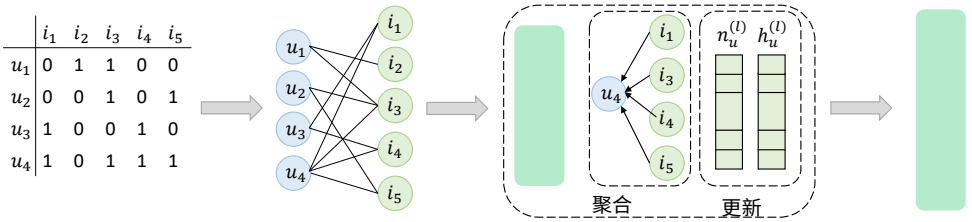


图 3. 用户-物品二部图中图神经网络的整体框架

以往的多模态推荐工作更注重利用多模态信息更好地表示物品，却忽略了用户与物品之间的信息交换。MMGCN [59] 框架基于图神经网络实现消息传递，并针对特定模态构建用户-物品二部图。在完成聚合与组合过程后，可以利用多跳邻居的信息更好地理解用户偏好，并生成更优的用户表示。

MMGCN 直接将图神经网络应用于交互图，并平等对待邻居以进行信息交换，因此容易受到冗余甚至噪声信息的影响。基于 MMGCN 框架的 MGAT [49]采用原始 GCN 进行聚合，并以相同的方式组合聚合结果。为管理每种模态的信息传递，它引入了一种新的门控注意力机制。因此，它能够记录复杂的用户交互模式，并提供更好的用户推荐。
DualGNN [53] 还引入了模型偏好学习模块，引导用户关注不同模态。此外，学习到的多模态融合模式还能减轻模态缺失带来的影响。

上述基于图神经网络的模型（如 MMGCN）通过拼接等操作将彼此孤立的模态特定图合并为一个图，这容易低估不同模态之间的相互关系。EgoGCN [5] 提出了一种有效的图融合方法——自我融合，以打破这种隔离。它构建模态特定图的方式与 MMGCN 相同，但不局限于在单模态图中传播信息，而是通过聚合邻居节点中有用的跨模态消息来传播信息。多模态传播的结果与 ID 嵌入传播的结果拼接后，作为最终表示。

利用分布不均的数据集，并将单模态特征粗略融合为多模态特征，会导致单模态偏差。EliMRec [31] 指出，可以通过因果推断和反事实分析消除单模态偏差。单模态信息对最终结果的因果效应（总效应，TE）由两部分构成：通过单模态 → 物品表示 → 最终结果产生的总间接效应（TIE），以及通过单模态 → 最终结果产生的自然直接效应（NDE）。从 TE 中减去 NDE 即可消除单模态偏差。该方法沿用了 MMGCN 的相同结构，但消除了单模态偏差。

多模态特征提取（MFE）和用户兴趣建模（UIM）是多模态推荐的两个主要模块。然而，现有研究忽略了两者之间的不匹配问题。MEGCF [28] 利用与用户行为和偏好密切相关的用户情感信息和语义实体，有助于将内容导向的多模态特征提取转变为用户导向的多模态特征提取。为解决这一不匹配问题，在转变为用户导向的多模态特征提取后，该方法采用分离式框架建立 MFE 与 UIM 之间的关联。图卷积网络是一种常用技术，能够捕捉节点之间的高阶关联，因此在此处得到应用；此外，该方法将情感信息纳入邻居聚合的权重计算，从而增强图卷积操作。

同质图融合。异构图是一种仅包含一种节点和边类型的信息网络。如图 2(c) 所示，基于 GNN 的微视频推荐方法仅考虑用户与物品之间的关系，却忽略了同质数据间丰富的关系。通过考虑每种模态下的物品间关系（Lattice [67]）来构建同质图，或如图 2(b) 所示，考虑用户共现频率来构建图，有助于捕捉同类型节点之间的关系。

包含多模态信息的异质用户-物品交互图利用了历史交互，但忽略了能够更好地刻画物品表示的语义物品间关系。LATTICE [67] 在模态感知结构学习层中利用模态特征学习物品结构，然后通过物品亲和度进行图卷积，以学习更优的表示。多模态融合引入注意力机制，以获取各模态的重要性差异，并融合物品嵌入，从而得到更强大的表示。此外，自监督对比辅助任务被用于最大化各模态物品嵌入与融合后的多模态表示之间的一致性。学习到物品表示后，现有的协同过滤方法会利用这些表示来提升性能。DualGNN [53]不仅利用语义物品间关系构建同质图，还考虑了用户共现时间。它基于用户-微视频二部图和用户共现图，引导用户关注各种模态。此外，通过学习到的多模态融合模式，可以减轻模态缺失带来的影响。DRAGON [71] 利用用户-物品二部图学习用户-物品关系。此外，它还利用物品语义图和用户共现图，捕捉两组内部的关系。用户和物品的双重表示有助于捕捉互补信息。

此前尚无推荐系统会考虑用户的意图。为了引导用户和物品表示的学习，HUIGN [57] 尝试学习多层次用户意图。它

构建了两种不同类型的信息聚合方法和一个共同交互项目图。层内聚合利用物品之间的相关性，获得能够向物品传达用户意图的表示。层间聚合对高层和低层用户意图之间的关系进行建模。超节点被构建为粒度更粗的用户意图。

HGCL [2]指出，大多数图神经网络技术（例如，GCN [21], GraphSage [13], GAT [51]）都是为同质图设计的，但推荐系统中的用户和物品属于异构数据。当前基于图神经网络的推荐系统的性能会受到两方面影响：无法准确捕捉用户-物品图中节点的异质性，以及过度平滑问题。在 HGCL [2], 中，异构图编码器网络利用随机游走模型生成特定节点类型的同质图，以获取用户-项目二部网络中的异质性；随后，在同质图上应用图神经网络来学习表示。此外，图对比学习框架也应用于具有特定节点类型的图，通过对全局表示与局部补丁表示之间的互信息取平均，学习节点的表达性表示。

常用的模态融合方法是将多模态信息融合到 ID 嵌入中，以学习有意义的物品表示。LATTICE 开始认识到，模态特征可用于构建物品之间的潜在结构，而不是将模态信息融合到用户-物品二部图中。FREEDOM [73] 发现，学习潜在结构的效率低下，因此在训练前冻结物品-物品图有助于学习更好的物品语义关系。它还根据度敏感边剪枝技术计算出的分数对图进行去噪。HCGCN [36] 也遵循 LATTICE 的一般流程，构建物品语义图和用户-物品交互图。此外，它还引入了协同聚类技术，在同一子图中学习用户和物品表示，从而有助于捕捉多种用户行为模式。相应的协同聚类损失和项目聚类损失用于增强用户-项目偏好反馈并调整模态的重要性。A2BM2GL [3]中也提到了利用特征构建同质图的类似结构。它利用多模态内容的相似性构建特征图，以学习语义表示。除语义表示学习模块外，它还构建了协同表示学习模块来学习协同表示，并设置了一个反瓶颈模块，利用注意力机制学习短程和长程节点的重要性权重。它将加权回归损失更新为自适应推荐损失，从而有助于捕捉用户偏好的不同强度。LUDP [24] 指出，目前仅通过多模态特征丰富物品表示的模型，无法很好地学习用户偏好。因此，它设计了相同的物品-物品模态相似度图，以捕捉潜在的物品关系；同时还设计了用户偏好图，以捕捉用户对不同模态的偏好。

采样子图融合。为每个节点采样一个子图。根据模态对节点进行分组。HHFAN [4] 构建了一个包含用户、物品和多模态信息的异构图。通过图进行随机游走，并根据节点类型采样邻居。在类型内特征聚合网络中，它利用全连接层将不同类型的节点向量转换到同一空间，并利用长短期记忆网络聚合同一组节点中同类型节点的嵌入。随后，在类型间特征聚合网络中，针对不同类型的节点使用属性感知自注意力和邻居感知注意力，以生成能够更好地传达结构和语义信息的最终表示。HCGCN [36] 为用户-物品交互图采样子图，将用户和物品划分为多个组，从而有助于捕捉每组对应的各种用户行为模式。HCGCN通过采样子图来捕捉用户行为模式，InvRL [10]也采用了类似的方法。这表明，伪

相关性问题在于，通用表示关注语义含义，却没有关注影响用户偏好的因素。为减轻虚假相关的影响，该方法通过采样子图学习不变掩码，将多模态表示划分为不变表示和变化表示。基于UltraGCN [35], ，利用不变表示能够消除虚假相关。

使用多模态信息细化图结构。推荐系统广泛利用用户-物品交互图来进行表示学习和信息传播。然而，其中存在一些假阳性边，会影响推荐性能。GRCN [58] 致力于识别并删除噪声边，以细化用户-物品交互图的结构。在图细化层中，采用邻居路由机制来细化原型网络。用户偏好与原型之间的匹配结果显示其为真阳性边的概率。根据匹配分数，这些边将被剔除。随后，利用图卷积层和预测层，基于细化后的用户-物品图进行推荐。LATTICE [67]构建了一个具有模态感知能力的图结构学习层，从多模态信息中学习物品图结构，并融合多模态图。通过使用图卷积，一个物品可以沿着学习到的图结构从其邻居获取有用的高阶亲和关系。

4.2.8 自监督学习。

多模态推荐旨在为每种模态学习合适的表示。它们主要基于历史交互和多模态侧信息进行监督学习。自监督学习技术也有助于探索潜在关系。SLMRec [48] 通过将自监督学习融入基于图神经网络的推荐模型来学习这些关系。该论文将 ID 视为一种特殊模态，与视觉、文本和音频模态并列。并且，它以与 [53, 59, 67]相同的方式构建模态特定图。随后，在模态特定图上使用图神经网络技术，得到模态表示。该方法将每种模态视为独立特征，并利用特征丢弃（FD）、特征掩码（FM）以及特征细粒度与粗粒度空间（FAC）这几种自监督学习任务生成物品的不同视图，然后使用对比学习来学习潜在的训练信号。

BM3 [80] 通过将自监督学习方法引入多模态推荐，解决了计算成本和噪声监督信号的问题。首先，它利用简单的丢弃技术从物品和用户表示中生成不同视图。然后优化三个多模态目标，以获得更好的表示。图重建损失通过重构用户-物品交互图，在给定用户的情况下最大化对物品的预测。模态间特征对齐损失和模态内特征掩码损失均用于对齐模态特征。该模型取得了更好的性能，同时消耗的计算资源也更少。

MMGCL [65] 将基于图的自监督学习方法引入微视频推荐，并提出了一种前沿的负采样方法，用以展现不同模态之间的关系。该模型采用了两种数据增强技术：模态边丢弃和模态掩码。另一种技术是扰动某个负样本的其中一种模态，使正样本与负样本之间仅有一种模态不同。这些方法结合起来，揭示了多种模态之间的关系，并确保每种模态都能发挥有意义的作用。

4.2.9 预训练。

预训练模型已被证明在自然语言和图像领域十分有用。图像和文本等侧信息已被用于提升推荐系统的效果。PMGT [32] 开发了一种预训练模型，同时考虑多模态特征和物品

关系，以学习物品表示。该模型采用注意力机制对多模态特征进行加权聚合，并根据共同的用户行为构建图结构，以表示物品之间的关系。两项预训练任务是图结构重建和掩码节点特征重建。该模型开发了一种小批量上下文邻居采样算法，用于采样负样本和正样本，以处理大规模图。完成预训练后，我们可以获得节点表示，并将其用于初始化下游任务的物品嵌入。本文中的推荐任务以神经协同过滤作为基础模型。与其他多模态推荐模型相比，训练神经协同过滤并根据用户-物品交互微调物品嵌入，取得了最高的准确率。

同质图（如物品图和用户图）已被证明有助于学习表示并提升推荐性能。MMCPR [33] 在同质图和异质图上进行预训练，旨在学习能够捕捉用户偏好和物品特征的更优用户表示和物品表示。为学习用户表示和物品表示，基于用户与物品的共同交互关系构建用户图和物品图。对于用户表示，模型考虑两种模态：评论文本和用户图。模型使用 LightGCN 作为图编码器，以获取用户图中的节点表示。对于评论文本，模型采用模态内聚合来获得聚合后的评论表示，并通过注意力机制进行模态间聚合，从而聚合两种模态的表示。对于物品表示，模型考虑三种模态：文本、图像和物品图。模型使用 CLIP 提取特征。由于图像和文本描述的是同一物品，但侧重点不同，因此采用自监督对比式模态间对齐（CIMA）任务，通过最大化二者之间的余弦相似度来对齐这两种嵌入。模态内聚合与用户侧相同，通过注意力机制聚合这三种表示，以获得最终的物品嵌入。基于对比式模态间对齐损失和用户-物品相似度损失进行预训练，以获得预训练嵌入。随后，选择基于 MixGCF 的 LightGCN 作为微调基础模型。

多模态特征反映了物品的共通知识和特定知识。近期的简单融合技术将共性和独特特征结合为单一向量表示。PAMD [14] 旨在理解对象的共性和独特特征。它使用解耦编码器将模态特征划分为共性表示和特定表示。随后，通过对比学习进行跨模态对齐，将模态共性表示映射到原始表示。接着，它通过衡量原始表示与特定模态表示之间的差异，将二者的迁移表示用于优化。经过预训练阶段后，每个物品都能获得四种表示。它利用注意力机制融合所有嵌入以获得最终表示，并采用 BPR 作为下游基础模型。

5 模态融合

对于多模态推荐系统而言，模态融合是一项重要的研究课题。找到一种可解释且具有互补性的融合多模态表示，能够带来显著提升。从技术角度来说，多模态融合是将不同模态的信息整合为单一表示，以用于预测和分类等不同任务。多模态融合有三大优势。首先，利用多模态信息能够捕捉各模态中的通用模式，并基于相同模式获得更鲁棒的表示。其次，多模态能够捕捉单一模态中不可见的互补信息。最后，多模态融合能够处理模态缺失问题，即使缺少一种模态，系统仍可运行。根据

融合步骤	融合方法	模型
早期融合	拼接	VBPR[15], UVCAN[30], MAML[27]
	基于注意力	VECF[7], PMGT[32], LATTICE[67], 自由-DOM [73],HCGCN[36], InvRL[10],LUDP[24]
中间融合	拼接	MMGCL[65]
	基于注意力	MMRec[61], MKGAT[46], DMRL[26]
	专家乘积	MVGAE[64]
	自我融合	EgoGCN[5]
晚期融合	拼接	MMGCN[59], GRCN[58], SLMRec[48], JRL[69], EliMRec[31]
	逐元素求和	MMGCN[59], SLMRec[48], A2BM2GL[3]
	基于注意力	DualGNN[53], MGAT[49], PAMD[14], MMCPR[33],DRAGON [71]

表 4. 模态融合

流程中操作发生的时机，大致可分为早期融合、晚期融合和中间融合 [1]。

- 早期融合也称为数据级融合，指将模态嵌入融合为单一特征表示，然后输入模型。但这种方法无法捕捉多模态的互补信息，并且可能引入数据冗余。早期融合方法通常与一些特征提取方法（如主成分分析和自编码器）结合使用，以解决这一问题。

- 晚期融合也称为决策级融合，它基于各模态的预测结果或分数进行融合。常用方法包括最大值融合、平均值融合、基于贝叶斯规则的方法和集成学习。这类方法能够学习互补信息，且各模态彼此独立，因此不会导致错误叠加。
- 中间融合是在获取每种模态的高维嵌入后，对模态信息进行融合，然后利用中间层进行融合。这种融合方式取决于模型设计。

根据不同的融合方法，我们可以将其分为逐元素求和、基于注意力的融合和拼接。逐元素求和和拼接只是对模态表示进行相加或拼接，形式化表示如下：

$$\boldsymbol{u} = \boldsymbol{u_v} + \boldsymbol{u_t} + \boldsymbol{u_a} \qquad \boldsymbol{u} = \boldsymbol{u_v} \parallel \boldsymbol{u_t} \parallel \boldsymbol{u_a} \tag{1}$$

基于注意力的方法用于捕捉各个模态的重要性差异，也可应用于逐元素求和等基础融合方法：

$$\boldsymbol{u} = \sum_{m \in \{v,t,a\}} \alpha_m \boldsymbol{u_m} \tag{2}$$

有些模型设计了自己的模态融合方式，例如 MVGAE [64] 采用专家乘积模块来区分具有不同不确定性水平的模态，从而解决拼接带来的问题。而 IMRec [63] 并不融合模态信息，而是将印象图像作为一个整体进行编码。我们根据模型的模态融合方法对其进行分类，如表 4 所示。

6 测量与优化

6.1 评估指标

推荐系统常用的评估指标包括准确率、精确率、召回率、F1 分数、命中率、归一化折损累计增益（NDCG）和平均精度均值（MAP）。我们将真实标签视为正例或负例；在预测标签中，会出现真正例（TP）、真负例（TN）、假正例（FP）和假负例（FN）。准确率表示正确预测的物品数占物品总数的比例：

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

在实际情况下，数据集中的正样本和负样本并不平衡。通常，正样本少于 10%，这意味着即使将所有物品都预测为负例，准确率也可能高于 0.9。我们的目标是衡量预测正例的能力。召回率和精确率分别表示正确预测的正样本数占预测为正的样本数以及真实正样本数的比例。F1 分数综合了召回率和精确率的结果：

$$Recall = \frac{TP}{TP + FN} \qquad Precision = \frac{TP}{TP + FP} \qquad F1 - score = \frac{2 * precision * recall}{precision + recall} \tag{4}$$

顾名思义，命中率表示预测的 top-k 列表中的命中物品数占测试集物品总数的比例。如果用户实际与我们推荐的 top-k 物品之一发生交互，就视为命中。

$$HR = \frac{NumberOfHits@K}{Test} \tag{5}$$

精确率和召回率不考虑排序。平均精度（AP）用于评估面向单个用户的排序敏感型推荐结果，而 MAP 是所有用户的 $AP@K$ 的平均值：

$$AP@K = \frac{1}{|K|} \sum_{k=1}^{K} \frac{p_k \cdot rel(k)}{k} \qquad MAP = \frac{1}{|U|} \sum_{u \in U} AP_u \tag{6}$$

归一化折损累计增益（NDCG）也是衡量排序质量的指标。累积增益（CG）表示推荐列表中的相关性得分；DCG 则弥补了 CG 仅利用相关性得分、却未考虑排名位置这一不足。iDCG 是理想排序的推荐列表的 DCG。NDCG 通过 DCG 除以 iDCG 计算，形式如下：

$$DCG = \sum_{i=1}^{K} \frac{2^{relevance_i} - 1}{log_2(i+1)} \qquad NDCG = \frac{DCG}{iDCG} \tag{7}$$

6.2 目标函数

典型的损失函数包括均方损失（MSE）、成对损失、交叉熵损失和贝叶斯个性化排序（BPR）损失。在 top-k 推荐中，通常会使用最后两种损失函数

方法	模型
BPR	VBPR[15], ACF[6], MMGCN[59],DualGNN[53],GRCN[58], MGAT[49], GraphCAR[62], PAMD[14], MMGCL[65], MKGAT[46], MVGAE[64],DVBPR[19],LATTICE[67],FREEDOM [73], DMRL[26],HCGCN[36],EgoGCN[5],EliMRec[31], InvRL[10], LUDP[24], DRAGON [71]
成对损失	MAML[27]
均方误差损失	VMCF[38]
交叉熵损失	UVCAN[30], MMRec[61],MMCPR[33], IMRec[63]
对比损失	PAMD[14], MMCPR[33]。MMGCL[65],BM3[80], SLMRec[48]
重构损失	PMGT[32], BM3[80]
自设计损失	A2BM2GL[3]

表5. 损失函数

，而较少使用 MSE 和成对损失。MSE 始终用于评分预测。

$$MSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(h(\boldsymbol{x}^{i}) - \boldsymbol{y}^{i})^{2}} \tag{8}$$

基本的成对损失通常不会被使用，但存在一些改进版本。加权近似排序成对损失（WARP） [60] 通过考虑排序改进了基本的成对损失：对排名较低的正向物品施加的惩罚，远重于对排名靠前的正向物品施加的惩罚。BPR [41] 将 Top-k任务视为排序问题，鼓励给定用户的正向物品排在负向物品之前；而交叉熵损失则将 Top-k任务视为分类问题，并尽可能提高目标用户 u 与目标物品 i 的相似度。BPR损失更适用于 Top-k推荐，大多数模型都采用这种损失进行优化，例如 [7, 53, 59, 67]。

$$\mathcal{L}_{bpr} = \sum_{(u,i,i')\in\mathcal{R}} -ln\mu(\boldsymbol{u}^{T}\boldsymbol{i} - \boldsymbol{u}^{T}\boldsymbol{i}') \tag{9}$$

$$\mathcal{L}_{CE} = -\sum_{(u,i)\in\mathcal{R}} y_{ui}\cdot log(\boldsymbol{u}^{T}\boldsymbol{i}) + (1-y_{ui})\cdot log(1-\boldsymbol{u}^{T}\boldsymbol{i}) \tag{10}$$

随着预训练和微调被引入多模态推荐，预训练任务损失已被用于获得更好的表示结果。[32] 利用边重构损失和节点特征重构损失进行优化。最近，随着自监督学习方法应用于推荐领域，对比损失也被引入其中。对比损失旨在提高相似物品之间的相似度，同时使物品表示与其他物品有所区别。而 InfoNCE损失是将该问题视为多分类的变体 [48, 65]。BM3 [80] 无需进行负采样

表6. 数据集统计信息。

数据集	# 用户	# 物品	# 交互	稀疏度
Baby	19,445	7,050	160,792	99.8827%
体育	35,598	18,357	296,337	99.9547%
FoodRec	61,668	21,874	1,654,456	99.8774%
Elec	192,403	63,001	1,689,188	99.9861%

并设计不同视图，通过计算视图之间的相似度来获得对比损失。

$$\mathcal{L}_{CL} = \frac{\boldsymbol{u}^T \boldsymbol{i}}{\|\boldsymbol{u}\|_2 \|\boldsymbol{i}\|_2} \tag{11}$$

SLMRec [48] 为自监督任务和推荐任务都设计了对比损失。

$$\mathcal{L}_{InfoNCE} = -log\frac{exp(\boldsymbol{u}^T \boldsymbol{i}/\tau)}{\sum_{i' \in N_u} exp(\boldsymbol{u}^T \boldsymbol{i}'/\tau)} \tag{12}$$

如表5所示，我们根据模型所采用的目标函数对模型进行了分类。

7 数据集与实验

7.1 实验数据集

一些常用数据集规模庞大，包含大量用户、物品和交互，例如快手和 TikTok；还有一些数据集不对公众开放。本文使用公开且最常用的亚马逊数据集，以及从某个网站收集的一个食品数据集作为实验数据集。我们在四个主要类别——婴儿、体育、电子产品和食品上进行实验。这四个数据集的统计信息汇总于表6。与以往工作相同 [58, 59, 67, 80], , 我们对所有亚马逊数据集采用 5-core 设置，对食品数据集采用 10核设置，并按 8:1:1 的比例将数据集划分为训练集、验证集和测试集。需要说明的是，划分时会根据该比例为每位用户随机选择训练样本和测试样本。在接下来的实验中，我们将讨论不同的训练集、测试集和验证集划分方式，例如基于用户时间或全局时间进行划分。

亚马逊数据集的元数据包含文本描述和图像的 URL，可将其用作文本和视觉模态。对于收集到的食品数据集，我们将食谱和食材用作文本模态，将食品图像用作视觉模态。我们使用预训练的 sentence-transformers [40] 提取维度为384的文本特征，并按照 [67] 使用已发布的4096维视觉特征。获得文本和视觉嵌入后，我们将用户评分视为正向交互，然后基于评论评分构建用户-物品图。

7.2 基线模型

我们为读者提供了一个框架，便于使用基线模型运行自己的数据集，或基于该框架实现自己的模型。该框架由 PyTorch 实现，输入格式简单，包括文本嵌入、视觉嵌入和交互图。除 ADDVAE [50] 外，以下模型均已在我们的框架中实现，并通过实验比较其推荐性能。请注意，BPR 和 LightGCN 不使用多模态信息，而仅使用历史用户-物品交互。由于 ADDVAE 模型使用 TF-IDF 提取方法生成的不同格式的文本特征作为输入，因此未在我们的框架中实现。我们单独实现了 ADDVAE，并使用相同的训练集-测试集划分以及不同提取方法生成的原始文本特征运行该模型

。在此，我们仅将它们作为多模态推荐性能的基线模型。

• BPR [41] 利用历史用户-物品交互，将用户和物品表示建模为潜因子，并根据这些表示之间的相似度预测用户偏好。该模型引入了最常用的损失函数——BPR损失。

• LightGCN [16] 是一种基于图的模型，通过简化图卷积网络（GCN） [21] 来提升推荐性能和效率。

• VBPR [15] 是首个在推荐中考虑视觉特征的模型，它将视觉嵌入整合到物品表示中。为保证公平比较，我们在学习表示时将视觉和文本特征拼接为多模态特征。此处使用BPR损失来学习用户偏好。

• MMGCN [59] 通过为每种模态构建模态特定图来学习模态特定偏好，并利用这些用户-物品图学习各模态的表示。随后聚合所有模态特定表示，得到最终的用户和物品嵌入。

• DuaGNN [53] 引入了一种新的用户-用户共现图，该图利用从模态特定图中学习到的表示，并融合邻居的表示。

• GRCN [58] 采用与之前基于 GCN 的模型生成的相同用户-物品图。它识别图中的假阳性边并将其移除，随后通过在优化后的图中执行聚合和信息传播来生成新的表示。

• LATTICE [67] 为每种模态引入一个新的物品-物品图，并通过聚合所有模态得到潜在物品图。物品-物品图和用户-物品图均利用图卷积操作，最终获得用户和物品的表示。

• BM3 [80] 引入了一种新的自监督学习方法，以帮助降低计算资源的开销，并解决监督信号错误的问题。它利用简单的丢弃技术生成对比视图，并使用三个对比损失函数来优化最终表示。

• SLMRec [48] 在基于图神经网络的模型中采用自监督学习技术，以学习潜在关系。自监督学习任务补充监督任务，从数据本身中挖掘隐藏信号。这里采用了 FD、FM 和 FAC 自监督数据增强操作，随后使用对比损失进行优化。

• ADDVAE [50] 利用 MacridVAE 网络，通过互信息最大化从用户-物品交互和用户的文本内容中获取两种解耦表示。随后利用注意力机制对齐这些解耦表示。

• FREEDOM [73] 利用为每种模态构建的物品-物品图，其构建方式与 LATTICE 相同，但在训练前冻结这些图；此外，还引入度敏感边剪枝技术，对用户-物品交互图进行去噪。

7.3 实验结果

本节首先分析所有已实现算法在四个数据集上的表现（第 7.3.1 节）。随后讨论不同类型的数据划分策略对性能的影响（第 7.3.2 节）。最后，利用不同的模态信息评估多模态表示学习的有效性（第 7.3.3 节）。我们采用广泛使用的指标评估偏好排序性能：Recall@k 和 NDCG@k。默认情况下，我们将 k 设为 $= 10,20,50$，并报告测试集中所有用户的平均指标。

7.3.1 使用不同模型进行推荐性能比较。
表 7 报告了性能比较结果，我们可以从中观察到以下几点，并了解如何选择合适的模型：

- 总体而言，内容感知方法的性能优于基于协同过滤的方法（例如，BPR、LightGCN），这表明多模态特征始终能为物品提供更丰富的信息，并帮助推荐模型达到更高的准确率。尽管基于 BPR 构建的 VBPR 的推荐准确率已经有了很大提升，但 MMGCN [59] 的性能始终不如未利用多模态信息的 LightGCN。这是因为正如 LightGCN [16]所指出的，不合理的设计会使大量噪声影响性能。MMGCN 是较早为多模态推荐设计的模型，当时 LightGCN 技术尚未发布；而 DualGNN [53], 采用了 LightGCN 的优势来学习模态用户-物品二部图，因此取得了更好的性能。

  此外，现有的内容感知方法在很大程度上依赖于多模态特征的表征，这会导致它们在不同数据集上的性能波动。

- 具体而言，FREEDOM [73] 在婴儿和体育数据集上的表现最佳，但在食品和电子数据集上的排名较低；后两者是包含约170万次交互的大型图。FREEDOM在小型图上表现良好，但不适用于大型图。除FoodRec数据集外，ADDVAE的排名始终较高，这可能是因为FoodRec数据集从网站中提取而来，该数据集的文本特征存在问题，相关信息可能难以得到利用。对于食品数据集，这些模型之间的差异很小。SLMRec的性能始终取决于数据集，并且该模型需要较长时间进行参数搜索。LATTICE和BM3在实验数据集上的表现较为稳定，其性能排名高于基于协同过滤的方法，以及MMGCN和VBPR等较早开发的多模态模型。

- 图的规模和资源消耗也会影响算法的选择。LATTICE会构建物品图，模型运行需要大量资源，因此尽管它在大多数数据集上表现良好，却无法应用于大型图。FREEDOM [73] 在小型数据集上能够取得较高的准确率，但应用于大型图时表现并非最佳。它比LATTICE消耗更少的计算资源，却能取得更好的性能。SLMRec可用于不同规模的图，但参数搜索过程耗时较长。通过分析多个数据集的结果，找出不敏感的参数并减少这些参数，可以简化搜索过程。BM3在各种规模的数据集上都表现良好，并且只需消耗较少资源就能达到较高的性能排名。

7.3.2 使用不同数据划分方法的推荐性能比较。
离线评估基于历史物品评分或隐式物品反馈。由于这种方法依赖用户-物品交互，且模型都是基于监督信号进行学习，因此我们需要将交互划分为训练集、验证集和测试集。我们采用了三种主要的划分策略来比较性能：

- 随机划分：顾名思义，这种划分策略会为每位用户随机选择训练集与测试集的分界点。此前，留一最后物品策略仅将最后一次交互作为测试数据，而随机划分则按照比例对交互进行划分。随机划分策略的缺点是，除非

表 7. 不同模型在召回率和归一化折损累计增益方面的性能，最佳结果以**粗体**显示，次佳结果带下划线。

数据集	模型	Recall@10	Recall@20	召回率@50	NDCG@10	NDCG@20	归一化折损累计增益@50
Baby	BPR	0.0357	0.0575	0.1054	0.0192	0.0249	0.0345
	LightGCN	0.0479	0.0754	0.1333	0.0257	0.0328	0.0445
	VBPR	0.0423	0.0663	0.1212	0.0223	0.0284	0.0396
	MMGCN	0.0378	0.0615	0.1100	0.0200	0.0261	0.0359
	DualGNN	0.0448	0.0716	0.1288	0.0240	0.0309	0.0424
	GRCN	0.0539	0.0833	0.1464	0.0288	0.0363	0.0490
	LATTICE	0.0547	0.0850	0.1477	0.0292	0.0370	0.0497
	BM3	0.0564	0.0883	0.1477	0.0301	0.0383	0.0502
	SLMRec	0.0529	0.0775	0.1252	0.0290	0.0353	0.0450
	ADDVAE	0.0598	0.091	0.1508	0.0323	0.0404	0.0525
	FREEDOM	**0.0627**	**0.0992**	**0.1655**	**0.0330**	**0.0424**	**0.0558**
体育	BPR	0.0432	0.0653	0.1083	0.0241	0.0298	0.0385
	LightGCN	0.0569	0.0864	0.1414	0.0311	0.0387	0.0498
	VBPR	0.0558	0.0856	0.1391	0.0307	0.0384	0.0492
	MMGCN	0.0370	0.0605	0.1078	0.0193	0.0254	0.0350
	DualGNN	0.0568	0.0859	0.1392	0.0310	0.0385	0.0493
	GRCN	0.0598	0.0915	0.1509	0.0332	0.0414	0.0535
	LATTICE	0.0620	0.0953	0.1561	0.0335	0.0421	0.0544
	BM3	0.0656	0.0980	0.1581	0.0355	0.0438	0.0561
	SLMRec	0.0663	0.0990	0.1543	0.0365	0.0450	0.0562
	ADDVAE	0.0709	0.1035	0.1663	**0.0389**	0.0473	0.0600
	FREEDOM	**0.0717**	**0.1089**	**0.1768**	0.0385	**0.0481**	**0.0618**
FoodRec	BPR	0.0303	0.0511	0.0948	0.0188	0.0250	0.0356
	LightGCN	0.0331	0.0546	0.1003	0.0210	0.0274	0.0386
	VBPR	0.0306	0.0516	0.0972	0.0191	0.0254	0.0365
	MMGCN	0.0307	0.0510	0.0943	0.0192	0.0253	0.0359
	DualGNN	0.0338	0.0559	0.1027	0.0214	0.0280	0.0394
	GRCN	**0.0356**	**0.0578**	**0.1063**	**0.0226**	**0.0295**	**0.0411**
	LATTICE	0.0336	0.0560	0.1012	0.0211	0.0277	0.0388
	BM3	0.0334	0.0553	0.0994	0.0208	0.0274	0.0381
	SLMRec	0.0323	0.0515	0.0907	0.0208	0.0266	0.0362
	ADDVAE	0.0309	0.0508	0.093	0.0186	0.0247	0.035
	FREEDOM	0.0333	0.0556	0.1009	0.0212	0.0279	0.0389
Elec	BPR	0.0235	0.0367	0.0621	0.0127	0.0161	0.0212
	LightGCN	0.0363	0.0540	0.0879	0.0204	0.0250	0.0318
	VBPR	0.0293	0.0458	0.0778	0.0159	0.0202	0.0267
	MMGCN	0.0213	0.0343	0.0610	0.0112	0.0146	0.0200
	DualGNN	0.0365	0.0542	0.0875	0.0206	0.0252	0.0319
	GRCN	0.0389	0.0590	0.0970	0.0216	0.0268	0.0345
	LATTICE	-	-	-	-	-	-
	BM3	0.0437	0.0648	0.1021	0.0247	0.0302	0.0378
	SLMRec	0.0443	0.0651	0.1038	0.0249	0.0303	0.0382
	ADDVAE	**0.0451**	**0.0665**	**0.1066**	**0.0253**	**0.0308**	**0.0390**
	FREEDOM	0.0396	0.0601	0.0998	0.0220	0.0273	0.0353

表 8. 不同数据划分下的性能（召回率和召回率和归一化折损累计增益方面的表现 回率和归一化折损累计增益

数据集	模型	随机	用户时间	全局时间	随机	用户时间	全局时间
		Recall@10			Recall@20		
体育	MMGCN	0.0384	0.0266	0.0140	0.0611	0.0446	0.0245
	BPR	0.0444	0.0322	0.0152	0.0663	0.0509	0.0258
	VBPR	0.0563	0.0385	0.0176	0.0851	0.0620	0.0298
	DualGNN	0.0576	0.0403	0.0181	0.0859	0.0611	0.0297
	GRCN	0.0604	0.0418	0.0167	0.0915	0.0666	0.0286
	LightGCN	0.0568	0.0405	0.0205	0.0863	0.0663	0.0336
	LATTICE	0.0641	0.0450	0.0207	0.0964	0.0699	0.0337
	BM3	0.0646	0.0447	0.0213	0.0955	0.0724	0.0336
	SLMRec	0.0651	0.0470	0.0220	0.0985	0.0733	0.0350
	FREEDOM	0.0708	0.0490	0.0226	0.1080	0.0782	0.0372
数据集	模型	NDCG@10			NDCG@20		
体育	MMGCN	0.0202	0.0134	0.0091	0.0261	0.0180	0.0125
	BPR	0.0245	0.0169	0.0102	0.0302	0.0218	0.0135
	VBPR	0.0304	0.0204	0.0115	0.0378	0.0265	0.0153
	DualGNN	0.0321	0.0214	0.0118	0.0394	0.0268	0.0155
	GRCN	0.0332	0.0219	0.0101	0.0412	0.0282	0.0138
	LightGCN	0.0315	0.0220	0.0139	0.0391	0.0286	0.0180
	LATTICE	0.0351	0.0238	0.0138	0.0434	0.0302	0.0177
	BM3	0.0356	0.0237	0.0144	0.0436	0.0308	0.0182
	SLMRec	0.0364	0.0253	0.0148	0.0450	0.0321	0.0189
	FREEDOM	0.0388	0.0255	0.0151	0.0485	0.0330	0.0197

作者公布了数据划分方式，但如果不考虑时间，这并不是一个现实的场景。

- 用户时间划分：
  时间划分策略根据交互时间戳，按比例（例如，训练集:验证集:测试集=8:1:1）划分历史交互。它将用户产生的最后一定比例的交互划为测试集。VAECF [25], SVAE [42] 和 NGCF [56] 均在此策略下进行评估。尽管该策略考虑了时间戳，但它仍然不是一个现实的场景，因为它仍然是在用户产生的所有交互中划分训练/测试集，却没有考虑全局时间。

- 全局时间划分：
  全局时间划分策略根据划分比例，为所有用户固定一个共同的时间点。最后一个时间点之后的交互被划分为测试集。此外，全局时间边界之后发生交互的用户必须出现在训练集中，这符合最贴近现实且最严格的设置。这种策略的局限在于，由于删除了训练集中不存在的用户，用户数量会减少。

为了比较采用不同数据划分方法时的推荐性能，我们在体育数据集上针对不同的数据划分策略进行了一些简单实验。表 8 展示了以 Recall@k 和 NDCG@k 衡量的性能比较结果，其中 k=10,20。表 9 展示了基于 Recall@20 和 NDCG@20 对模型进行的性能排名。

我们希望了解数据划分策略会对最先进模型产生怎样的影响。因此，我们在体育数据集上采用三种数据划分策略，对 10 个最先进模型的表现进行比较：随机划分、用户时间划分和全局时间划分。如表 8

所示，即使使用相同的数据集和评估指标，不同的数据划分策略也会导致模型得到不同的性能数值。使用不同数据划分策略设计的不同模型，其性能难以进行比较，因此这些模型之间不可比。

表 9 列出了 10 种最先进模型在不同划分策略下的排名。各行按照模型在随机划分策略下的表现排序，上下箭头表示相较于随机划分时排名位置的变化。可以看到，不同划分策略下，各模型的排名有所变化。例如，在全局时间划分下，GRCN [58] 的排名会比随机划分时低四位，而 BPR、VBPR 和 LightGCN 的表现则超过了它。在我们的实验中，表现最好的模型和最弱模型的排名不会受到影响，只有中等表现模型的排名会因划分策略不同而变化。此外，基于时间的划分会导致推荐系统出现冷启动问题，而人们认为可以利用多模态信息解决这一问题。从表现排名来看，在所有划分策略下，基于内容的模型整体表现都优于基于协同过滤的模型。不过，我们还发现了一种相反的趋势：排名变化大致呈现一定规律——基于协同过滤的模型在时间评估中的排名更有可能提升，而部分基于内容的模型（例如，DualGNN 和 GRCN）排名则会下降。我们推测，原因可能是部分基于内容的模型未能正确利用多模态特征；这些特征不仅无助于解决冷启动问题，在应对冷启动问题时反而会成为噪声信息。

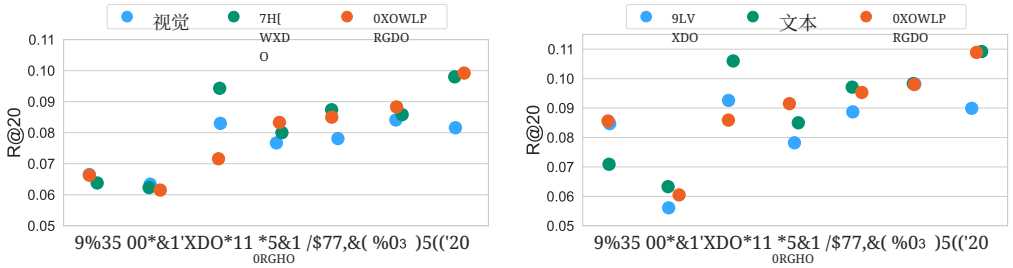
大多数多模态推荐系统采用随机划分策略，较少的模型采用用户时间划分和全局时间划分。然而，时间划分设置更接近现实场景，因为推荐系统无法获取未来数据作为训练数据。稳健的推荐系统应在所有划分策略下都表现稳定。模型的行为取决于表示的学习方式以及实例的选择方式。因此，我们可以得出结论：数据划分策略是影响模型性能的关键因素，针对不同情况的模型需要选择合适的划分策略。不过，我们只能从简单实验中获得一些初步启发，而且仅在体育数据集上运行了 10 个模型。如果你对不同划分方式的影响感兴趣，可以使用更多模型和数据集进行研究，以发现潜在问题。

表 9. 体育数据集的性能排名

模型	体育，NDCG@20		
	随机划分	用户时间划分	全局时间划分
MMGCN	10	10	10
BPR	9	9	8↑1
VBPR	8	8	7↑1
LightGCN	7	5↑2	4↑3
DualGNN	6	7↓1	6
GRCN	5	6↓1	9↓4
LATTICE	4	4	5↓1
BM3	3	3	3
SLMRec	2	2	2
FREEDOM	1	1	1

模型	体育，Recall@20		
	随机划分	用户时间划分	全局时间划分
MMGCN	10	10	10
BPR	9	9	9
VBPR	8	7↑1	6↑2
DualGNN	7	8↓1	7
LightGCN	6	6	5↑1
GRCN	5	5	8↓3
BM3	4	3↑1	4
LATTICE	3	4↓1	3
SLMRec	2	2	2
FREEDOM	1	1	1

7.3.3 不同模态下的推荐性能对比。



(a) 婴儿 (b) 体育

图4. 不同模态的性能比较

多模态信息已被引入推荐系统，有助于缓解冷启动问题。一些模型将模态信息作为辅助信息来丰富物品表示，另一些模型则将其与交互信息结合，以增强用户表示或捕捉用户偏好。

我们关注模态信息如何促进推荐，以及哪种模态的贡献更大。我们希望探究每种模态的影响，并为研究人员选择所使用的模态信息提供指导。

我们通过输入单模态信息，评估多模态信息如何影响最先进模型的推荐性能，并比较同时使用两种模态与仅使用单一模态时的性能。表10展示了不同模态的实验结果。粗体表示输入其他模态时特定模型所取得的最佳性能。我们根据 Recall@20 绘制了图4，以展示其他模态的概况和趋势，直观呈现不同模态对不同模型的影响差异。橙色圆圈表示多模态的性能，绿色圆圈表示文本模态的性能，蓝色圆圈则表示视觉模态的性能。

我们首先在婴儿和体育数据集上进行实验。结果表明，只有加入多模态特征的 GRCN [58] 在所有数据集上的表现优于仅使用文本或视觉特征的情况。对于其他模型，使用单模态信息可能比融合模态取得更好的性能，这表明，仅通过求和、加权求和和拼接来融合模态表示，并不能有效学习物品表示。这种简单的融合会破坏从各个模态中学到的模态特定信息和模态偏好。尤其是对于 DualGNN [53], , 单独使用文本和视觉模态所取得的性能远优于融合模态，这表明所采用的融合方法会影响学到的表示质量。

此外，我们发现，这些表示与文本模态的关系更为密切。如图4所示，文本模态的表现优于或接近多模态表现。我们在思考数据集是否会影响结果，以及我们测试的这些数据集是否更侧重于文本模态。于是，我们尝试了视觉特征更为重要的服装数据集，但结果仍然表明文本模态的重要性。

对于融合问题，我们需要找出合适的方法，在保留模态特定信息的同时融合多模态表示，并捕捉单一模态所不包含的互补信息。理想结果是多模态表示优于单模态表示。对于文本模态的表现优于或接近多模态的情况，首要原因是融合方法

会影响表示的质量；第二个原因可能是，高维视觉嵌入经过多层感知机层映射到潜在空间时会压缩数据，并丢失某些模态特定属性。此外，我们发现，基线模型的一些消融研究也包含多模态与单模态的性能比较 [53, 59] ，且多模态表现优于单模态，但数据集尚未发布。因此，最后一个原因可能是不同数据集侧重于不同模态。微视频数据集更侧重于视觉模态，其中包含更多视觉模态信息，有助于学习有意义的表示。

8 挑战与未来研究方向

在推荐系统中利用多模态信息与实际应用密切相关。近年来，多模态内容已广泛用于电子商务和社交平台上的产品或相关物品推荐。

在本综述中，我们回顾并归类了多模态推荐领域中近期最常用的模型。具体而言，我们根据模型所采用的技术对多模态推荐系统进行了分类，这些技术能够利用有助于推荐的多模态信息。除此之外，我们还介绍了多模态推荐的流程，并列举了特征提取、模态融合和评估中常用的技术。此外，我们构建了一个通用框架，帮助新入行的研究人员轻松运行基线模型，并了解推荐模型的设计方式。读者也可以利用该框架更轻松地开发自己的模型，并将其与基线模型进行比较。

在综述和研究过程中，我们发现了一些挑战以及可能的未来研究方向。

8.1 如何有效融合多模态信息?

既然模态信息已被证明有助于推荐，那么如何设计模型以更好地利用多模态便成为一项研究挑战。正确利用多模态特征能够提升性能，但错误利用则会引入噪声信息。正如上述实验所示，利用模态信息能够提升推荐系统的性能；但对于某些模型，单一模态也能取得良好的表现，甚至优于联合利用多模态信息。我们认为，对于这些模型而言，问题可能出在多模态融合上。每种模态都可能捕捉物品的不同方面，因此我们应设法将它们融合，同时保留模态特定信息，从而学习到包含互补信息、而单模态表示无法涵盖的多模态表示。如果模型能够高效融合多模态特征，其推荐准确率理应高于仅利用单一模态时的准确率。

此外，现实世界中普遍存在模态缺失问题。然而，一些模型假设训练和推理期间所有模态信息都可用，因此在遇到不完整或缺失的模态时便无法发挥作用。LRMM [52] 通过利用生成模型重构模态特定嵌入并补全缺失模态，从而缓解模态缺失问题和冷启动问题。

未来研究方向应聚焦于如何高效利用多模态特征：（1）找出能够捕捉单一模态无法提供的互补信息的高效模态融合方法。（2）如何解决模态缺失问题并重构有意义的表示。

表10. 不同特征的召回率表现

数据集	模型	Both	文本	视觉	Both	文本	视觉
		Recall@10			Recall@20		
Baby	VBPR	0.0423	0.0400	**0.0428**	0.0663	0.0638	**0.0665**
	MMGCN	0.0378	0.0365	**0.0384**	0.0615	0.0623	**0.0634**
	DualGNN	0.0448	**0.0612**	0.0511	0.0716	**0.0943**	0.0830
	GRCN	**0.0539**	0.0517	0.0488	**0.0833**	0.0800	0.0767
	LATTICE	**0.0547**	0.0546	0.0492	0.0850	**0.0874**	0.0781
	BM3	0.0564	**0.0571**	0.0544	**0.0883**	0.0858	0.0841
	ADDVAE	-	0.0598	-	-	0.091	-
	FREEDOM	**0.0627**	0.0622	0.0501	**0.0992**	0.0980	0.0816
体育	VBPR	**0.0558**	0.0472	**0.0558**	**0.0856**	0.0709	0.0847
	MMGCN	0.0370	**0.0401**	0.0358	0.0605	**0.0633**	0.0561
	DualGNN	0.0568	**0.0697**	0.0615	0.0859	**0.1060**	0.0926
	GRCN	**0.0598**	0.0545	0.0495	**0.0915**	0.0850	0.0782
	LATTICE	0.0620	**0.0625**	0.0572	0.0953	**0.0971**	0.0887
	BM3	**0.0656**	0.0652	0.0649	0.0980	**0.0983**	0.0982
	ADDVAE	-	0.0709	-	-	0.1035	-
	FREEDOM	0.0717	**0.0725**	0.0589	0.1089	**0.1092**	0.0899
服装	VBPR	0.0281	**0.0328**	0.0283	0.0415	**0.0506**	0.0416
	MMGCN	**0.0218**	0.0211	0.0203	**0.0345**	0.0342	0.0328
	DualGNN	0.0454	**0.0524**	0.0420	0.0683	**0.0798**	0.0636
	GRCN	**0.0424**	0.0392	0.0351	**0.0662**	0.0612	0.0542
	LATTICE	0.0492	**0.0521**	0.0408	0.0733	**0.0749**	0.0614
	BM3	0.0422	0.0413	**0.0423**	0.0621	0.0613	**0.0626**
	FREEDOM	**0.0629**	0.0625	0.0420	**0.0941**	0.0921	0.0640
数据集	模型	NDCG@10			NDCG@20		
Baby	VBPR	0.0223	0.0213	**0.0230**	0.0284	0.0275	**0.0290**
	MMGCN	**0.0200**	0.0189	0.0197	**0.0261**	0.0256	0.0261
	DualGNN	0.0240	**0.0331**	0.0278	0.0309	**0.0417**	0.0360
	GRCN	**0.0288**	0.0269	0.0258	**0.0363**	0.0342	0.0330
	LATTICE	**0.0292**	0.0287	0.0265	0.0370	**0.0371**	0.0339
	BM3	0.0301	**0.0303**	0.0290	**0.0383**	0.0377	0.0367
	ADDVAE	-	0.0323	-	-	0.0404	-
	FREEDOM	**0.0330**	0.0325	0.0274	**0.0424**	0.0417	0.0355
体育	VBPR	**0.0307**	0.0256	0.0306	**0.0384**	0.0317	0.0380
	MMGCN	0.0193	**0.0212**	0.0193	0.0254	**0.0272**	0.0245
	DualGNN	0.0310	**0.0379**	0.0335	0.0385	**0.0473**	0.0415
	GRCN	**0.0332**	0.0289	0.0258	**0.0414**	0.0368	0.0332
	LATTICE	0.0335	**0.0336**	0.0312	0.0421	**0.0425**	0.0393
	BM3	**0.0355**	0.0352	0.0352	**0.0438**	0.0437	**0.0438**
	ADDVAE	-	0.0389	-	-	0.0473	-
	FREEDOM	0.0385	**0.0387**	0.0318	0.0481	**0.0482**	0.0397
服装	VBPR	0.0158	**0.0173**	0.0159	0.0192	**0.0218**	0.0193
	MMGCN	**0.0110**	0.0109	0.0107	**0.0142**	0.0142	0.0139
	DualGNN	0.0241	**0.0281**	0.0229	0.0299	**0.0351**	0.0283
	GRCN	**0.0223**	0.0204	0.0178	**0.0283**	0.0260	0.0226
	LATTICE	0.0268	**0.0290**	0.0221	0.0330	**0.0348**	0.0273
	BM3	0.0231	0.0225	**0.0235**	0.0281	0.0276	**0.0286**
	FREEDOM	**0.0341**	0.0336	0.0227	**0.0420**	0.0411	0.0283

8.2 如何标准化数据划分策略和所用的通用数据集 ?

在训练/测试集划分过程中采用了多种数据划分策略。正如前述实验所示，即使使用相同的数据集和评估指标，模型采用不同的数据划分策略也会得到不同的性能数值；不同的划分策略也会影响模型性能的排名。尽管最常用的是随机划分，但基于时间的划分更贴近实际场景。稳健的多模态推荐模型应当在随机划分或时间划分下都表现良好。

用于多模态推荐的数据集尚未标准化，而且综述论文中使用的大多数数据集都未公开，例如快手、TikTok 和大众点评。即使研究使用相同的公开数据集，数据集的设置也不统一。有些人会采用 5-core 设置，而另一些人可能采用 10-core，这使得数据集的预处理各不相同。正如我们之前所展示的，特征提取技术也会影响最终学到的表示，但每篇论文使用的技术都不一样。

未来的工作中，我们应找出最佳的预处理方法，并采用合适的特征提取技术和数据集划分策略，以规范实验中的训练/测试集。

8.3 评估指标

推荐模型的评估是一项重要的研究课题。推荐系统不仅应考虑准确率，还应考虑其他推荐质量，例如推荐列表的多样性和独特物品的出现情况，这些因素可能会显著影响推荐系统的整体质量 [18]。此外， [11] 指出，通用推荐的评估方法未必能完全适用于多模态推荐。评估效果时不应只考虑准确率，还应考虑用户体验和公平性。

8.4 研究与应用差距

学术研究人员和工业界都在研究多模态推荐。双方所使用的数据、可用的信息，甚至应用场景都各不相同。此外，数据量不断增长，也给现实场景带来了巨大挑战。可扩展性和时间复杂度是工业界选择模型时需要重点考虑的因素。正如前文所述，工业研究的设定更加贴近现实。学术研究中的一些设定应加以规范，并增强其实用性，以便开发出适用于工业场景的模型。理想的推荐系统应具备鲁棒性，并能轻松应用于现实场景。未来还应进一步研究如何高效地进行推荐：（1）面对大型数据集时，如何平衡模型复杂度与可扩展性；（2）如何提升高维张量和多模态信息的计算效率。

8.5 多模态序列推荐

序列推荐系统不同于采用协同过滤和基于内容的过滤的推荐系统，它旨在理解并建模用户随时间变化的序列行为。多模态信息可能会显著影响用户的偏好，但现有的大多数序列推荐模型都忽略了这类有用信息。MML [37] 将物品的多模态侧信息融入其中，以改进并稳定元学习过程，并帮助解决冷启动问题。因此，在序列推荐系统中利用多模态信息将是未来研究的一个重要方向。

8.6 跨域推荐

跨域推荐系统利用来自其他域或平台的行为信息，提升目标域的性能。常用技术依赖显式重叠数据（例如，共同用户和物品）在不同域之间迁移数据。近期，一些研究学习用户和物品的通用表示，并将其应用于跨域推荐。UniSRec [17]利用文本信息学习通用物品表示，无需共同用户和物品，因而可应用于不同域。未来，利用多模态信息或许有助于模型学习通用表示。

9 结论

在本综述中，我们全面回顾了多模态推荐系统。我们为 MMRec 提出了清晰的流程，并列出了各步骤中常用的技术。我们根据用于整理和聚类现有文献的学习方法对模型进行了分类。我们还在四个常用数据集上进行了实验，以评估模型的性能，并提供了一个通用框架，便于用户运行模型和开发自己的模型。此外，我们详细阐述了一些亟待解决的挑战和有前景的未来研究方向。每年都会涌现大量新技术，我们希望本综述能为研究人员提供有益且详尽的概述，所提供的框架能让模型运行更加便捷高效，并推动未来的进步。

参考文献

[1] Tadas Baltrušaitis、Chaitanya Ahuja 和 Louis-Philippe Morency。2018。多模态机器学习：综述与分类体系。《*IEEE*模式分析与机器智能汇刊》41, 2 (2018), 423–443。[2] Desheng Cai、Shengsheng Qian、Quan Fang、Jun Hu、Wenkui Ding 和 Changsheng Xu。2022。用于个性化微视频推荐的异构图对比学习网络。*IEEE*多媒体汇刊 (2022)。[3] Desheng Cai、Shengsheng Qian、Quan Fang、Jun Hu 和 Changsheng Xu。2022。用于个性化微视频推荐的自适应反瓶颈多模态图学习网络。载于第*30*届*ACM*国际多媒体会议论文集。581–590。[4] Desheng Cai、Shengsheng Qian、Quan Fang 和 Changsheng Xu。2021。用于个性化微视频推荐的异构层次特征聚合网络。*IEEE*多媒体汇刊 24 (2021), 805–818。[5] Feiyu Chen、Junjie Wang、Yinwei Wei、Hai-Tao Zheng 和 Jie Shao。2022。打破隔离：通过边级调制实现多媒体推荐的多模态图融合。载于第*30*届*ACM*国际多媒体会议论文集。385–394。[6] Jingyuan Chen、Hanwang Zhang、Xiangnan He、Liqiang Nie、Wei Liu 和 Tat-Seng Chua。2017。注意力协同过滤：通过物品级和组件级注意力实现多媒体推荐。载于第*40*届国际*ACM SIGIR*信息检索研究与发展会议论文集。335–344。[7] Xu Chen、Hanxiong Chen、Hongteng Xu、Yongfeng Zhang、Yixin Cao、Zheng Qin 和 Hongyuan Zha。2019。基于多模态注意力网络并提供视觉解释的个性化时尚推荐：迈向视觉可解释推荐。载于第*42*届国际*ACM SIGIR*信息检索研究与发展会议论文集。765–774。[8] Yashar Deldjoo、Markus Schedl、Paolo Cremonesi 和 Gabriella Pasi。2020。利用多媒体内容的推荐系统。《*ACM*计算调查》（*CSUR*）53, 5 (2020), 1–38。[9] Shiyao Ding、Donghui Lin 和 Xin Zhou。2021。用于边缘计算中依赖任务分配的图卷积强化学习。载于*2021*年*IEEE*智能体国际会议（*ICA*）。IEEE，25–30。[10] Xiaoyu Du、Zike Wu、Fuli Feng、Xiangnan He 和 Jinhui Tang。2022。多媒体推荐中的不变表示学习。载于第*30*届*ACM*国际多媒体会议论文集。619–628。[11] Mouzhi Ge 和 Fabio Persia。2018。多媒体推荐系统评估：实用指南。载于*2018*年*IEEE*第*12*届语义计算国际会议（*ICSC*）。IEEE，294–297。[12] Qipeng Guo、Xipeng Qiu、Xiangyang Xue 和 Zheng Zhang。2021。通过图神经网络实现语法引导文本生成。《中国科学：信息科学》64, 5 (2021), 1–10。

[13] Will Hamilton、Zhitao Ying 和 Jure Leskovec。2017。大型图上的归纳式表示学习。《神经信息处理系统进展》30 (2017)。[14] Tengyue Han、Pengfei Wang、Shaozhang Niu 和 Chenliang Li。2022。模态匹配模态：为推荐预训练模态解耦的物品表示。载于《2022年ACM Web会议论文集》。2058–2066。[15] Ruining He 和 Julian McAuley。2016。VBPR：基于隐式反馈的视觉贝叶斯个性化排序。载于《AAAI人工智能会议论文集》，卷30。[16] Xiangnan He、Kuan Deng、Xiang Wang、Yan Li、Yongdong Zhang 和 Meng Wang。2020。LightGCN：简化并增强用于推荐的图卷积网络。载于《第43届国际ACM SIGIR信息检索研究与发展会议论文集》。639–648。[17] Yupeng Hou、Shanlei Mu、Wayne Xin Zhao、Yaliang Li、Bolin Ding 和 Ji-Rong Wen。2022。迈向推荐系统的通用序列表示学习。载于《第28届ACM SIGKDD知识发现与数据挖掘会议论文集》。585–593。[18] Marius Kaminskas 和 Derek Bridge。2016。多样性、惊喜性、新颖性与覆盖率：推荐系统中超越准确率目标的综述与实证分析。《ACM交互式智能系统汇刊》（TiiS）7, 1 (2016), 1–42。[19] Wang-Cheng Kang、Chen Fang、Zhaowen Wang 和 Julian McAuley。2017。利用生成式图像模型实现视觉感知时尚推荐与设计。载于《2017年IEEE国际数据挖掘会议》（ICDM）。IEEE，207–216。[20] Donghyun Kim、Chanyoung Park、Jinoh Oh、Sungyoung Lee 和 Hwanjo Yu。2016。用于文档上下文感知推荐的卷积矩阵分解。载于《第10届ACM推荐系统会议论文集》。233–240。[21] Thomas N Kipf 和 Max Welling。2016。基于图卷积网络的半监督分类。arXiv预印本 arXiv:1609.02907 (2016)。[22] Claire Lauer。2009。厘清术语：“多模态”与“多媒体”在学术界和公共领域中的含义。《计算机与写作》26, 4 (2009), 225–239。[23] Xiangnan He、Le Wu、Xiang Wang、Kun Zhang 和 Meng Wang。2022。准确率导向的神经推荐综述：从协同过滤到信息丰富型推荐。《IEEE知识与数据工程汇刊》(2022), 1–1。[24] Fei Lei、Zhongqi Cao、Yuning Yang、Yibo Ding 和 Cong Zhang。2022。学习用户在多模态推荐系统中的深层偏好。《ACM多媒体计算、通信与应用汇刊》(2022)。[25] Dawen Liang、Rahul G Krishnan、Matthew D Hoffman 和 Tony Jebara。2018。用于协同过滤的变分自编码器。载于《2018年万维网会议论文集》。689–698。[26] Fan Liu、Zhiyong Cheng、Huilin Chen、Anan Liu、Liqiang Nie 和 Mohan Kankanhalli。2022。用于推荐的解耦多模态表示学习。arXiv预印本 arXiv:2203.05406 (2022)。[27] Fan Liu、Zhiyong Cheng、Changchang Sun、Yinglong Wang、Liqiang Nie 和 Mohan Kankanhalli。2019。通过多模态注意力度量学习建模用户多样化偏好。载于《第27届ACM国际多媒体会议论文集》。1526–1534。[28] Kang Liu、Feng Xue、Dan Guo、Le Wu、Shujie Li 和 Richang Hong。 [未注明日期。]MEGCF：用于个性化推荐的多模态实体图协同过滤。《ACM信息系统汇刊》（TOIS）([未注明日期])。[29] Qidong Liu、Enguang Yao、Chaoyue Liu、Xin Zhou、Yafei Li 和 Mingliang Xu。2022。M2GCN：用于建模多药联用副作用的多模态图卷积网络。《应用智能》(2022), 1–12。[30] Shang Liu、Zhenzhong Chen、Hongyi Liu 和 Xinghai Hu。2019。用于个性化微视频推荐的用户-视频协同注意力网络。载于《万维网会议论文集》。3020–3026。[31] Xiaohao Liu、Zhulin Tao、Jiahong Shao、Lifang Yang 和 Xianglin Huang。2022。EliMRec：消除多媒体推荐中的单模态偏差。载于《第30届ACM国际多媒体会议论文集》。687–695。[32] Yong Liu、Susen Yang、Chenyi Lei、Guoxin Wang、Haihong Tang、Juyong Zhang、Aixin Sun 和 Chunyan Miao。2021。利用多模态侧信息预训练图变换器以实现推荐。载于《第29届ACM国际多媒体会议论文集》。2853–2861。[33] Zhuang Liu、Yunpu Ma、Matthias Schubert、Yuanxin Ouyang 和 Zhang Xiong。2022。用于推荐的多模态对比预训练。载于《2022年国际多媒体检索会议论文集》。99–108。[34] Linyuan Lu、Matus Medo、Chi Ho Yeung、Yi-Cheng Zhang、Zi-Ke Zhang 和 Tao Zhou。2012。推荐系统。《物理学报告》519, 1 (2012), 1–49。[35] Kelong Mao、Jieming Zhu、Xi Xiao、Biao Lu、Zhaowei Wang 和 Xiuqiang He。2021。UltraGCN：面向推荐的图卷积网络极简化。载于《第30届ACM信息与知识管理国际会议论文集》。1253–1262。

[36] Zongshen Mu、Yueting Zhuang、Jie Tan、Jun Xiao 和 Siliang Tang。2022。学习多媒体推荐的混合行为模式。载于第30届ACM国际多媒体会议论文集。376–384。[37] Xingyu Pan、Yushuo Chen、Changxin Tian、Zihan Lin、Jinpeng Wang、He Hu 和 Wayne Xin Zhao。2022。面向冷启动序列推荐的多模态元学习。载于第31届ACM信息与知识管理国际会议论文集。3421–3430。[38] Chanyoung Park、Donghyun Kim、Jinoh Oh 和 Hwanjo Yu。2017。“也看过”的产品是否有助于用户评分预测？载于第26届万维网国际会议论文集。1113–1122。[39] Amir Hossein Nabizadeh Rafsanjani、Naomie Salim、Atae Rezaei Aghdam 和 Karamollah Bagheri Fard。2013。推荐系统：综述。《国际计算工程研究杂志》3卷5期（2013），47–52。[40] Nils Reimers 和 Iryna Gurevych。2019。Sentence-BERT：使用孪生BERT网络的句子嵌入。arXiv预印本 arXiv:1908.10084（2019）。[41] Steffen Rendle、Christoph Freudenthaler、Zeno Gantner 和 Lars Schmidt-Thieme。2012。BPR：基于隐式反馈的贝叶斯个性化排序。arXiv预印本 arXiv:1205.2618（2012）。[42] Noveen Sachdeva、Giuseppe Manco、Ettore Ritacco 和 Vikram Pudi。2019。用于协同过滤的序列变分自编码器。载于第十二届ACM网络搜索与数据挖掘国际会议论文集。600–608。[43] Lipi Shah、Hetal Gaudani 和 Prem Balani。2016。推荐系统综述。《国际计算机应用杂志》137卷7期（2016），43–49。[44] Yue Shi、Martha Larson 和 Alan Hanjalic。2014。超越用户-物品矩阵的协同过滤：现状综述与未来挑战。《ACM计算调查》（CSUR）47卷1期（2014），1–45。[45] Umut Sulubacak、Ozan Caglayan、Stig-Arne Grönroos、Aku Rouhe、Desmond Elliott、Lucia Specia 和 Jörg Tiedemann。2019。通过视觉信息与语音实现多模态机器翻译。arXiv预印本 arXiv:1911.12798（2019）。[46] Rui Sun、Xuezhi Cao、Yan Zhao、Junchen Wan、Kun Zhou、Fuzheng Zhang、Zhongyuan Wang 和 Kai Zheng。2020。用于推荐系统的多模态知识图谱。载于第29届ACM信息与知识管理国际会议论文集。1405–1414。[47] Jinhui Tang、Xiaoyu Du、Xiangnan He、Fajie Yuan、Qi Tian 和 Tat-Seng Chua。2019。面向鲁棒多媒体推荐系统的对抗训练。《IEEE知识与数据工程汇刊》32卷5期（2019），855–867。[48] Zhulin Tao、Xiaohao Liu、Yewei Xia、Xiang Wang、Lifang Yang、Xianglin Huang 和 Tat-Seng Chua。2022。用于多媒体推荐的自监督学习。《IEEE多媒体汇刊》（2022）。[49] Zhulin Tao、Yinwei Wei、Xiang Wang、Xiangnan He、Xianglin Huang 和 Tat-Seng Chua。2020。MGAT：用于推荐的多模态图注意力网络。《信息处理与管理》57卷5期（2020），102277。[50] Nhu-Thuat Tran 和 Hady W Lauw。2022。对齐评分与文本内容中的双重解耦用户表示。载于第28届ACM SIGKDD知识发现与数据挖掘会议论文集。1798–1806。[51] Petar Veličković、Guillem Cucurull、Arantxa Casanova、Adriana Romero、Pietro Lio 和 Yoshua Bengio。2017。图注意力网络。arXiv预印本 arXiv:1710.10903（2017）。[52] Cheng Wang、Mathias Niepert 和 Hui Li。2018。LRMM：学习在缺失模态条件下进行推荐。arXiv预印本 arXiv:1808.06791（2018）。[53] Qifan Wang、Yinwei Wei、Jianhua Yin、Jianlong Wu、Xuemeng Song、Liqiang Nie 和 Min Zhang。2021。DualGNN：用于多媒体推荐的双图神经网络。《IEEE多媒体汇刊》（2021）。[54] Shoujin Wang、Liang Hu、Yan Wang、Xiangnan He、Quan Z Sheng、Mehmet A Orgun、Longbing Cao、Francesco Ricci 和 Philip S Yu。2021。基于图学习的推荐系统：综述。arXiv预印本 arXiv:2105.06339（2021）。[55] Xin Wang、Hong Chen 和 Wenwu Zhu。2021。用于推荐的多模态解耦表示。载于2021年IEEE多媒体与博览会国际会议（ICME）。IEEE，1–6。[56] Xiang Wang、Xiangnan He、Meng Wang、Fuli Feng 和 Tat-Seng Chua。2019。神经图协同过滤。载于第42届国际ACM SIGIR信息检索研究与发展会议论文集。165–174。[57] Yinwei Wei、Xiang Wang、Xiangnan He、Liqiang Nie、Yong Rui 和 Tat-Seng Chua。2021。用于多媒体推荐的分层用户意图图网络。《IEEE多媒体汇刊》（2021）。[58] Yinwei Wei、Xiang Wang、Liqiang Nie、Xiangnan He 和 Tat-Seng Chua。2020。基于隐式反馈的多媒体推荐图精炼卷积网络。载于第28届ACM国际多媒体会议论文集。3541–3549。[59] Yinwei Wei、Xiang Wang、Liqiang Nie、Xiangnan He、Richang Hong 和 Tat-Seng Chua。2019。MMGCN：用于微视频个性化推荐的多模态图卷积网络。载于第27届ACM国际多媒体会议论文集。1437–1445。[60] Jason Weston、Samy Bengio 和 Nicolas Usunier。2010。大规模图像标注：利用联合词图嵌入进行学习排序。《机器学习》81卷1期（2010），21–35。

[61] Chuhan Wu、Fangzhao Wu、Tao Qi 和 Yongfeng Huang。2021。MM-Rec：多模态新闻推荐。*arXiv*预印本 *arXiv:2104.07407*（2021）。[62] Qidi Xu、Fumin Shen、Li Liu 和 Heng Tao Shen。2018。GraphCAR：利用图自编码器实现内容感知多媒体推荐。载于第*41*届国际*ACM SIGIR*信息检索研究与发展会议。981–984。[63] Jiahao Xun、Shengyu Zhang、Zhou Zhao、Jieming Zhu、Qi Zhang、Jingjie Li、Xiuqiang He、Xiaofei He、Tat-Seng Chua 和 Fei Wu。2021。我们为何点击：视觉印象感知新闻推荐。载于第*29*届*ACM*国际多媒体会议论文集。3881–3890。[64] Jing Yi 和 Zhenzhong Chen。2021。用于推荐系统的多模态变分图自编码器。*IEEE*多媒体汇刊（2021）。[65] Zixuan Yi、Xi Wang、Iadh Ounis 和 Craig Macdonald。2022。用于微视频推荐的多模态图对比学习。载于第*45*届国际*ACM SIGIR*信息检索研究与发展会议论文集。1807–1811。[66] Fuzheng Zhang、Nicholas Jing Yuan、Defu Lian、Xing Xie 和 Wei-Ying Ma。2016。面向推荐系统的协同知识库嵌入。载于第*22*届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集。353–362。[67] Jinghao Zhang、Yanqiao Zhu、Qiang Liu、Shu Wu、Shuhui Wang 和 Liang Wang。2021。挖掘多媒体推荐的潜在结构。载于第*29*届*ACM*国际多媒体会议论文集。ACM。[68] Lingzi Zhang、Yong Liu、Xin Zhou、Chunyan Miao、Guoxin Wang 和 Haihong Tang。2022。基于扩散的图对比学习：利用隐式反馈进行推荐。载于高级应用数据库系统：第*27*届国际会议，*DASFAA 2022*，线上活动，*2022*年*4*月*11*日至*14*日，论文集，第二部分。Springer，232–247。[69] Yongfeng Zhang、Qingyao Ai、Xu Chen 和 W Bruce Croft。2017。利用异构信息源进行Top-N推荐的联合表示学习。载于*2017*年*ACM*信息与知识管理会议论文集。1449–1458。[70] Lei Zheng、Vahid Noroozi 和 Philip S Yu。2017。利用评论对用户和物品进行联合深度建模，以实现推荐。载于第十届*ACM*网络搜索与数据挖掘国际会议论文集。425–434。[71] Hongyu Zhou、Xin Zhou、Lingzi Zhang 和 Zhiqi Shen。2023。利用同质图增强多模态推荐中的二元关系。*arXiv*预印本 *arXiv:2301.12097*（2023）。https://arxiv.org/abs/2301.12097[72] Jie Zhou、Ganqu Cui、Shengding Hu、Zhengyan Zhang、Cheng Yang、Zhiyuan Liu、Lifeng Wang、Changcheng Li 和 Maosong Sun。2020。图神经网络：方法与应用综述。*AI Open* 1（2020），57–81。[73] Xin Zhou。2022。两个图的故事：为多模态推荐冻结和去噪图结构。*arXiv*预印本 *arXiv:2211.06924*（2022）。[74] Xin Zhou。2023。MMRec：简化多模态推荐。*arXiv*预印本 *arXiv:2302.03497*（2023）。[75] Xin Zhou、Donghui Lin 和 Toru Ishida。2016。在评分稀缺的情况下评估网络服务的声誉。载于*2016*年*IEEE*服务计算国际会议（*SCC*）。IEEE，211–218。[76] Xin Zhou、Donghui Lin、Yong Liu 和 Chunyan Miao。2022。用于推荐的层细化图卷积网络。*arXiv*预印本 *arXiv:2207.11088*（2022）。[77] Xin Zhou、Shigeo Matsubara、Yuan Liu 和 Qidong Liu。2022。评分系统中的贿赂：博弈论视角。载于知识发现与数据挖掘进展：第*26*届亚太知识发现与数据挖掘会议，*PAKDD 2022*，中国成都，*2022*年*5*月*16*日至*19*日，论文集，第三部分。Springer，67–78。[78] Xin Zhou、Aixin Sun、Yong Liu、Jie Zhang 和 Chunyan Miao。2021。Selfcf：用于自监督协同过滤的简易框架。*arXiv*预印本 *arXiv:2107.03019*（2021）。[79] Xin Zhou、Jinglong Wang、Yong Liu、Xingyu Wu、Zhiqi Shen 和 Cyril Leung。2022。用于配送时间估计的归纳式图Transformer。*arXiv*预印本 *arXiv:2211.02863*（2022）。[80] Xin Zhou、Hongyu Zhou、Yong Liu、Zhiwei Zeng、Chunyan Miao、Pengwei Wang、Yuan You 和 Feijun Jiang。2022。多模态推荐的自举潜在表示。*arXiv*预印本 *arXiv:2207.05969*（2022）。