arXiv:2302.03883v2 [cs.IR] 4 Sep 2024

多模态推荐系统：综述

Qidong Liu*，西安交通大学与香港城市大学，中国Jiaxi Hu*，香港城市大学，中国Yutian Xiao*，香港城市大学，中国Xiangyu Zhao†，香港城市大学，中国Jingtong Gao，香港城市大学，中国Wanyu Wang，香港城市大学，中国Qing Li，香港理工大学，中国Jiliang Tang，密歇根州立大学，美国

推荐系统（RS）一直是在线服务不可或缺的工具。推荐系统利用各种深度学习技术，根据标识符和属性信息对用户偏好进行建模。随着多媒体服务的出现，例如短视频、新闻等，在进行推荐的同时理解这些内容变得至关重要。此外，多模态特征还有助于缓解推荐系统中的数据稀疏性问题。因此，**多模态推荐系统**（MRS）近年来受到了学术界和工业界的广泛关注。本文将主要从技术视角出发，对多模态推荐系统模型进行全面综述。首先，我们总结多模态推荐系统的一般流程和主要挑战。随后，我们根据四个类别介绍现有的多模态推荐系统模型，即：**模态编码器**、**特征交互**、**特征增强** 和**模型优化**。此外，为方便希望研究这一领域的读者，我们还总结了数据集和代码资源。最后，我们讨论了多模态推荐系统的一些有前景的未来方向，并对本文进行了总结。若要获取所综述论文的更多详细信息，例如实现代码，我们开源了代码仓库[1]。

CCS 概念：• **信息系统** → **推荐系统**；**多媒体和多模态检索**。

附加关键词和短语：推荐系统；多模态；多媒体

ACM参考格式：
Qidong Liu、Jiaxi Hu、Yutian Xiao、Xiangyu Zhao、Jingtong Gao、Wanyu Wang、Qing Li 和 Jiliang Tang。2024。多模态推荐系统：综述。1，1（2024年9月），17页。https://doi.org/XXXXXXX。XXXXXXX

*这些作者对本研究作出了同等贡献。
†通讯作者。[1]
https://github.com/Applied-Machine-Learning-Lab/Awesome-Multimodal-Recommender-Systems

作者联系信息：Qidong Liu，liuqidong@stu.xjtu.edu.cn，西安交通大学和香港城市大学，中国西安；Jiaxi Hu，jiaxihu2-c@my.cityu.edu.hk，香港城市大学，中国香港；Yutian Xiao，yutiaxiao2-c@my.cityu.edu.hk，香港城市大学，中国香港；Xiangyu Zhao，xianzhao@cityu.edu.hk，香港城市大学，中国香港；Jingtong Gao，jt.g@my.cityu.edu.hk，香港城市大学，中国香港；Wanyu Wang，wanyuwang4-c@my.cityu.edu.hk，香港城市大学，中国香港；Qing Li，qing-prof.li@polyu.edu.hk，香港理工大学，中国香港；Jiliang Tang，tangjili@msu.edu，密歇根州立大学，美国东兰辛。

允许免费为个人或课堂使用而制作本作品全部或部分内容的数字或纸质副本，前提是这些副本不得用于营利或商业用途，也不得为此进行分发，并且副本必须带有本声明和首页的完整引文。对于本作品中由 ACM 以外的其他方拥有版权的部分，必须遵守其版权规定。注明出处后允许摘录。以其他方式复制、再版、发布到服务器或分发到列表，均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。© 2024 ACM。
ACM XXXX-XXXX/2024/9-ART https://doi.org/XXXXXXX.XXXXXXX

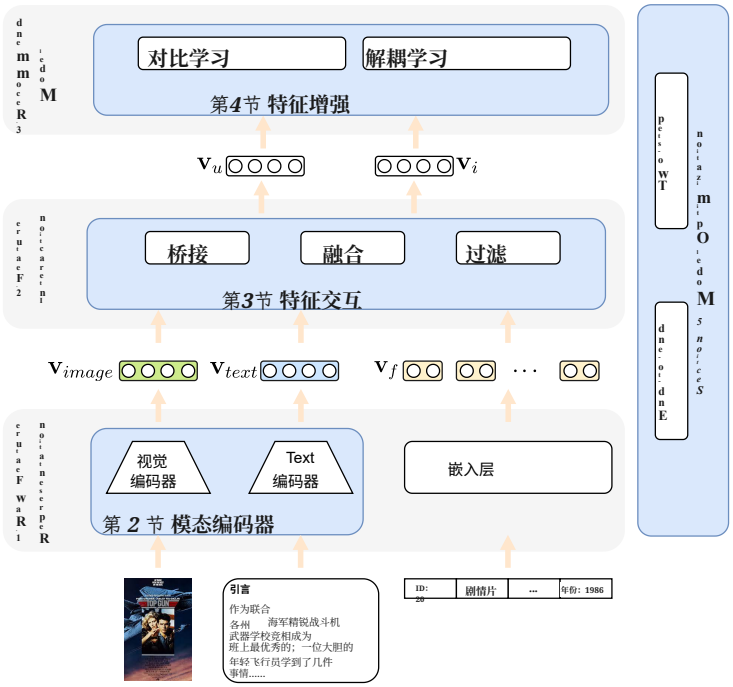

图1. 多模态推荐系统的一般流程。

1 引言

随着互联网的发展，许多多媒体在线服务不断涌现，例如时尚推荐 [9] 和等。这些多媒体应用为推动推荐系统朝着理解推荐物品的方向发展提供了契机，这具有重要意义。一方面，理解能力可以帮助推荐系统利用物品丰富的多模态信息，从而缓解数据稀疏性问题 [9]。另一方面，它能够帮助推荐系统从语义层面更深入地了解用户偏好。考虑到多媒体服务的普及，多模态推荐系统（MRS）有望成为未来推荐系统的一种通用模式。因此，近期越来越多的研究聚焦于多模态推荐系统，亟需对相关研究进行综述和分类。

一般而言，推荐系统主要关注协同信息或侧信息，侧信息指物品的标识符（简称 ID）和表格特征，例如类别和出版年份。相比之下，在多模态推荐系统中，图像、音频和文本等多模态特征发挥着至关重要的作用。为简明起见，我们将多模态推荐系统定义为：面向具有多模态特征的物品的推荐系统。在接下来的小节中，我们将介绍一般流程和分类体系，以使本综述更易于阅读。

1.1 多模态推荐系统的流程

根据多模态推荐系统的输入物品，我们总结了其统一流程，如图 1 所示。该流程包含三个步骤：**原始特征表示**、**特征交互** 和 **推-荐**。下面以电影推荐为例，说明一般流程：

原始特征表示。每部电影都具有两类特征：表格特征通过数值或分类（如类别或年份）描述其重要属性；多模态特征则通过多种表示模态（如海报图像和文本介绍）描绘电影。处理表格特征时，通用推荐系统通常采用嵌入层，将稀疏离散特征转换为稠密向量 [18]。具体而言，嵌入层将每个特征字段视为一个离散集合，并将其映射为固定长度的表示。然而，多模态特征的格式通常各不相同，无法共用这样的嵌入层。因此，通常会将多模态特征输入不同的模态编码器，以提取全面的表示。模态编码器通常采用其他领域中使用的通用架构，例如用于图像的 ViT [12] 和用于文本的 BERT [11]。随后，我们可以得到每个物品的表格特征和多模态特征（即 图像和文本）的表示，分别记为 $\mathbf{v}f$、$\mathbf{v}^{image}$ 和 $\mathbf{v}_{text}$。

特征交互。我们获取了每个物品不同模态的表示，但它们处于不同的语义空间中。此外，不同用户对模态也有不同的偏好 [65]。因此，在这一过程中，多模态推荐系统旨在融合并交互多模态表示$\mathbf{v}f$，$\mathbf{v}^{image}$ 以及 $\mathbf{v}_{text}$，从而获得统一的物品表示和用户表示，这些表示通常用于生成推荐列表 [13, 38]。

推荐。经过第二个过程，我们获得了用户和物品的表示，分别记为 $\mathbf{v}_u$ 和 $\mathbf{v}_i$。通用推荐模型吸收这两种表示，并为不同物品给出推荐概率 [22]。然而，数据稀疏性问题始终会降低推荐性能。因此，许多研究 [34, 40]提出通过引入多模态信息来增强这些表示。

1.2 分类体系

多模态特征为缓解数据稀疏性问题带来了可能。然而，由于多模态特征的复杂性和异质性，多模态推荐系统的每个过程以及整个流程都面临一些挑战。这些挑战阻碍了多模态特征发挥作用，甚至会对推荐系统产生不利影响。因此，为了充分利用多模态信息，现有研究主要致力于应对其中一个或多个挑战。下面将按照多模态推荐系统的流程列出主要挑战及相应的解决方案。

• 挑战***1***：对于原始多模态输入，例如图像或文本，如何从复杂的模态特征中获取表示。嵌入层已被广泛应用于通用推荐系统中，以从原始特征中学习表示。然而，从复杂的图像或文本中获取表示并不容易。例如，显然，从电影的原始海报或简介中提取有用信息既困难又至关重要。得益于计算机视觉和自然语言处理领域的进步，多模态推荐系统可以借用许多编码器来获取表示，例如用于图像的 ViT [12]。在本文中，我们将所有处理多模态特征的编码器称为**模态编码器**（见第2节）。

• 挑战 ***2***：对于特征交互过程，如何融合不同语义空间中的模态特征，并获取每种模态的多种偏好。多模态特征的异质性使多模态推荐系统中的物品表示和用户偏好学习变得困难。例如，电影海报和引言的特征处于截然不同的表示空间，这阻碍了对用户偏好的推断。为应对这一挑战，许多研究通过提取用户、物品与模态之间的关系来设计模型。这些研究被归类为**特征交互** 技术（见第3节）。

表1. 模态编码器类别

模态	类别	相关工作
视觉编码器	CNN	[6], [36], [20], [76] [45], [61], [7], [5], [34], [31], [30], [14], [35], [37], [33], [84], [26]
	ResNet	[38], [47], [46], [39], [58], [52], [66], [65], [59], [16], [17], [1]
	Transformer	[8], [13]
文本编码器	Word2vec	[36], [2], [52], [61], [65], [59], [14]
	RNN	[68], [30]
	CNN	[62], [5]
	句子转换器	[20], [74], [73], [45], [81], [84], [26], [63], [64]
	Bert	[38], [47], [32], [46], [39], [58], [8], [66], [41], [34], [13], [17], [1], [33], [29]
其他模态编码器	已发布特征	[53], [74], [73], [71], [65], [59], [70], [72], [51], [83], [63], [64]

• 挑战***3***：对于推荐流程，如何在数据稀疏条件下为推荐模型获取全面的表示。尽管多模态特征丰富了物品的信息，但由于推荐系统中的交互记录数量较少，稀疏性问题依然存在。例如，Movielens-1M2这一典型电影推荐数据集的稀疏率超过95%。与通用推荐系统相比，多模态特征可进一步用于增强用户和物品的表示。我们将这类研究称为**特征增强**（见第4节）。

• 挑战***4***：对于多模态推荐系统的整个流程，如何优化轻量级推荐模型和参数化模态编码器。以电影推荐模型 [29]为例，其电影简介文本编码器的参数规模约为1.1亿，而基础推荐系统的参数规模不到1000万。为了解决优化问题，一些多模态推荐系统相关工作提出了新颖的技术，并将其归入**模型优化** （第5节）。

基于上述四项挑战，我们根据相应的技术解决方案组织本文的其余内容，即 **模态编码器、特征交互、特征增强** 和**模型优化**。据我们所知，本综述与现有的两篇多模态推荐系统综述完全不同。一篇综述 [10] 按照实际应用中的不同模态组织研究内容。另一篇最新的综述 [80] 更加关注推荐系统本身，而忽略了多模态推荐系统的特点。相比之下，我们的综述围绕各种类型的技术组织相关描述，尤其是多模态技术，这有助于读者更好地理解整体的多模态推荐系统架构。此外，我们还尝试收集所有近期工作，帮助读者了解该领域的最新进展。

2 模态编码器

物品的多模态特征对于构建更具体的用户兴趣以及增强模型可解释性至关重要。在推荐任务中，物品ID信息通常由通过可训练嵌入表获得的稠密向量表示。然而，物品的多模态特征需要相应的编码器，以获得稠密特征表示，从而提取更全面的信息。本节除介绍已发布的特征向量外，还简要介绍三类常用多模态特征的编码器，即即，**视觉**、**文本**和**其他模态**，并在表1中详细列出现有多模态推荐系统模型所使用的编码器信息。

• **视觉编码器**：视觉特征是MRS中的一种重要模态，例如电影推荐中的海报和时尚推荐中的服装图像。早期的大多数MRS使用基于CNN的预训练模型作为图像编码器。它通过

[2]https://grouplens.org/datasets/movielens/

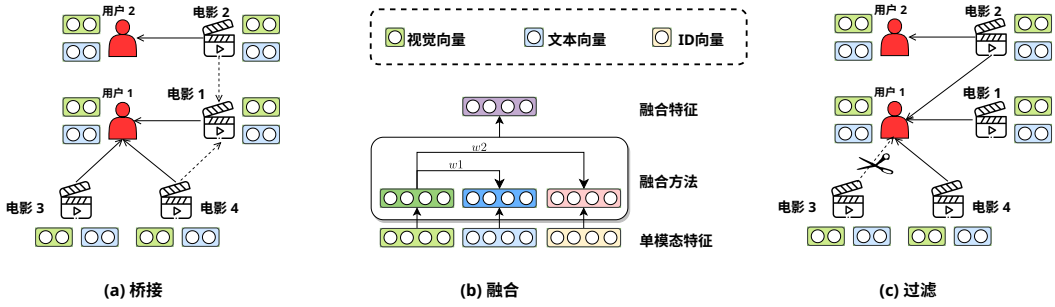

图 2. 三种特征交互类型的示意图。

卷积和池化操作从原始像素信息中压缩并提取特征。例如，MMGCN [65]采用ResNet [15] 从新闻的视觉信息中提取信息。POG [6]使用VGG作为服装图像的图像编码器。近年来，基于Transformer [54] 架构的视觉预训练模型取得了更好的性能，一些MRS模型[8, 13] 也开始使用ViT [12] 提取视觉特征。因此，我们主要将视觉编码器分为基于CNN、基于ResNet和基于Transformer三类。

- **文本编码器**：文本信息（如物品描述）通常比图像包含更多语义信息，因此更适合用于增强用户兴趣建模 [82]。一些多模态推荐系统模型 [14, 52] 通常使用 GloVe [48] 或 Word2Vec [44] 提取文本特征。此外，一些模型还会使用 Text-CNN [21]等编码器。随着自然语言模型的发展，文本编码器逐渐统一为 BERT [11]。总而言之，我们将文本编码器分为五类，即，基于 BERT、基于RNN、基于CNN、基于Sentence Transformer 和基于Word2Vec。

- **其他模态编码器**：对于声学和视频数据，通常会先将其转换为文本或视觉信息，再输入文本或视觉编码器。此外，一些方法会直接使用原始数据集中包含的预先发布的特征向量，对声学和视频模态进行建模。

3 特征交互

多模态数据指各种模态的描述信息。由于这些信息较为稀疏且处于不同的语义空间，因此，将其连接到推荐任务至关重要。特征交互能够将不同模态的各种特征空间非线性地变换到公共空间，最终提升推荐模型的性能和泛化能力。如图2所示，我们将交互分为三种类型：**桥接**、**融合**和**过滤**。这三类技术从不同视图实现交互，因此可以同时应用于同一个多模态推荐系统模型。为便于阅读，我们还在表2中根据交互类型对现有工作进行了分类。

3.1 桥接

这里的桥接指的是构建多模态信息传递通道，重点在于结合多模态信息，捕捉用户与物品之间的相互关系。多媒体推荐与传统推荐的区别在于，物品包含丰富的多媒体信息。早期的大多数工作只是利用多模态内容来增强物品表示，但往往忽略了用户与物品之间的交互。图神经网络的消息传递机制可以通过用户与物品之间的信息交换来增强用户表示，并进一步捕捉用户对不同模态信息的偏好。

图 2(a)给出了一个示例：许多模型通过聚合每种模态下发生交互的物品来获取用户 1 的偏好。此外，电影 1 的模态表示可以从潜在物品-物品图中推导得出。本小节将介绍多模态推荐系统中构建桥接的方法。

3.1.1 用户-物品图。 利用用户与物品之间的信息交换，可以捕捉用户对不同模态的偏好。因此，一些研究采用用户-物品图。MMGCN [65] 为每种模态构建一个用户-物品二部图。对于每个节点，可以利用相邻节点的拓扑结构和物品的模态信息来更新该节点的特征表示。在 MMGCN 的基础上，GRCN [71] 通过在模型训练期间自适应地修改图的结构，删除错误的交互数据（用户点击了不感兴趣的视频），从而提升推荐性能。尽管这些方法在性能上取得了巨大成功，但它们仍受限于使用统一的方式融合用户对不同模态的偏好，忽视了用户对不同模态偏好程度的差异。换言之，对每种模态赋予相同权重可能导致模型性能不够理想。为了解决这一问题，DualGNN [59] 利用用户之间的相关性，基于二部图和用户共现图学习用户偏好。此外，MMGCL [70] 设计了一种新的多模态图对比学习方法来解决这一问题。MMGCL 使用模态边损失和模态掩蔽来生成用户-物品图，并引入一种新颖的负采样技术来学习模态之间的相关性。MGAT[53] 在 MMGCN 的基础上引入注意力机制，有助于自适应地捕捉用户对不同模态的偏好。此外，MGAT 使用门控注意力机制判断用户对不同模态的偏好，从而捕捉用户行为中包含的相对复杂的交互模式。

3.1.2 物品-物品图。 上述研究主要致力于利用多模态特征对用户-物品交互进行建模，却忽略了潜在的语义物品-物品结构。合理利用物品-物品结构有助于更好地学习物品表示并提升模型性能。例如，LATTICE [73] 基于用户-物品二部图为每种模态构建物品-物品图，并将其聚合以获得潜在物品图。MICRO [74] 同样为每种模态构建物品-物品图。不同于 LATTICE，MICRO 在执行图卷积后采用一种新的对比方法来融合特征。然而，这些研究没有考虑不同特定用户群体之间偏好的差异。此外，HCGCN [45]提出了一种聚类图卷积网络，首先对物品-物品图和用户-物品图进行分组，然后通过动态图聚类学习用户偏好。除此之外，受预训练模型近期成功的启发，PMGT [39]提出了一种参考 BERT 结构的预训练图 Transformer，以多模态形式提供项目关系及其相关侧信息的统一视图。BGCN [3], 作为一种捆绑推荐模型，将用户-物品交互、用户-捆绑交互和捆绑-物品关联统一到一个异构图中，并利用图卷积提取细粒度特征。Cross-CBR [43] 构建了用户-捆绑图、用户-物品图和物品-捆绑图，并利用对比学习从捆绑和物品视图对它们进行对齐。

3.1.3 知识图谱。 知识图谱（KG）因能够为推荐系统提供辅助信息而得到广泛应用。为将知识图谱与多模态推荐系统相结合，许多研究人员将物品的各个模态引入知识图谱，并将其作为实体。MKGAT [52]是首个将知识图谱引入多模态推荐的模型。MKGAT提出了一种多模态图注意力技术，从实体信息聚合和实体关系推理两个方面对多模态知识图谱进行建模

表2. 特征交互类别

交互	Goal	类别	相关工作
桥接	捕捉相互关系 用户与物品之间	用户-物品图 物品-物品图 知识图谱	[53], [46], [59], [65], [70], [72], [33], [83], [63], [64] [74], [73], [45], [39], [3], [43], [74], [17], [72], [84] [62], [55], [2], [58], [8], [52], [35]
融合	结合各种偏好 与模态	粗粒度注意力 细粒度注意力 组合注意力 其他融合方法	[40], [47], [38], [7], [72], [84], [63] [6], [20], [67], [28], [53], [46], [39], [8], [14], [34], [20], [31], [24], [35], [5], [16],[30], [85], [66], [25] [36], [32], [13], [17] [61], [4], [41], [41], [76], [68], [1], [83], [29]
过滤	过滤掉噪声数据	过滤	[52], [81], [71], [37], [69], [51], [26], [79]

。此外，该模型采用一种新颖的图注意力网络，在考虑知识图谱中关系的同时聚合邻近实体。SI-MKR [62] 基于交替训练以及MKR [55]提出的知识图谱表示，提出了一种增强型多模态推荐方法。此外，大多数多模态推荐系统都忽略了数据类型多样性问题。SI-MKR通过添加来自知识图谱的用户和物品属性信息解决了这一问题。相比之下，MMKGV[35] 采用图注意力网络在知识图谱上进行信息传播和信息聚合，将多模态信息相结合，并利用知识图谱的三元组推理关系。CMCKG [2] 将描述性属性和结构连接中的信息视为两种模态，并通过最大化这两种视图之间的一致性来学习节点表示。

3.2 融合

在多模态推荐场景中，用户和物品的多模态信息类型繁多、数量庞大。因此，有必要融合不同的多模态信息，为推荐任务生成特征向量 [27]。与桥接相比，融合更关注物品的多模态内部关系。具体而言，其旨在将各种偏好整合到不同模态中。由于内部关系和跨模态关系对于学习全面的表示至关重要，许多多模态推荐系统模型 [46, 53] 甚至会同时采用融合与桥接。注意力机制是应用最广泛的特征融合方法，它能够灵活地以不同权重融合多模态信息，并将重点放在不同方面。本小节如图2(b)所示，首先按照融合粒度对注意力机制进行分类，然后介绍多模态推荐系统中其他一些现有的融合方法。

3.2.1 粗粒度注意力。 一些模型在粗粒度层面应用注意力机制来融合来自多个模态的信息。例如，UVCAN [38] 将多模态信息划分为用户侧和物品侧，其中包括各自的 ID信息和侧信息。它利用用户侧的多模态数据，通过自注意力为物品侧生成融合权重。除了用户侧和物品侧之外，一些模型还会融合不同模态方面的信息。CMBF [7] 引入交叉注意力机制，分别共同学习图像和文本模态的语义信息，然后将二者拼接起来。此外，在一些模型中，各种模态所占的比例也有所不同。MML [47] 设计了一个基于 ID信息的注意力层，并辅以视觉和文本信息。Liu 等人 [40] 指出，每种模态所占的位置相同，而融合权重由自注意力机制决定。相比之下，HCGCN [45] 更关注物品本身的视觉和文本信息。此外，MGCN [72] 提出了行为感知注意力

以结合重要性各不相同的模态。Zhou等人 [84] 强调身份与其他模态之间的差异，并据此设计了用于融合的多层注意力。

3.2.2 细粒度注意力。 多模态数据包含全局特征和细粒度特征，例如音频录音的音调或服装的图案。由于粗粒度融合通常具有侵入性且不可逆 [32], ，会损害原始模态信息并降低推荐性能，因此一些研究考虑采用细粒度融合，有选择地融合不同模态间的细粒度信息。

细粒度融合在时尚推荐场景中意义重大。POG [5] 是一个基于Transformer架构的大型在线服装推荐系统。编码器通过多层注意力挖掘时尚图像中属于搭配方案的深层特征，从而持续实现细粒度整合。与POG相比，NOR [30] 采用了编码器-解码器Transformer架构，其中包含细粒度自注意力结构。它能够根据搭配信息生成相应的方案描述。此外，为了提高可解释性，EFRM [16] 还设计了一个语义提取网络（SEN），用于提取局部特征，最后通过细粒度注意力偏好融合这两种特征。VECF [6]通过图像分割，将各图像块的图像特征与其他模态整合。UVCAN [31] 像VECF一样对视频截图进行图像分割，并分别通过注意力机制，将图像块与ID信息和文本信息融合。MM-Rec [66] 首先通过目标检测算法Mask-RCNN从新闻图像中提取感兴趣区域，然后使用协同注意力将兴趣点与新闻内容融合。MINER [25],、DMIN [67], 和SUM [28] 均通过多模态信息构建用户不同方面的兴趣表示。

一些其他模型设计了独特的内部结构，以实现更好的细粒度融合。例如，MKGformer [8] 通过共享部分 QKV参数并采用相关的感知融合模块来实现细粒度融合。MGAT [53] 使用门控注意力机制来关注用户的局部偏好。MARIO [20] 通过考虑每种模态对每次交互的单独影响来预测用户偏好。因此，它设计了一种模态感知注意力机制，用于识别不同模态对每次交互的影响，并对不同模态进行逐点乘法。

3.2.3 组合注意力. 在细粒度融合的基础上，一些模型设计了组合融合结构，希望在融合细粒度特征的同时，也能保留全局信息的聚合。NOVA [32]将侧信息引入序列推荐。它指出，使用标准注意力直接融合不同模态的特征通常收效甚微，甚至会导致性能下降。因此，它提出了一种包含两个分支的非侵入式注意力机制，并将身份嵌入分别置于两个分支中，以在融合过程中保留交互信息。NRPA [36] 提出了一种个性化注意力网络，该网络考虑了由用户评论所表征的用户偏好。它使用个性化词级注意力，为每个用户/物品筛选出评论中更重要的词语，并依次通过细粒度与粗粒度融合，将评论信息传递至注意力层。VLSNR [13] 是序列推荐的另一种应用，即即 新闻推荐。它能够对用户的短期与长期兴趣进行建模，并通过多头注意力和GRU网络实现细粒度与粗粒度融合。此外，MMSR [17] 采用双重注意力，在序列推荐中保留模态的时间顺序。

3.2.4 其他融合方法。 除了通过注意力权重融合多模态信息外，一些研究还采用了其他简单方法，包括拼接操作 [76], 和门控机制 [32]。然而，它们很少单独出现，通常与上述图和

注意力机制结合使用。已有研究 [32] 表明，如果使用得当，简单的交互不会损害推荐效果，还能降低模型的复杂度。此外，一些早期模型采用循环神经网络 [13], 等结构，尝试通过多模态信息对用户的时间偏好进行建模。其他模型则通过线性层和非线性层融合多模态特征。Lv 等人 [41]在相应位置设置了一个线性层，用于融合文本和视觉特征。在 MMT-Net [23], 中，餐厅数据的三个上下文不变量被人为标记，并通过多层感知机进行交互。近年来，研究者还开发了更多用于融合的人工设计架构，例如专家混合 [1] 和MLP混合器 [29]。

3.3 过滤

多模态数据不同于用户交互数据，其中包含大量与用户偏好无关的信息。例如，如图2(c) 所示，电影 3 与用户 1 之间的交互存在噪声，应当将其移除。在多模态推荐任务中，过滤噪声数据通常能够提升推荐性能。值得注意的是，噪声可能存在于交互图或多模态特征本身，因此，过滤可以分别嵌入桥接和融合过程中。

一些模型使用图像处理进行去噪。例如，VECF [6] 和 UVCAN [31] 通过图像分割去除图像中的噪声，从而更好地建模用户的个性化兴趣。MM-Rec [66] 使用目标检测算法选取图像中的重要区域。

此外，许多基于图神经网络的结构也被用于去噪。由于用户-物品交互的稀疏性以及物品特征中的噪声，通过图聚合学习到的用户和物品表示本身就包含噪声。FREEDOM [81] 设计了一种度敏感边剪枝方法，用于对用户-物品交互图进行去噪。GRCN[71] 检测用户是否意外地与噪声物品发生交互。与图卷积网络模型不同，GRCN能够在训练过程中自适应地调整交互图的结构，以识别并剪除错误的交互信息。PMGCRN [19] 也考虑了用户与不感兴趣物品之间的交互，但不同于 GRCN，它通过主动注意力机制处理不匹配的交互，以纠正用户错误的偏好。此外，MEGCF [37] 关注多模态特征提取与用户兴趣建模之间的不匹配问题。它首先构建一个多模态用户-物品图，然后利用评论数据中的情感信息，在图卷积网络模块中对邻居聚合进行细粒度加权，从而过滤噪声。MAGAE [69] 是一种旨在处理多模态数据不确定性的模型。此外，Shang 等人。 [51] 首先发现了模态不平衡问题，并提出过滤掉具有敏感模态的物品。随后，MCLN [26] 整合了一种新颖的反事实框架来消除噪声。

4 特征增强

同一对象的不同模态表示具有独特或共同的语义信息。因此，如果能够区分独特特征与共同特征，多模态推荐系统的推荐性能和泛化能力将得到显著提升。近期，为了解决这一问题，一些模型采用了**解耦表示学习 (DRL)**和**对比学习 (CL)** ，基于交互进行特征增强，如图 3(a) 和 (b) 所示。

4.1 解耦表示学习

不同模态的特征对于推荐系统中用户对目标物品某一特定因素的偏好，其重要性各不相同。然而，每种模态中不同因素的表示往往相互纠缠，因此许多研究者引入了分解学习技术，以挖掘

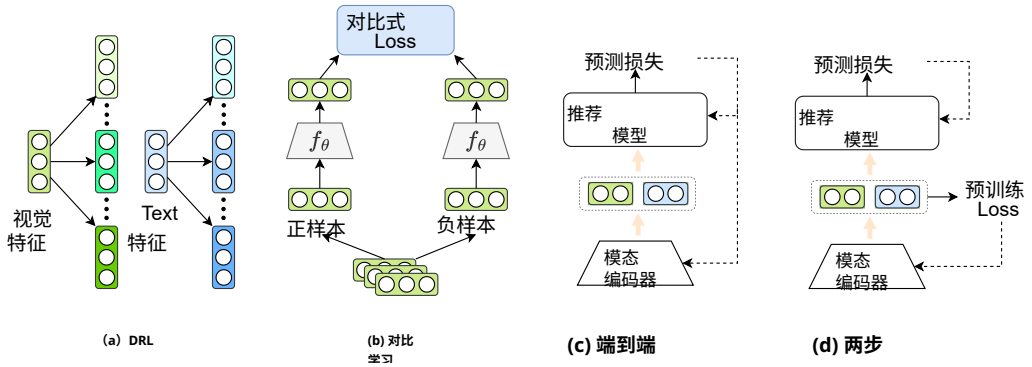

图 3. 特征增强和模型优化示意图。

用户偏好中细致的因素，例如 DICER [77], MacridVAE [42], CDR [4]。此外，多模态推荐旨在从多模态数据中发现由各种隐藏因素形成的有用信息，而这些因素以复杂的方式高度纠缠。MDR [61] 提出了一种多模态解耦推荐方法，能够学习充分解耦的表示，这些表示承载了来自不同模态的互补信息和通用信息。DMRL [34] 考虑了各种模态特征对每个解耦因子的不同贡献，以捕捉用户偏好。此外，PAMD [14] 设计了一种解耦编码器，可自动提取模态共有特征，同时保留模态特定特征。此外，所设计的对比学习保证了分离后的模态表示之间的一致性和差异。与 MacridVAE 相比，SEM-MacridVAE [60]在从用户行为中学习解耦表示时考虑了物品语义信息。

4.2 对比学习

与 DRL 不同，对比学习方法通过数据增强来提升表示，这也有助于解决稀疏性问题。此外，许多研究提出了对比学习损失函数，主要用于模态对齐。

Liu 等 [40] 提出了一种新颖的对比学习损失，使同一物品的不同模态表示具有语义相似性。此外，GHMFC [58] 基于图神经网络得到的实体嵌入表示，构建了两个对比学习模块。这两个对比学习损失函数分别针对两个方向，即，文本到图像和图像到文本。Cross-CBR [43]提出了一种对比学习损失，用于对齐捆绑视图和物品视图中的图表示。MICRO [74]同时关注共享的模态信息和特定的模态信息。在 CMCKG [2], 中，通过知识图谱获取描述性属性和结构链接信息中的实体嵌入，用于计算对比损失。在 HCGCN [45], 中，为了使视觉和文本物品特征映射到同一语义空间，模型借鉴了 CLIP [49]，采用对比学习来最大化一个批次中正确视觉—文本对的相似度。此外，Zhou 等 [83] 和 Wei等 [63] 均设计了基于模态的对比学习损失，用于对齐不同模态并挖掘其相互关系。

由于对比学习的核心是挖掘正负样本之间的关系，许多模型采用数据增强方法，在推荐场景中构造正样本。MGMC [78] 设计了图增强来扩充样本，并引入元学习以提升模型泛化能力。MML [47] 是一种序列推荐模型，它通过构造用户历史购买物品序列的一个子集来扩充训练数据。LHBPMR [45] 从图卷积中选取偏好相似的物品，以

构造正样本。MMGCL [70] 通过模态边丢弃和模态掩蔽来构造正样本。此外，Victor [24] 首次通过中文语义构造样本。Combo-Fashion [85] 是一种套装时尚推荐模型，因此会构造负样本和正样本时尚匹配方案。大多数现有模型都考虑移除多模态数据中不属于用户偏好的信息。相比之下，UMPR [68]直接构造了一种损失，用来描述视觉正负样本之间的差异。

5 模型优化

与传统推荐任务不同，由于存在多模态信息，在多模态编码器与推荐系统一同训练时，模型训练所需的计算量会大幅增加。因此，训练时的多模态推荐系统可分为两类：**端到端训练** 和**两阶段训练**。如图 3(c) 所示，端到端训练能够利用反向传播得到的每个梯度更新模型所有层的参数。相比之下，两阶段训练包括两个阶段：第一阶段是预训练多模态编码器，第二阶段是面向任务的优化，如图 3(d) 所示。

5.1 端到端训练

由于多模态推荐系统会使用图片、文本、音频及其他多媒体信息，因此在处理这些多模态数据时，通常会采用其他领域的一些常见编码器，例如 ViT [12], ResNet [15], BERT [11], 。这些预训练模型的参数量往往非常庞大。例如，ViT-Base [12] 的参数量高达 86M，这对计算资源构成了巨大挑战。为解决这一问题，大多数多模态推荐系统会直接采用预训练编码器，仅以端到端模式训练推荐模型。NOVA [32]和 VLSNR [13] 使用预训练编码器对图像和文本特征进行编码，再将得到的多模态特征向量输入模型，并为用户生成推荐。研究表明，即使不更新编码器参数，仅引入多模态数据也能提升推荐性能。Liu 等人 [40] 提出仅通过推荐损失和对比损失进行 100 个轮次的训练，以微调编码器参数。具体而言，MG [79] 通过梯度策略消除多模态输入中包含的噪声。近期，一些研究者 [56, 75]仅利用预训练编码器提取的物品多模态特征，从而使多模态推荐系统摆脱特定数据集中物品标识的限制。

一些端到端方法还致力于在提升推荐性能的同时减少计算量。训练时，它们通常会减少需要更新的参数数量。例如，MKGformer [8] 是一种多层 Transformer 结构，其中多个注意力层的参数共享，以减少计算量。FREEDOM [81] 旨在冻结图结构中的部分参数，从而大幅降低内存开销，并实现去噪效果。

5.2 两阶段训练

与端到端模式相比，两阶段训练方案能更好地针对下游任务，但需要更多的计算资源。因此，很少有多模态推荐系统采用两步训练。PMGT [39] 提出了一种参照 BERT 结构的预训练图 Transformer。它通过两个目标学习物品表示：图结构重建和掩码节点特征重建。InPOG [5], 预训练了一个 Transformer 来学习时尚匹配知识，然后通过服装生成模型为用户提供推荐。此外，这种方法在序列推荐中很常见，因为序列推荐难以采用端到端方案训练模型。例如，在预训练阶段，MML [47] 首先通过元学习训练元学习器，以增强模型泛化能力，然后在第二阶段训练物品嵌入生成器。此外，

表3. 多模态推荐系统数据集汇总。

Data	领域	模态	规模	Link
TikTok	微视频	V、T、M、A	726K+	https://paperswithcode.com/dataset/tiktok-dataset
Kwai	微视频	V、T、M	100 万+	https://zenodo.org/record/4023390#.Y9YZ6XZBw7c
Movielens + IMDB	电影	V,T	10 万~25米	https://grouplens.org/datasets/movielens/
豆瓣	电影、图书、音乐	V,T	100万+	https://github.com/FengZhu-Joey/GA-DTCDR/tree/main/数据
Yelp	POI	V、T、兴趣点	100万+	https://www.yelp.com/dataset
亚马逊	电子商务	V,T	1亿+	https://cseweb.ucsd.edu/ jmcauley/数据集.html#亚马逊_reviews
Book-Crossings	Book	V,T	100万+	http://www2.informatik.uni-freiburg.de/ cziegler/BX/
亚马逊图书	Book	V,T	300万	https://jmcauley.ucsd.edu/data/亚马逊/
亚马逊时尚	时尚	V,T	100 万	https://jmcauley.ucsd.edu/data/amazon/
POG	时尚	V,T	100 万+	https://drive.google.com/drive/folders/1xFdx5xuNXHGsUVG2VIohFTXf9S7G5veq
天猫	时尚	V,T	800万+	https://tianchi.aliyun.com/dataset/43
淘宝	时尚	V,T	100万+	https://tianchi.aliyun.com/dataset/52
天池新闻	News	T	300万+	https://tianchi.aliyun.com/competition/entrance/531842/introduction
MIND	News	V,T	1500万+	https://msnews.github.io/
Last.FM	音乐	V,T,A	186 k+	https://www.heywhale.com/mw/dataset/5cfe0526e727f8002c36b9d9/content
MSD	音乐	T,A	4800万+	http://millionsongdataset.com/challenge/

[1] “V”“T”“M”“A”分别表示视觉数据、文本数据、视频数据和声学数据。

TESM [46] 和 Victor [24] 分别预训练了精心设计的图神经网络和视频Transformer。近年来，一些更先进的技术被用于提高训练效率，例如知识蒸馏和提示调优。对于前一种技术，SGFD [33]在针对推荐任务进行微调时，从预训练的模态编码器中蒸馏出一个更轻量的模态编码器。此外，PromptMM [64] 提出了一种预训练提示方案，以实现更轻松的微调和更强的任务适应性。

6 应用与资源

如今，当用户浏览在线购物平台时，他们会接收到大量有关物品的多模态信息，这些信息会在不知不觉中影响用户的行为。尽管大多数研究人员旨在提出一个适用于所有应用的通用多模态推荐系统模型，但针对一些典型应用设计独特的模型会更好。例如，在时尚推荐应用中，用户通常会因为服装的款式而受到诱惑，购买原本并不需要的服装。POG [5] 提出利用服装物品的图像和标题来预测其款式是否与用户匹配并受到用户偏好。此外，物品的内容对于新闻推荐非常重要，因此相关研究着重关注文本特征 [13, 25]。通常，通用推荐系统也可以适用于这些应用，但与多模态推荐系统模型相比，其性能明显较差 [5, 25, 29, 85]。

数据集是研究多模态推荐系统的一项必要资源，尤其是针对这些典型应用。因此，为了便于获取这些重要资源，我们根据多模态推荐系统的应用情况，在表3中总结了若干热门数据集。这将帮助研究人员方便地获取这些多模态推荐系统数据集。任何希望使用这些数据集的人，都可以查阅相应的引文和网站以了解更多详情。对于多模态推荐系统模型的评估指标，它们通常与通用推荐系统相同，例如命中率和归一化折损累积增益。

如前所述，多模态推荐系统通常由多个架构组件构成，这给实际系统的实现带来了技术难题。例如，参数量庞大的模态编码器难以部署。此外，由于存在许多不同分支，例如不同的交互模块，因此很难为多模态推荐系统模型设计统一的训练流程。对于

数据预处理而言，数据划分和过滤的设置各不相同，这给复现带来了挑战。为了解决这些问题，以下两个开源基准很有帮助：

• **MMRec**[3]：MMRec 是一款基于 PyTorch 的多模态推荐工具箱。它集成了十多种优秀的多模态推荐系统模型，例如 MMGCN [65]。

• **Cornac**[4] [50]：Cornac 是一个面向多模态推荐系统的比较框架。它为多模态推荐系统提供了完整的实验流程，即，也就是数据、模型、指标和实验。此外，Cornac 与 TensorFlow 和 PyTorch 等主流深度学习框架具有很高的兼容性。

7 挑战与未来方向

为了激励有志于投身这一领域的研究人员，我们列出了若干亟待开展研究的挑战：

• **通用解决方案**。值得注意的是，尽管已有研究针对模型不同阶段提出了一些方法 [24], ，但目前尚无将这些技术组合起来的最新通用解决方案。

• **模型可解释性**。多模态模型的复杂性可能使人难以理解和解释系统生成的推荐，从而限制系统的可信度和透明度。尽管少数先行者 [6, 16] 提到了这一点，但仍有待进一步探索。

• **计算复杂度**。由于模态编码器参数密集，多模态推荐系统需要大量计算资源，因此难以在大型数据集和大规模用户群体上训练模型。此外，多模态数据和模型的复杂性会增加推理成本及生成推荐所需的时间，从而给实时应用带来挑战。

• **过拟合风险**。由于多模态推荐系统数据的稀疏性，以及从人工构造的模态编码器中获得的丰富表征，多模态推荐系统模型容易出现过拟合。

• **隐私**。尽管多模态信息能够缓解数据稀疏性，从而使推荐系统受益，但也会增加隐私泄露的风险。在多模态信息丰富的情况下如何保护个人隐私，也是研究人员面临的一大挑战。

• **大型通用多模态推荐系统数据集**。目前，多模态推荐系统数据集的规模仍然有限，涵盖的模态也不够广泛。此外，不同模态数据的质量和可用性可能有所差异，从而影响推荐的准确性和可靠性。因此，亟需构建一个模态丰富的大型高质量多模态推荐系统数据集。

• **不完整和有偏数据**。在实际应用中，多模态数据往往不完整或存在偏差。例如，在数据收集过程中可能会遗漏某种特定模态。此外，实际交互往往受到流行度的影响而呈现偏斜。应对这两项数据挑战，将推动多模态推荐系统在工业领域的应用。

此外，我们还提出以下若干未来方向：

• **跨模态表示学习**。预计开发更好的跨模态表示学习方法，例如迁移学习和大规模预训练编码器，将使多模态推荐系统更加有效、高效。

• **与其他技术的融合**。预计与增强现实和虚拟现实等技术的融合将提升用户体验，并为多模态推荐带来新的机遇。

[3]https://github.com/enoche/MMRec [4]
https://github.com/PreferredAI/cornac

• **多模态大语言模型的应用**。多模态大语言模型（MLLM）展现出了出色的理解和推理能力。因此，将 MLLM 应用于多模态推荐系统具有相当大的前景 [57]。

8 结论

多模态推荐系统凭借聚合不同模态的优势，正成为推荐系统领域的主要研究方向之一。本文根据不同建模阶段面临的挑战，对多模态推荐系统进行了分类，涵盖模态编码器、特征交互、特征增强和模型优化。我们还总结了数据集和开源代码。最后，提出多模态推荐系统面临的一些挑战与未来方向，以期推动进一步研究。

致谢

本研究部分受研究影响基金（No.R1015-23）、APRC - CityU 新研究计划（No.9610565, CityU 新进教师启动基金）、CityU - HKIDS 青年研究人员研究资助（No.9360163）、香港创新科技署创新及科技基金大学中游研究计划项目（No.ITS/034/22MS）、香港环境及自然保育基金（No. 88/2022），以及 SIRG - CityU 战略跨学科研究资助（No.7020046）资助；还获得华为（华为创新研究计划）、腾讯（CCF-腾讯开放基金、腾讯犀牛鸟专项研究计划）、蚂蚁集团（CCF-蚂蚁科研基金、蚂蚁集团科研基金）、阿里巴巴（CCF-Alimama Tech KangarooFund（No. 2024002））、CCF-百川-易百纳基础模型基金和快手的支持。

参考文献

[1] Shuqing Bian、Xingyu Pan、Wayne Xin Zhao、Jinpeng Wang、Chuyuan Wang 和 Ji-Rong Wen。2023年。用于序列推荐的多模态专家混合表示学习。载于*CIKM*论文集。[2] Xianshuai Cao、Yuliang Shi、Ji hu Wang、Han Yu、Xinjun Wang 和 Zhongmin Yan。2022年。用于机器学习方法推荐的跨模态知识图谱对比学习。载于*ACM MM*论文集。[3] Jianxin Chang、Chen Gao、Xiangnan He、Depeng Jin 和 Yong Li。2020年。基于图卷积网络的捆绑推荐。载于*SIGIR*论文集。[4] Hong Chen、Yudong Chen、Xin Wang、Ruobing Xie、Rui Wang、Feng Xia 和 Wenwu Zhu。2021年。基于噪声多反馈的课程式解耦推荐。*NeurIPS*论文集（2021年）。[5] Wen Chen、Pipei Huang、Jiaming Xu、Xin Guo、Cheng Guo、Fei Sun、Chao Li、Andreas Pfadler、Huan Zhao 和 Binqiang Zhao。2019年。POG：面向阿里巴巴 iFashion 时尚推荐的个性化穿搭生成。载于*KDD*论文集。[6] Xu Chen、Hanxiong Chen、Hongteng Xu、Yongfeng Zhang、Yixin Cao、Zheng Qin 和 Hongyuan Zha。2019年。基于多模态注意力网络的视觉解释个性化时尚推荐：迈向视觉可解释推荐。载于*SIGIR*论文集。[7] Xi Chen、Yangsiyi Lu、Yuehai Wang 和 Jianyi Yang。2021年。CMBF：基于跨模态融合的推荐算法。*Sensors*（2021年）。[8] Xiang Chen、Ningyu Zhang、Lei Li、Shumin Deng、Chuanqi Tan、Changliang Xu、Fei Huang、Luo Si 和 Huajun Chen。2022年。用于多模态知识图谱补全的混合 Transformer 与多层级融合。*arXiv*预印本 *arXiv:2205.02357*（2022年）。[9] Yashar Deldjoo、Fatemeh Nazary、Arnau Ramisa、Julian Mcauley、Giovanni Pellegrini、Alejandro Bellogin 和 Tommaso Di Noia。2022年。现代时尚推荐系统综述。*arXiv*预印本 *arXiv:2202.02757*（2022年）。[10] Yashar Deldjoo、Markus Schedl、Paolo Cremonesi 和 Gabriella Pasi。2020年。利用多媒体内容的推荐系统。*ACM*计算调查（*CSUR*）（2020年）。[11] Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。2018年。BERT：用于语言理解的深度双向 Transformer 预训练。*arXiv*预印本 *arXiv:1810.04805*（2018年）。[12] Alexey Dosovitskiy、Lucas Beyer、Alexander Kolesnikov、Dirk Weissenborn、Xiaohua Zhai、Thomas Unterthiner、Mostafa Dehghani、Matthias Minderer、Georg Heigold、Sylvain Gelly 等。2020年。一张图像胜过 16×16 个词：用于大规模图像识别的 Transformer。*arXiv*预印本 *arXiv:2010.11929*（2020年）。[13] Songhao Han、Wei Huang 和 Xiaotian Luan。2022年。VLSNR：视觉-语言协调的时间序列感知新闻推荐。*arXiv*预印本 *arXiv:2210.02946*（2022年）。

[14] Tengyue Han、Pengfei Wang、Shaozhang Niu 和 Chenliang Li。2022年。模态匹配模态：面向推荐的模态解耦物品表示预训练。载于 WWW 论文集。[15] Kaiming He、Xiangyu Zhang、Shaoqing Ren 和 Jian Sun。2016年。用于图像识别的深度残差学习。载于 CVPR 论文集。[16] Min Hou、Le Wu、Enhong Chen、Zhi Li、Vincent W Zheng 和 Qi Liu。2019年。可解释时尚推荐：一种由语义属性区域引导的方法。arXiv预印本 arXiv:1905.12862（2019年）。[17] Hengchang Hu、Wei Guo、Yong Liu 和 Min-Yen Kan。2023年。序列推荐系统中的自适应多模态融合。载于 CIKM 论文集。[18] Umair Javed、Kamran Shaukat、Ibrahim A Hameed、Farhat Iqbal、Talha Mahboob Alam 和 Suhuai Luo。2021年。基于内容和上下文的推荐系统综述。International Journal of Emerging Technologies in Learning（iJET）（2021年）。[19] Xiangen Jia、Yihong Dong、Feng Zhu、Yu Xin 和 Jiangbo Qian。2022年。偏好校正多模态图卷积推荐网络。Applied Intelligence（2022年）。[20] Taeri Kim、Yeon-Chang Lee、Kijung Shin 和 Sang-Wook Kim。2022年。MARIO：用于多媒体推荐的模态感知注意力与模态保持解码器。载于 CIKM 论文集。[21] Yoon Kim。2014年。用于句子分类的卷积神经网络。载于 EMNLP 论文集。1746–1751。https: //doi.org/
10 3115 1 14 1181 。/v /D - Yehuda Koren、Robert Bell 和 Chris Volinsky。2009年。推荐系统的矩阵分解技术。Computer（2009年）。[23] Adit Krishnan、Mahashweta Das、Mangesh Bendre、Hao Yang 和 Hari Sundaram。2020年。通过上下文不变量实现迁移学习，用于一对多跨域推荐。载于 SIGIR 论文集。[24] Chenyi Lei、Shixian Luo、Yong Liu、Wanggui He、Jiamang Wang、Guoxin Wang、Haihong Tang、Chunyan Miao 和 Houqiang Li。2021年。通过对比式多模态预训练理解中文视频与语言。载于 ACM MM 论文集。[25] Jian Li、Jieming Zhu、Qiwei Bi、Guohao Cai、Lifeng Shang、Zhenhua Dong、Xin Jiang 和 Qun Liu。2022年。MINER：用于新闻推荐的多兴趣匹配网络。载于 ACL Findings。343–352。[26] Shuaiyang Li、Dan Guo、Kang Liu、Richang Hong 和 Feng Xue。2023年。用于基于多媒体的推荐的多模态反事实学习网络。载于 SIGIR 论文集。[27] Xinhang Li、Xiangyu Zhao、Jiaxing Xu、Yong Zhang 和 Chunxiao Xing。2023年。IMF：用于链接预测的交互式多模态融合模型。载于 WWW 论文集。[28] Jianxun Lian、Iyad Batal、Zheng Liu、Akshay Soni、Eun Yong Kang、Yajun Wang 和 Xing Xie。2021年。面向大规模序列推荐的多兴趣感知用户建模。arXiv预印本 arXiv:2102.09211（2021年）。[29] Jiahao Liang、Xiangyu Zhao、Muyang Li、Zijian Zhang、Wanyu Wang、Haochen Liu 和 Zitao Liu。2023年。MMMLP：用于序列推荐的多模态多层感知机。载于 WWW 论文集。[30] Yujie Lin、Pengjie Ren、Zhumin Chen、Zhaochun Ren、Jun Ma 和 Maarten De Rijke。2019年。通过联合穿搭匹配与评论生成实现可解释穿搭推荐。IEEE Transactions on Knowledge and Data Engineering（2019年）。[31] Bo Liu。2022年。基于隐式语义的个性化微视频推荐。arXiv预印本 arXiv:2205.03297（2022年）。[32] Chang Liu、Xiaoguang Li、Guohao Cai、Zhenhua Dong、Hong Zhu 和 Lifeng Shang。2021年。用于序列推荐中侧信息融合的非侵入式自注意力。载于 AAAI 论文集。[33] Fan Liu、Huilin Chen、Zhiyong Cheng、Liqiang Nie 和 Mohan Kankanhalli。2023年。用于多模态推荐的语义引导特征蒸馏。载于 ACM MM 论文集。[34] Fan Liu、Zhiyong Cheng、Huilin Chen、Anan Liu、Liqiang Nie 和 Mohan Kankanhalli。2022年。用于推荐的解耦多模态表示学习。arXiv预印本 arXiv:2203.05406（2022年）。[35] Huizhi Liu、Chen Li 和 Lihua Tian。2022年。用于视频推荐的多模态图注意力网络。载于 CCET 论文集。[36] Hongtao Liu、Fangzhao Wu、Wenjun Wang、Xianchen Wang、Pengfei Jiao、Chuhan Wu 和 Xing Xie。2019年。NRPA：采用个性化注意力的神经推荐。载于 SIGIR 论文集。[37] Kang Liu、Feng Xue、Dan Guo、Le Wu、Shujie Li 和 Richang Hong。2022年。MEGCF：用于个性化推荐的多模态实体图协同过滤。TOIS（2022年）。[38] Shang Liu、Zhenzhong Chen、Hongyi Liu 和 Xinghai Hu。2019年。用于个性化微视频推荐的用户-视频协同注意力网络。载于 WWW 论文集。[39] Yong Liu、Susen Yang、Chenyi Lei、Guoxin Wang、Haihong Tang、Juyong Zhang、Aixin Sun 和 Chunyan Miao。2021年。利用多模态侧信息预训练图 Transformer，以实现推荐。载于 ACM MM 论文集。[40] Zhuang Liu 、 Yunpu Ma 、 Matthias Schubert 、 Yuanxin Ouyang 和 Zhang Xiong。2022年。用于推荐的多模态对比预训练。载于 ICMR 论文集。

[41] Junmei Lv、Bin Song、Jie Guo、Xiaojiang Du 和 Mohsen Guizani。2019。基于多模态数据的 Top-N 推荐兴趣相关物品相似度模型。IEEE Access（2019）。[42] Jianxin Ma、Chang Zhou、Peng Cui、Hongxia Yang 和 Wenwu Zhu。2019。面向推荐的解耦表示学习。NeurIPS 论文集（2019）。[43] Yunshan Ma、Yingzhi He、An Zhang、Xiang Wang 和 Tat-Seng Chua。2022。CrossCBR：用于捆绑推荐的跨视图对比学习。arXiv 预印本 arXiv:2206.00242（2022）。[44] Tomas Mikolov、Kai Chen、Greg Corrado 和 Jeffrey Dean。2013。在向量空间中高效估计词表示。arXiv 预印本 arXiv:1301.3781（2013）。[45] Zongshen Mu、Yueting Zhuang、Jie Tan、Jun Xiao 和 Siliang Tang。2022。学习用于多媒体推荐的混合行为模式。载于 ACM MM 论文集。[46] Juan Ni、Zhenhua Huang、Yang Hu 和 Chen Lin。2022。结合多模态辅助信息的推荐两阶段嵌入模型。Information Sciences（2022）。[47] 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 CIKM 论文集。[48] Jeffrey Pennington、Richard Socher 和 Christopher D Manning。2014。GloVe：用于词表示的全局向量。载于 EMNLP 论文集。1532–1543。[49] Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等。2021。通过自然语言监督学习可迁移视觉模型。载于 ICML 论文集。[50] Aghiles Salah、Quoc-Tuan Truong 和 Hady W Lauw。2020。Cornac：多模态推荐系统的比较框架。Journal of Machine Learning Research（2020）。[51] Yu Shang、Chen Gao、Jiansheng Chen、Depeng Jin、Huimin Ma 和 Yong Li。2023。通过模态平衡增强多模态推荐的对抗鲁棒性。载于 ACM MM 论文集。[52] Rui Sun、Xuezhi Cao、Yan Zhao、Junchen Wan、Kun Zhou、Fuzheng Zhang、Zhongyuan Wang 和 Kai Zheng。2020。推荐系统中的多模态知识图谱。载于 CIKM 论文集。[53] 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。Information Processing & Management（2020）。[54] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。2017。Attention is All You Need。NeurIPS 论文集（2017）。[55] Hongwei Wang、Fuzheng Zhang、Miao Zhao、Wenjie Li、Xing Xie 和 Minyi Guo。2019。用于知识图谱增强推荐的多任务特征学习。载于 WWW 论文集。[56] Jie Wang、Fajie Yuan、Mingyue Cheng、Joemon M Jose、Chenyun Yu、Beibei Kong、Zhijin Wang、Bo Hu 和 Zang Li。2022。TransRec：从混合模态反馈中学习可迁移推荐。arXiv 预印本 arXiv:2206.06190（2022）。[57] Maolin Wang、Yao Zhao、Jiajia Liu、Jingdong Chen、Chenyi Zhuang、Jinjie Gu、Ruocheng Guo 和 Xiangyu Zhao。2023。蚂蚁集团通过高效剪枝与蒸馏压缩大型多模态模型。arXiv 预印本 arXiv:2312.05795（2023）。[58] Peng Wang、Jiangheng Wu 和 Xiaohang Chen。2022。利用门控层次融合与对比训练进行多模态实体链接。载于 SIGIR 论文集。[59] Qifan Wang、Yinwei Wei、Jianhua Yin、Jianlong Wu、Xuemeng Song 和 Liqiang Nie。2021。DualGNN：用于多媒体推荐的双图神经网络。IEEE Transactions on Multimedia（2021）。[60] Xin Wang、Hong Chen、Yuwei Zhou、Jianxin Ma 和 Wenwu Zhu。2023。面向推荐的解耦表示学习。IEEE Transactions on Pattern Analysis and Machine Intelligence（2023）。[61] Xin Wang、Hong Chen 和 Wenwu Zhu。2021。用于推荐的多模态解耦表示。载于 ICME 论文集。[62] Yuequn Wang、Liyan Dong、Hao Zhang、Xintao Ma、Yongli Li 和 Minghui Sun。2020。基于知识图谱表示交替训练的增强型多模态推荐。IEEE Access（2020）。[63] Wei Wei、Chao Huang、Lianghao Xia 和 Chuxu Zhang。2023。用于推荐的多模态自监督学习。载于 WWW 论文集。[64] Wei Wei、Jiabin Tang、Lianghao Xia、Yangqin Jiang 和 Chao Huang。2024。PromptMM：结合提示调优的推荐多模态知识蒸馏。载于 WWW 论文集。[65] 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 ACM MM 论文集。[66] 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）。[67] Zhibo Xiao、Luwei Yang、Wen Jiang、Yi Wei、Yi Hu 和 Hao Wang。2020。用于点击率预测的深度多兴趣网络。载于 CIKM 论文集。2265–2268。

[68] Cai Xu、Ziyu Guan、Wei Zhao、Quanzhou Wu、Meng Yan、Long Chen 和 Qiguang Miao。2020。面向智慧城市应用的基于用户多模态偏好的推荐。*IEEE工业信息学汇刊* （2020）。[69] Jing Yi 和 Zhenzhong Chen。2021。用于推荐系统的多模态变分图自编码器。*IEEE Transactions on Multimedia* （2021）。[70] 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 *SIGIR* 论文集。[71] Wei Yinwei、Wang Xiang、Nie Liqiang、He Xiangnan 和 Chua Tat-Seng。2021。GRCN：结合隐式反馈进行多媒体推荐的图细化卷积网络。*arXiv* 预印本 *arXiv:2111.02036* （2021）。[72] Penghang Yu、Zhiyi Tan、Guanming Lu 和 Bing-Kun Bao。2023。用于多媒体推荐的多视图图卷积网络。载于 *ACM MM* 论文集。[73] Jinghao Zhang、Yanqiao Zhu、Qiang Liu、Shu Wu、Shuhui Wang 和 Liang Wang。2021。挖掘用于多媒体推荐的潜在结构。载于 *ACM MM* 论文集。[74] Jinghao Zhang、YanqiaoZhu、Qiang Liu、Mengqi Zhang、Shu Wu 和Liang Wang。2022。通过对比模态融合挖掘潜在结构，实现多媒体推荐。*IEEE Transactions on Knowledge and Data Engineering* （2022）。[75] Lingzi Zhang、Xin Zhou 和 Zhiqi Shen。2023。通过对比学习实现序列推荐的多模态预训练框架。*arXiv* 预印本 *arXiv:2303.11879* （2023）。[76] Xiaoyan Zhang、Haihua Luo、Bowei Chen 和 Guibing Guo。2020。用于餐厅推荐的多视图视觉贝叶斯个性化排序。*Applied Intelligence* （2020）。[77] Yin Zhang、Ziwei Zhu、Yun He 和 James Caverlee。2020。用于增强推荐的内容协同解耦表示学习。载于 *RecSys* 论文集。[78] Feng Zhao 和 Donglin Wang。2021。多模态图元对比学习。载于 *CIKM* 论文集。[79] Shanshan Zhong、Zhongzhan Huang、Daifeng Li、Wushao Wen、Jinghui Qin 和 Liang Lin。2024。镜像梯度：通过探索平坦局部极小值实现鲁棒多模态推荐系统。载于 *WWW* 论文集。[80] Hongyu Zhou、Xin Zhou、Zhiwei Zeng、Lingzi Zhang 和 Zhiqi Shen。2023。多模态推荐系统综合综述：分类体系、评估与未来方向。*arXiv* 预印本 *arXiv:2302.04473* （2023）。[81] Xin Zhou。2022。双图之说：冻结并去噪图结构以实现多模态推荐。*arXiv* 预印本 *arXiv:2211.06924* （2022）。[82] Xin Zhou。2023。MMRec：简化多模态推荐。*arXiv* 预印本 *arXiv:2302.03497* （2023）。[83] Xin Zhou、Hongyu Zhou、Yong Liu、Zhiwei Zeng、Chunyan Miao、Pengwei Wang、Yuan You 和 Feijun Jiang。2023。自举多模态推荐的潜在表示。载于 *WWW* 论文集。[84] Yan Zhou、Jie Guo、Hao Sun、Bin Song 和 Fei Richard Yu。2023。注意力引导的多步融合：用于多模态推荐的分层融合网络。载于 *SIGIR* 论文集。[85] Chenxu Zhu、Peng Du、Weinan Zhang、Yong Yu 和 Yang Cao。2022。Combo-Fashion：结合物品历史记录的时尚服装匹配点击率预测。载于 *KDD* 论文集。