CAMMSR：面向多模态序列推荐的类别引导注意力专家混合

1st Jinfeng Xu香港大学中国香港特别行政区
jinfeng@connect.hku.hk
ORCID：0009-0001-7876-3740

2nd Zheyu Chen北京理工大学中国北京
zheyu.chen@bit.edu.cn
ORCID：0009-0003-5779-3523

3rd Shuo Yang香港大学中国香港特别行政区
shuoyang.ee@gmail.com
ORCID：0000-0003-1638-9623

4th Jinze Li香港大学中国香港特别行政区
lijinze-hku@connect.hku.hk
ORCID：0009-0002-6749-5442

5th Hewei Wang卡内基梅隆大学美国匹兹堡
stephenw2026@gmail.com
ORCID：0000-0002-6952-0886

6th Yijie Li卡内基梅隆大学匹兹堡，美国
yijie.li2022@gmail.com
ORCID: 0000-0003-2118-2280

7th Jianheng Tang北京大学北京，中国
tangentheng@gmail.com
ORCID: 0000-0002-4762-5943

8th Yunhuai Liu北京大学北京，中国
yunhuai.liu@pku.edu.cn
ORCID: 0000-0002-1180-8078

9th Edith C. H. Ngai*香港大学中国香港特别行政区，中国
chngai@eee.hku.hk ORCID: 0000-0002-3454-8731

arXiv:2603.04320v1 [cs.IR] 4 Mar 2026

摘要——**信息丰富环境中多媒体数据的激增加剧了个性化内容发现的挑战，使推荐系统成为被动数据管理的一种重要形式。多模态序列推荐利用文本和图像等多样化物品信息，在丰富物品表示和深化对用户兴趣的理解方面展现出巨大潜力。然而，现有大多数模型依赖启发式融合策略，无法捕捉用户与模态的交互所具有的动态性和上下文敏感性。在现实场景中，用户对模态的偏好不仅因人而异，同一用户在不同物品或类别上也会有所不同。此外，模态之间的协同效应——即组合信号以孤立模态无法实现的方式激发用户兴趣——在很大程度上仍未得到充分探索。**

为此，我们提出CAMMSR，一种面向多模态序列推荐的类别引导的注意力专家混合模型。我们的框架整合来自行为序列和时间戳的多模态内容，以提升推荐的鲁棒性与上下文对齐。其核心在于，CAMMSR引入注意力专家混合（CAMoE）模块，该模块从多个视角学习专门化的物品表示，并显式建模模态间的协同作用。该组件在辅助类别预测任务的引导下动态分配模态权重，从而实现多模态信号的自适应融合。此外，我们设计了模态交换对比学习任务，通过序列级增强来加强跨模态表示对齐。在四个公开数据集上的大量实验表明，CAMMSR持续优于最先进的基线方法，验证了其在实现自适应、协同且以用户为中心的多模态序列推荐方面的有效性。

索引词—**多模态、推荐、序列推荐、混合专家**

I. 引言

互联网规模数据的指数级增长，使用户进行主动信息搜寻和个人数据管理日益困难；他们难以准确表达自己的长期偏好，也难以有效筛选异质性强且质量参差不齐的多媒体内容 [1], [2], [3],[4], [5], [6]。在此背景下，推荐系统已成为一种重要的被动式数据管理形式——通过从历史行为交互中推断用户兴趣，智能过滤无关信息，并呈现个性化内容。这类系统将原始交互数据转化为结构化且可执行的洞见，从而减轻用户负担，实现高质量、自动化的信息管理 [7], [8], [9]。其中，序列推荐系统 [10], [11] 通过建模用户行为的时序动态脱颖而出，为动态信息环境中的以用户为中心的数据管理提供了可扩展的机制。通过从交互序列 [12], [13],中捕捉不断演变的偏好，这些系统为信息日益丰富的数字环境中的以用户为中心的数据管理提供了一种可扩展且可持续的方法。

在序列建模的基础上，近期研究致力于通过纳入辅助多模态数据信号（如物品描述和图像），进一步增强推荐系统的表达能力与判别能力，以提升推荐保真度及其与细微用户偏好的对齐 [14], [15]。通过利用与

物品相关的丰富多模态信息——涵盖文本元数据、视觉内容和类别属性——研究人员开发了多模态序列推荐模型 [16], [17], [18], [19]。这些工作旨在构建更细粒度的物品表示，从而更深入地理解用户意图以及序列交互中的兴趣变化 [20], [21], [22], [23]。因此，多模态序列推荐系统已获得学术界与工业界的广泛关注，并被视为通向更智能、更贴近人类的数据管理的一条有前景的路径。

尽管已有这些进展，现有的多模态序列模型往往难以动态适应用户与模态交互所固有的上下文依赖特性。用户对不同模态的兴趣本质上是动态的，不仅因人而异，即便是同一用户，也会随具体物品或情境上下文而变化 [24], [22]。例如，有些用户在浏览时尚物品时可能优先关注视觉外观，而另一些用户——甚至同一用户在不同会话中——可能更依赖文本描述来了解产品规格或评价。此类模态偏好转变往往受到物品类别、购买意图或其他上下文因素的影响 [25], [12]。使这一挑战更为复杂的是，模态之间的协同效应在以往研究中仍 largely 未得到充分探索 [22], [23]。一件产品可能无法仅凭任一单一模态吸引兴趣，但互补信号的同时出现——例如品牌文本与独特视觉图案并存——却能引发显著的用户参与。有效建模此类跨模态协同，对于弥合静态多模态融合与真实世界用户决策过程之间的差距至关重要。

为此，我们提出了 CAMMSR，一种用于多模态序列推荐的类别引导注意力混合专家模型。我们的框架全面整合物品序列中的多模态内容及其对应的时间戳，以实现更加稳健且具备上下文感知能力的推荐。CAMMSR 的核心是引入类别引导注意力混合专家（CAMoE）组件，该组件从多角度专家中捕获专门化物品表示和用户兴趣，同时显式建模模态间协同效应。该模块在辅助类别预测任务的引导下动态分配特定模态的权重，使模型能够根据物品类型和用户行为自适应地平衡多模态信号。此外，我们设计了模态交换对比学习目标，通过序列级增强来提升跨模态对齐，并挖掘模态之间的潜在相关性。在四个公开真实世界数据集上的大量实验验证了 CAMMSR 相较于最先进基线方法的优越性，表明其能够有效推动多模态序列推荐朝着更加自适应、协同化和以用户为中心的数据管理方向发展。总而言之，我们的工作作出以下贡献：

- 我们提出了一种用于多模态序列推荐的新型 CAMMSR 模型。据我们所知，我们首次利用辅助类别预测任务来分配模态权重，以提升多模态序列推荐 .
- 我们提出一种类别引导的注意力专家混合，能够从多个视角高效捕获富有信息的物品表示和用户偏好，同时探索不同模态之间的协同交互。该方法引入辅助类别预测任务，以动态且自适应地平衡不同模态信息的贡献。
- 我们进一步引入一项定制的模态交换对比学习任务，以更好地利用模态之间的相关性，并加强不同模态表示之间的对齐。
- 在四个真实世界数据集上的全面实验证明了CAMMSR的有效性。

II. 相关工作

在本节中，我们回顾与本研究相关的先前研究，重点关注传统序列推荐和多模态序列推荐。

A. 传统序列推荐

序列推荐旨在通过分析用户的历史行为序列，预测用户的下一个交互物品。早期的基础研究主要致力于开发专用的序列编码器，以有效建模用户交互数据中的时间依赖关系。例如，GRU4Rec [26] 率先采用门控循环单元（RNN）来捕捉短期序列模式；Caser [27] 则引入卷积神经网络（CNN），通过水平和垂直滤波器对点级和联合级序列行为进行建模。SASRec [11] 通过采用单向自注意力机制，为整个序列中的相关物品分配权重，从而能够捕捉长程依赖关系，大幅推动了该领域的发展。在Transformer架构的基础上，BERT4Rec [10]通过双向自注意力进一步增强了序列建模能力，使模型能够利用掩码序列中来自过去和未来交互的上下文信息。

认识到时序动态在用户兴趣演化中的关键作用后，后续研究系统地集成了时间感知机制，以改进序列推荐。TiSASRec [28]将交互之间的相对时间间隔显式纳入自注意力计算，使模型能够根据内容与时间邻近性自适应地为物品赋予权重。MEANTIME [29]通过引入多种粒度的时间嵌入——包括绝对、相对和类别时间特征——扩展了这一思路，从而更全面地表示时间上下文。最近，TGCL4SR [30] 采用时序图对比学习，并配合精心设计的时间扰动增强，以强化时间转移模式的编码并提升对时间噪声的鲁棒性。与之并行的进展包括FEARec [13],，该方法引入了一种新颖的频率渐变结构，将时域与频域自注意力模块相结合，以

同时捕捉低频稳定偏好与高频动态兴趣。类似地，TiCoSeRec[12] 提出了五种不同的基于时间间隔的序列增强策略，以归一化用户行为的时间分布，并结合对比学习目标来增强时间表示学习与泛化性能。尽管取得了这些进展，上述方法仅关注行为序列，而忽视了多模态数据的潜力。因此，它们未能充分利用能够提供更丰富用户兴趣表示的多模态特征。

B. 多模态序列推荐

多模态序列推荐已成为一种颇具前景的范式，它利用物品丰富的多模态信号——包括视觉特征和文本特征——来建模复杂的用户偏好与序列行为，从而显著提升推荐质量 [31],[32], [22], [33], [34]。现有研究的核心方向之一，是集成这些异构模态，以构建表达能力更强、能更好感知上下文的物品表示。早期工作如 MV-RNN [35],探索了直接的融合策略，包括逐元素加法、拼接和特征重构，以合并多模态输入。UniSRec [36] 通过采用混合专家（MoE）架构推进了这一研究方向，促进文本表示中的语义迁移；随后再通过参数化加权将其与 ID嵌入融合，从而实现跨域语义对齐。

随后，研究人员开发了更精细且自适应的融合机制，以提升多模态集成的灵活性和可扩展性。MISS-Rec [20] 引入了一种轻量级动态融合模块，通过高效的门控网络自适应地重新校准用户对各模态的注意力，从而实现细粒度的模态加权。MMSR [37] 利用异构图神经网络对模态之间的结构依赖进行建模，并通过专门设计的消息传递机制支持灵活的跨模态交互。为进一步拓展混合专家范式，M3SRec [21] 同时引入模态专属和跨模态MoE模块，使模型能够捕获各模态中的专有模式与共享模式，从而实现更全面的表示学习。作为互补方向，TedRec [23] 应用快速傅里叶变换（FFT），将 ID 和文本嵌入序列投影到频域，在降低高频噪声的同时，高效实现序列级语义融合。最近，FindRec [38]提出了一种以 Stein 核为基础、具有理论依据的方法来确保跨模态一致性，并通过集成信息协调模块对多模态对齐进行形式化，该模块可在模态专属流形之间维持统计相干性。

尽管这些方法展现出了一定进展，但大多数方法都忽略了不同模态之间的协同效应。为此，我们提出了类别引导的注意力MoE，并结合定制的模态交换对比学习，以有效挖掘模态特征并利用不同模态之间的协同效应。

表 I：本文中使用的关键符号。

符号	描述
$\mathcal{I}$	物品集合
$\mathcal{U}$	用户集合
S	历史行为序列
T	时间戳序列
$\mathbf{E}_m$	模态 m的完整表示
$\bar{S}^m$	增强的物品序列表示
$\mathbf{Z}^m$	最终物品序列表示
$f_{i_{id}}$	物品 i的ID表示
f_{i_t}	物品 i的文本表示
f_{i_v}	物品的视觉表示 i
$\mathbf{e}_{i_m}$	模态 m的物品表示 i
$\hat{\mathbf{e}}_{i_m}$	模态 m的物品修订表示 i
$\bar{\mathbf{e}}_{i_m}$	模态 m的物品最终表示 i
$\hat{y}_c^m$	类别 c 的概率
d_i^m	真实分类与预测标签之间的差异
w_i^m	物品 i的模态权重
$\hat{y}_i$	最终预测分数

III. 预备知识

多模态序列推荐旨在利用多模态物品数据和用户过去的行为，为其下一次交互提供个性化建议。设 $\mathcal{U}$ 表示用户集合， $\mathcal{I}$ 表示物品集合。在多模态场景中，每个物品 $i \in \mathcal{I}$表示为 $i = \langle i_{id},\ i_t,\ i_v \rangle$，其中$i_{id}$ 表示物品 i的ID嵌入，而 i_t 和 i_v 分别表示物品 i关联的文本和图像内容。接着，用户 $u \in \mathcal{U}$的历史行为序列 S 表示为 $S = \{i_1, i_2, ..., i_{|S|}\}$，其中 $i_n \in \mathcal{I}$表示第 n个时间步交互的物品。时间戳序列表示为 $T = \{t_1, t_2, ..., t_{|S|}\}$。

问题定义： 给定用户 u 及其包含多模态物品信息的历史行为序列 S ，多模态序列推荐的目标是预测用户 u 在第（$|S|$+1）个时间步可能感兴趣的物品 i 。这可以表述为：寻找给定交互序列时条件概率最大的物品 i ：

$$i = \text{argmax}_{i \in \mathcal{I}} P\left(i_{|S|+1} = i \mid S\right). \tag{1}$$

关键符号汇总于表 I。

IV. 方法

本节介绍 CAMMSR[1] 的技术细节。如图1所示，CAMMSR 由四个主要部分组成：物品表示初始化、类别引导的注意力专家混合、用户兴趣学习和模态交换对比学习。

A. 物品表示初始化

多模态信息包含丰富的语义特征[14], [39], [40],，能够有效表征物品。因此，

[1]代码可在**补充材料**中找到。

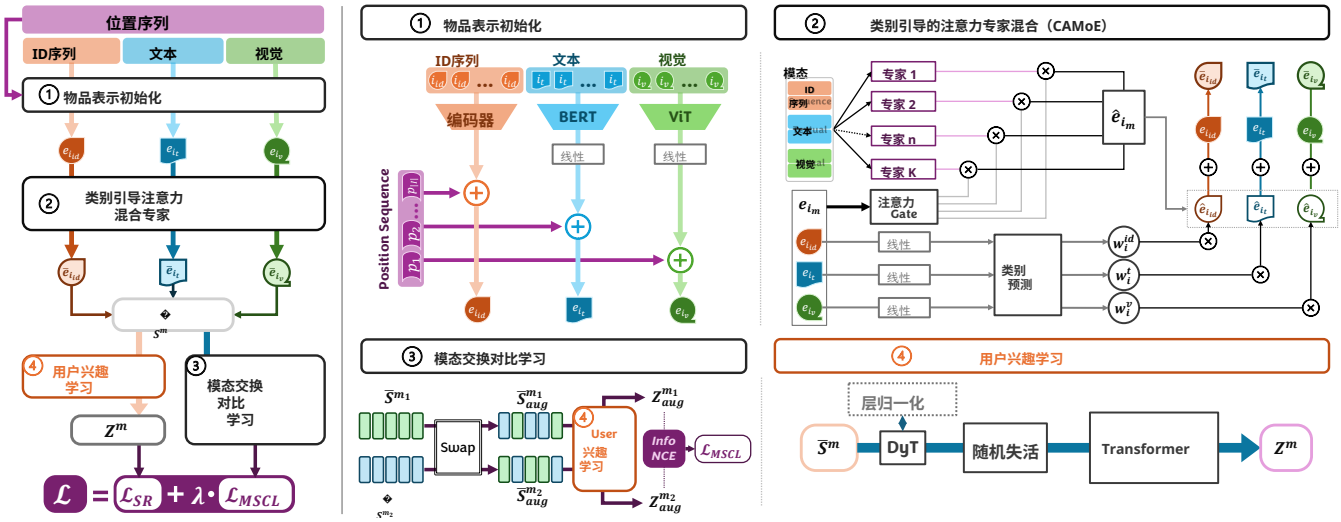


图 1：CAMMSR 的整体框架。左侧给出流程描述，右侧详细说明各个组件。1）物品表示初始化使用预训练提取器结合位置序列初始化物品多模态表示。2）类别引导的注意力专家混合利用额外的类别预测任务和注意力机制，有效分配模态权重。3）模态交换对比学习通过模态交换操作构建增强对比视图，并应用对比学习以增强不同模态表示之间的对齐。4）用户兴趣学习使用基于 Transformer 的编码器对用户行为序列进行建模，并探索将 DyT 作为 Transformer 中层归一化的替代方案。

我们从真实世界推荐中广泛使用的三种模态获取物品表示：ID、文本和视觉。

ID 模态：我们为每个物品 i 初始化一个 ID 表示 $f_{i_{id}} \in \mathbb{R}^d$，其中 d是 ID 模态的维度。**文本模态：**遵循先前研究 [41], [21], ，我们采用广泛使用的预训练 BERT [42] 提取文本特征，以从文本语义中捕获用户偏好。我们将每个物品 i的所有文本描述 $i_{[\text{Words}]}$ 与特殊符号 $i_{[\text{CLS}_t]}$ 拼接。由于 $i_{[\text{CLS}_t]}$ 能够传达整句的语义，我们使用 $i_{[\text{CLS}_t]}$ 的嵌入来表示文本特征。在此，我们将拼接后的句子输入预训练 BERT，以获得文本特征，如下所示：

$$f_{i_t} = \text{BERT}(\text{Con}(i_{[\text{CLS}_t]}, i_{[\text{Words}]})), \tag{2}$$

其中 $f_{i_t} \in \mathbb{R}^{d_t}$ 是 $i_{[\text{CLS}_t]}$的最终隐藏嵌入，Con(·, ·)表示拼接操作。

视觉模态：遵循先前的研究 [41], [21]，我们采用广泛使用的预训练ViT [43] 提取视觉特征，以从视觉语义中捕捉用户偏好。我们将每个物品 i 的图像划分为若干图像块，然后将这些图像块转换为带有特殊符号 $i_{[\text{CLS}_v]}$的序列 $i_{[\text{Patches}]}$。在此，我们将组合后的句子输入预训练ViT，以如下方式获得视觉特征：

$$f_{i_v} = \text{ViT}(\text{Con}(i_{[\text{CLS}_v]}, i_{[\text{Patches}]})), \tag{3}$$

其中 $f_{i_v} \in \mathbb{R}^{d_v}$ 是 $i_{[\text{CLS}_v]}$的最终隐藏嵌入。

为了使文本和视觉模态的维度与 ID 模态对齐，我们采用简单的线性投影来改变它们的维度，形式化表示如下：

$$x_{i_t} = \mathbf{W}_t f_{i_t} + \mathbf{b}_t, \quad x_{i_v} = \mathbf{W}_v f_{i_v} + \mathbf{b}_v, \tag{4}$$

其中$\mathbf{W}_t \in \mathbb{R}^{d_t \times d}$、$\mathbf{W}_v \in \mathbb{R}^{d_v \times d}$、$\mathbf{b}_t \in \mathbb{R}^d$和$\mathbf{b}_v \in \mathbb{R}^d$是可训练参数，而 f_{i_t} 和 f_{i_v} 是冻结的。

此外，位置信息在序列推荐中的序列建模里起着至关重要的作用。为了融入位置信息，我们为所有模态的物品表示引入相应的位置嵌入 $p_i \in \mathbb{R}^d$：$\mathbf{e}_{i_m} = x_{i_m} + p_i$，其中 $m \in \{id, t, v\}$ 表示模态。模态 m 的全部表示记为 $\mathbf{E}_m \in \mathbb{R}^{|\mathcal{I}| \times d}$。

B. 类别引导的注意力专家混合

用户对多模态信息的兴趣是动态的，且并非同等加权。不同用户对不同模态的重视程度各不相同，即便是同一用户，针对不同物品也可能优先考虑不同的模态。例如，在购买美容化妆品时，视觉模态起主导作用，而在购买游戏时，文本模态更具指导意义。为了建模用户对源自多模态信息的推荐兴趣，并为不同模态分配适当的权重，我们提出了类别引导的注意力专家混合（CAMoE）。具体而言，CAMoE进一步利用额外的类别预测任务，通过采用注意力机制学习不同模态的权重，从而更有效地引导模态权重的分配。先前的研究[21], [44], [24], [45], [46] 已证明混合专家（MoE）框架在各种推荐场景中的有效性。这归因于其能够

通过从多个视角捕获专门化物品表示和用户偏好，增强推荐模型的学习能力与灵活性。

现有研究 [21], [36]通常独立处理每种模态的特征，然后应用加权机制。然而，这种范式往往忽视不同模态之间的协同效应。例如，文本模态中描述的折扣或品牌等营销信息可能难以直接说服用户购买某件物品，而动漫图案等视觉元素单独存在时也可能无法直接触发购买。然而，带有特定动漫图案的折扣物品往往会带来购买量激增。为此，我们将其他模态中的有价值信息融入与每种模态对应的专家中。具体而言，给定某件物品的 ID、文本和图像表示 $\mathbf{e}^{i_{id}}$、$\mathbf{e}^{i_t}$ 和$\mathbf{e}^{i_v}$，对目标模态 $m \in \{id, t, v\}$ 的处理如下：

$$\hat{\mathbf{e}}_{i_m} = \sum_{k=1}^{K} g_k^m \, \mathrm{Expert}_k(\mathrm{Con}(\mathbf{e}_{\mathrm{i_{id}}}, \mathbf{e}_{\mathrm{i_t}}, \mathbf{e}_{\mathrm{i_v}})), \tag{5}$$

其中，Con(·, ·)表示拼接操作，Expert$_*$(·)表示专家。 K 是一个超参数，表示每种模态的专家总数，我们在**消融研究**小节中对其进行了实证分析。我们采用前馈网络（ FFN ）作为专家，形式化表示为 Expert$_*$(·) = FFN(x) = (GeLU($x\mathbf{W}_1 + \mathbf{b}_1$))$\mathbf{W}_2 + \mathbf{b}_2$,，其中$\mathbf{W}_1 \in \mathbb{R}^{3d \times 4d}$、$\mathbf{W}_2 \in \mathbb{R}^{4d \times d}$、$\mathbf{b}_1 \in \mathbb{R}^{4d}$和$\mathbf{b}_2 \in \mathbb{R}^{d}$是可学习参数。 g_k^m表示由注意力门控路由器生成的路由向量 $G^m \in \mathbb{R}^K$ 中，模态 m 上第 k个专家对应的组合权重：

$$G^m = \mathrm{Softmax}(Q^m (K^m)^T / \sqrt{K}) V^m, \tag{6}$$

$$Q^m, K^m, V^m = \mathbf{W}_Q^m \mathbf{E}_m, \mathbf{W}_K^m \mathbf{E}_m, \mathbf{W}_V^m \mathbf{E}_m, \tag{7}$$

其中，$\mathbf{W}_Q^m \in \mathbb{R}^{d \times K}$、$\mathbf{W}_K^m \in \mathbb{R}^{d \times K}$和$\mathbf{W}_V^m \in \mathbb{R}^{d \times K}$ 是可学习参数。

由于用户数据和物品数量庞大，用户难以与所有物品进行交互，因此推荐系统通常面临数据稀疏问题 [40], [47],，这也带来了依赖训练不同模态的可学习权重所面临的挑战。受先前研究[48], [49]利用物品类别信息增强对物品特征理解的启发，我们使用物品类别来指导模态权重的分配。具体而言，我们将物品类别视为真实分类标签，并通过衡量真实分类标签与根据相应模态信息得出的预测标签之间的差异来计算模态权重。为生成更贴近用户序列的信号，我们在序列级别使用二元交叉熵损失 [50], [51]来衡量真实分类标签与预测标签之间的差异。对于物品$i \in S$ 在模态 m上的表示，我们使用$\mathbf{e}^{i_m}$ of$i$ 来计算其属于各类别的概率，公式如下：

$$\hat{y}_c^m = \mathbf{W}_c^m \mathbf{e}_{i_m} + \mathbf{b}_c^m, \tag{8}$$

其中， $\mathcal{C}$ 表示类别集合。$\mathbf{W}_c$ 和$\mathbf{b}_c$ 为可学习参数。$\hat{}y_c^m$表示 c 属于 $c \in \mathcal{C}$的似然。因此，通过二元交叉熵损失 d_i^m 计算真实分类标签与预测标签之间的差异，公式如下：

$$d_i^m = -\sum_{c \in \mathcal{C}} y_c \log(\hat{y}_c^m) + (1 - y_c) \log(1 - \hat{y}_c^m), \tag{9}$$

其中， y_c 是真实值[2]。接着，我们进一步计算模态权重，具体形式如下：

$$w_i^m = \frac{\exp(d_i^m)}{\sum_{n \in \{id,t,v\}} \exp(d_i^n)}. \tag{10}$$

获得模态权重 w_i^m、每个物品 i的专家输出$\hat{}\mathbf{e}^{i_m}$和物品表示$\mathbf{e}^{i_m}$ 后，我们得到物品 i 在模态 m上的最终表示 $\bar{}\mathbf{e}^{i_m}$，具体形式如下：

$$\bar{\mathbf{e}}_{i_m} = \mathbf{e}_{i_m} + w_i^m \hat{\mathbf{e}}_{i_m}. \tag{11}$$

C. 用户兴趣学习

为有效捕捉序列交互中用户兴趣的动态演变特性，我们使用基于 Transformer 的编码器 [52],对用户行为序列进行建模，该编码器在序列推荐任务中展现出强大的长程依赖关系和时间模式捕捉能力 [11], [10]。具体而言，对于每种模态m，我们将增强的物品序列表示$\bar{S}^m = \{\bar{}\mathbf{e}^{i_{m,1}}, \bar{}\mathbf{e}^{i_{m,2}}, \ldots\ldots \bar{}\mathbf{e}^{i_{m,|S|}}\}$（即我们的类别引导注意力专家混合（CAMoE）的输出）输入模态专属 Transformer 模型进行处理。编码过程形式化定义如下：

$$\mathbf{Z}^m = \mathrm{Transformer}(\mathrm{Dropout}(\mathrm{LayerNorm}(\bar{S}^m))), \tag{12}$$

其中，Transformer(·) 表示 Transformer 模型，Dropout(·) 表示随机失活操作，LayerNorm(·) 表示层归一化操作。$\mathbf{Z}^m$ 表示模态 m 下与最后一个位置对应的用户兴趣最终表示。

受近期研究发现 [53] 的启发，这些研究强调了 LayerNorm(·) 在对输入进行非线性缩放并通过类似 Tanh(·) 的 S 形函数压缩离群值方面的作用。我们探索了一种轻量且自适应的替代方案：动态双曲正切（DyT）操作，其定义如下：

$$\mathrm{DyT}(x) = \mathrm{Tanh}(\alpha x), \tag{13}$$

其中， α 是一个可学习参数。DyT 模拟了 Layer Norm 的归一化和激活行为，同时通过自适应缩放输入并借助有界双曲正切函数限制输出幅度，提供了更高的效率和灵活性。受其提升表示质量和训练稳定性的潜力启发，我们在式 12 中用 DyT(·) 替换 LayerNorm(·)，得到：

$$\mathbf{Z}^m = \mathrm{Transformer}(\mathrm{Dropout}(\mathrm{DyT}(\bar{S}^m))). \tag{14}$$

[2]所有基线方法均可获取类别信息。所有基线方法都将标题、类别和品牌字段拼接为文本信息。

在**消融研究**小节中，我们通过实证分析探讨以 DyT(·) 替换 LayerNorm(·) 所带来的性能提升。为了预测用户 u与物品 i之间的交互概率，我们首先计算每种模态的得分，然后将这些得分相加，得到最终预测 $\hat{y}_i$，形式如下：

$$\hat{y}_i = \sum_{m \in \{id,t,v\}} \mathbf{Z}^m \mathbf{e}_{i_m}. \tag{15}$$

D. 模态交换对比学习

为增强对多模态信息之间相关性的学习，我们提出了一项模态交换对比学习任务，以加强不同模态表示之间的对齐。受近期研究 [54],启发，我们使用模态交换操作构造增强对比视图。具体而言，对于每一对模态 $\langle m_1, m_2 \rangle$, 其中 $m_1 \neq m_2$和 $m_1, m_2 \in \{id, t, v\}$，我们通过在两种模态之间随机交换一半的序列物品 $\bar{S}^{m_1/m_2} = \{\bar{\mathbf{e}}_{i_{m_1/m_2,1}}, \ldots\ldots, \bar{\mathbf{e}}_{i_{m_1/m_2,|S|}}\}$，构造一对模态交换对比视图 $\langle \mathbf{Z}^{m_1}_{aug}, \mathbf{Z}^{m_2}_{aug} \rangle$:

$$\mathbf{Z}^{m_1/m_2}_{aug} = \text{Transformer}(\text{Dropout}(\text{DyT}(\bar{S}^{m_1/m_2}_{aug}))), \tag{16}$$

$$\bar{S}^{m_1}_{aug}, \bar{S}^{m_2}_{aug} = \text{Swap}(\bar{S}^{m_1}, \bar{S}^{m_2} | 0.5), \tag{17}$$

其中，Swap($S_1, S_2|\rho$)表示以 ρ的概率在序列 S_1 和 S_2 之间随机交换物品表示。为更好地对齐不同模态的表示，我们在增强对比视图之间采用 InfoNCE [55] 。具体而言，在一个训练批次 $\mathcal{B}$中，我们可以得到 $\{\mathbf{Z}^{m_1/m_2}_{aug\,1}, \mathbf{Z}^{m_1/m_2}_{aug\,2}, \ldots, \mathbf{Z}^{m_1/m_2}_{aug,|\mathcal{B}|}\}$，并将 $\langle \mathbf{Z}^{m_1}_{aug,i}, \mathbf{Z}^{m_2}_{aug,i} \rangle$ 视为正样本对，而将 $\{\langle \mathbf{Z}^{m_1}_{aug,i}, \mathbf{Z}^{m_2}_{aug,j} \rangle | i \neq j\}$ 视为负样本对。对比学习损失可表示为：

$$\mathcal{L}^{m_{1,2}}_{MSCL} = -\sum_{i=1}^{|\mathcal{B}|} \log \frac{\exp(\text{Sim}(\mathbf{Z}^{m_1}_{aug,i}, \mathbf{Z}^{m_2}_{aug,i}))}{\sum_{j=1}^{|\mathcal{B}|} \exp(\text{Sim}(\mathbf{Z}^{m_1}_{aug,i}, \mathbf{Z}^{m_2}_{aug,j}))}, \tag{18}$$

其中 Sim(·) 表示余弦相似度。在**消融研究**部分，我们通过实证分析评估不同模态配对选择的有效性。

E. 优化

我们使用两个目标来优化模型：模态交换对比学习以及主要的序列推荐任务。给定用户行为序列 S，序列推荐的首要目标是预测用户在第（$|S|$+1）个时间步可能交互的物品。利用我们提出的模型，我们得到预测分数$\hat{y}_i$，它表示用户在下一时间步与物品i 交互的概率。随后采用交叉熵损失函数作为主要训练目标，以量化预测分数$\hat{y}_i$ 与真实值y_i之间的差异，定义为：

$$\mathcal{L}_{SR} = -\sum_{i \in \mathcal{I}} y_i \log(\hat{y}_i). \tag{19}$$

最终学习目标定义为：

$$\mathcal{L} = \mathcal{L}_{SR} + \lambda \mathcal{L}_{MSCL}, \tag{20}$$

其中，λ 是权重超参数。

为更清晰地概述 CAMMSR，我们在算法1中总结了 CAMMSR 的学习过程。

算法1： CAMMSR 的流程

输入： 用户集合 $\mathcal{U}$、物品集合 $\mathcal{I}$、模态 $\mathcal{M} = \{id, t, v\}$。
初始化物品 ID 表示 I_{id}、物品文本表示 f_{i_t}、物品视觉表示 f_{i_v}，以及每种模态的完整表示 $m$as $\mathbf{E}_m$;
while尚未收敛 **do**
  获取表示 $\hat{\mathbf{e}}_{i_m}$ 对于每种模态 m，并使用通过公式 5-7 计算注意力专家混合。
  计算物品 i在 w^m_i上的类别引导权重模态 m，通过公式 8-10。
  获取物品的最终表示 $\bar{\mathbf{e}}_{i_m}$ 针对物品 i在模态 m，通过公式 11。
  通过公式 13-14 获取序列表示 $\mathbf{Z}^m$ 。
  计算模态交换对比学习损失 $\mathcal{L}_{MSCL}$ 通过公式 16-18。
  计算主要序列推荐损失 $\mathcal{L}_{SR}$ 通过公式 15-19。
  通过公式 20 获取最终优化损失 $\mathcal{L}$ 。
  通过损失函数优化模型 $\mathcal{L}$。
结束循环

,

V. 实验

本节通过全面的实验，评估 CAMMSR 框架在四个广泛使用的真实世界数据集上的性能。实验结果可以很好地回答以下五个研究问题：**RQ1:** CAMMSR 是否优于最先进的传统序列推荐和多模态序列推荐方法?
RQ2:CAMMSR 的关键组件对其整体性能有何影响?
RQ3:类别噪声如何影响 CAMMSR 的性能？**RQ4:** 利用多模态是否有助于 CAMMSR 获得更理想的性能？**RQ5:** 与其他多模态序列推荐方法相比，CAMMSR 的效率如何?
RQ6: CAMMSR 的类别引导机制如何有效分配模态权重?
RQ7: 不同的超参数设置会如何影响 CAMMSR 的整体性能?

总之，我们在 V-A 节介绍实验设置，在 V-B 节展示 CAMMSR 的性能优势，在 V-C 节评估关键组件的必要性，在 V-D 节分析 CAMMSR 在类别噪声场景下的表现，在 V-E 节讨论多模态的重要性，在 V-F 节分析 CAMMSR 的效率，在 V-G 节提供真实世界案例研究，在 V-H 节进行超参数分析，并在 V-K 节将 CAMMSR 扩展至不同的类别标签场景。

表 II：四个数据集的统计信息。

数据集	Toys	游戏	美妆	Home
# 用户	19,412	24,303	22,363	66,520
# 物品	11,924	10,672	12,101	28,238
# 交互次数	167,597	231,780	198,502	551,682
# 平均每位用户交互次数	14.58	9.53	8.88	8.29
# 平均每个项目交互次数	17.03	21.72	16.40	19.54
# 稀疏度	99.93%	99.91%	99.93%	99.97%

A. 实验设置

1）数据集：为了验证 CAMMSR 的有效性，我们在四个公开的真实世界数据集上开展实验[56],，分别是**玩具和游戏（玩具）**、**电子游戏（游戏）**、**美妆**和**家居与厨房（家居）**。对于每个数据集，我们移除重复交互，并按时间戳对用户交互进行时间排序，以构建行为序列。遵循先前研究 [57], [36], ，我们采用 5-core过滤策略，从每个数据集中排除交互次数少于五次的用户和物品。对于文本数据，我们按照先前研究 [20], [51]的做法，将标题、类别和品牌字段中的短语拼接起来。对于图像数据，我们直接从提供的商品 URL 下载每件物品的首张图像。表 II 展示了这四个数据集的详细统计信息。

2）基线方法：为评估 CAMMSR 的有效性，我们将 CAMMSR 与多种最先进的基线方法进行比较：

（1）序列方法

- **GRU4Rec** [26] 利用门控循环单元（GRU）对会话推荐中的用户点击序列建模。遵循以往研究 [21], [10], ，我们使用嵌入向量而非独热向量表示物品。

- **SASRec** [11] 是一种基于自注意力的序列推荐模型，通过多头注意力机制预测下一个物品。

- **LRURec** [58] 引入线性循环单元，以提升对长用户行为序列的编码质量。
- **FEARec** [13] 通过引入斜坡结构进行频率级序列分析，从而有效处理频域信息。

- **TiCoSeRec** [12] 提出五种基于时间间隔的序列级增强方法，以创建均匀序列，并利用对比学习提升性能。

（2）多模态序列方法

- **NOVA** [17]提出了一种非侵入式自注意力机制，以有效利用辅助信息。
- **DIF-SR** [51] 引入了一种解耦的辅助信息融合模块，以解决秩瓶颈问题，并增强非侵入式注意力矩阵的表达能力。

- **UniSRec** [36]利用物品文本信息学习更具迁移性的表示，以进行序列推荐。
- **MISSRec** [20]提出了一个兴趣发现模块，用于捕捉物品之间的深层关系以及用户偏好。

- **M3SRec** [21] 采用特定模态和跨模态混合专家机制，以增强 Transformer 中的模态学习。

- **IISAN** [25] 在模态基础模型上利用解耦的参数高效微调，同时实现模态内适配和模态间适配。

- **MMMLP** [22] 提出了一种纯粹基于多层感知机的模型，利用 FeatureMixer 和 Fusion Mixer 处理多模态数据，以线性复杂度实现了出色的性能。
- **TedRec** [23] 在频域中使用快速傅里叶变换，对文本和 ID 进行序列级语义融合。
- **FindRec** [38] 通过基于 Stein 核的集成信息协调模块，从理论上保证了跨模态一致性。

3) 评估指标： 为公平评估推荐任务的性能，我们遵循先前研究 [22],[20], 采用两种成熟的评估指标：归一化折损累计增益 (NDCG@K) 和平均倒数排名 (MRR@K)，其中 K $\in \{5, 10\}$。我们采用留一评估策略开展实验。排名分数是在整个物品集合上计算的，不进行采样。

4）实现细节：为确保公平比较，我们使用 RecBole [59]实现 CAMMSR 和所有基线方法。RecBole 是一个统一且全面的推荐系统框架，并实现了多种推荐方法。我们将训练批次大小设为 1024，并将所有基线方法的隐藏嵌入维度设为 64。每个行为序列的最大长度限制为 50。对于编码器结构，Transformer 的多头数量和自注意力层数均根据经验设为 2。具体而言，我们采用相同的 Adam [60] 优化器和默认参数进行 Xavier 初始化 [61] 。我们对所有基线模型进行了详细的超参数搜索。CA MoE 组件的专家数量 K 在$\{1, 2, 4, 8, 16\}$ 范围内进行搜索，权重超参数 λ 在$\{0.25, 0.5, 0.75, 1\}$范围内进行搜索。此外，我们根据经验将 $1e^{-3}$固定为学习率。所有实验均在单个 NVIDIA GeForce RTX 3090 图形处理器上进行。

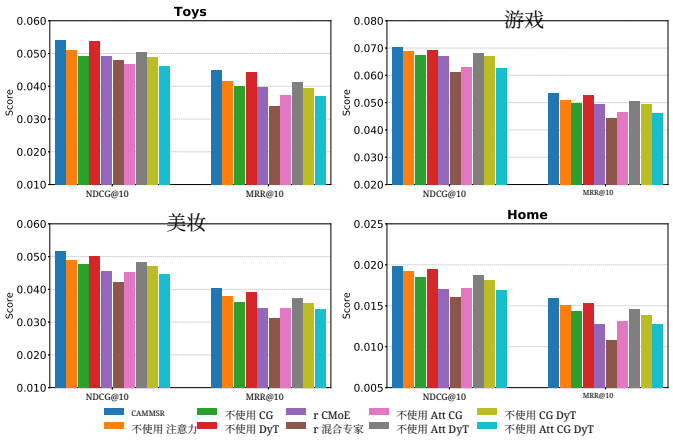


图 2：CAMMSR 及其所有变体在全部四个数据集上的性能比较。

表 III：不同数据集上各基线方法在 NDCG@K 和 MRR@K 指标上的性能比较。上标 * 表示改进具有统计显著性，即 p 值小于 0.05。“负向比例”表示测试集中移除类别引导后推荐准确率有所提升的比例。

基线	Toys				游戏				美妆				Home			
	NDCG		MRR		NDCG		MRR		NDCG		MRR		NDCG		MRR	
	@5	@10	@5	@10	@5	@10	@5	@10	@5	@10	@5	@10	@5	@10	@5	@10
GRU4Rec	0.0236	0.0289	0.0201	0.0222	0.0385	0.0514	0.0311	0.0365	0.0271	0.0337	0.0223	0.0260	0.0067	0.0087	0.0056	0.0064
SASRec	0.0348	0.0411	0.0257	0.0295	0.0391	0.0546	0.0286	0.0349	0.0325	0.0412	0.0250	0.0286	0.0118	0.0148	0.0089	0.0100
LRURec	0.0362	0.0446	0.0278	0.0312	0.0410	0.0557	0.0315	0.0376	0.0323	0.0412	0.0250	0.0286	0.0112	0.0141	0.0084	0.0096
FEARec	0.0354	0.0442	0.0266	0.0302	0.0411	0.0558	0.0306	0.0367	0.0342	0.0434	0.0265	0.0304	0.0121	0.0152	0.0092	0.0105
TiCoSeRec	0.0336	0.0404	0.0280	0.0311	0.0463	0.0589	0.0382	0.0434	0.0335	0.0413	0.0285	0.0317	0.0108	0.0134	0.0091	0.0102
NOVA	0.0381	0.0478	0.0288	0.0328	0.0417	0.0576	0.0303	0.0368	0.0347	0.0442	0.0273	0.0312	0.0120	0.0147	0.0092	0.0102
DIF-SR	0.0359	0.0444	0.0284	0.0318	0.0416	0.0575	0.0311	0.0375	0.0340	0.0436	0.0262	0.0302	0.0092	0.0112	0.0078	0.0086
UniSRec	0.0276	0.0384	0.0194	0.0239	0.0386	0.0535	0.0291	0.0353	0.0287	0.0393	0.0212	0.0288	0.0121	0.0160	0.0092	0.0108
MISSRec	0.0323	0.0414	0.0235	0.0270	0.0398	0.0506	0.0312	0.0353	0.0315	0.0397	0.0240	0.0271	0.0140	0.0174	0.0107	0.0119
M3SRec	0.0416	0.0486	0.0363	0.0391	0.0493	0.0643	0.0404	0.0464	0.0345	0.0428	0.0294	0.0328	0.0122	0.0144	0.0105	0.0115
IISAN	0.0395	0.0489	0.0354	0.0393	0.0525	0.0671	0.0432	0.0491	0.0372	0.0464	0.0309	0.0347	0.0152	0.0187	0.0129	0.0142
MMMLP	0.0378	0.0466	0.0344	0.0377	0.0486	0.0644	0.0397	0.0464	0.0351	0.0433	0.0295	0.0333	0.0133	0.0165	0.0107	0.0118
TedRec	0.0318	0.0397	0.0265	0.0297	0.0468	0.0604	0.0384	0.0439	0.0330	0.0419	0.0275	0.0311	0.0113	0.0140	0.0094	0.0105
FindRec	0.0390	0.0480	0.0357	0.0389	0.0501	0.0663	0.0411	0.0477	0.0359	0.0447	0.0304	0.0341	0.0137	0.0171	0.0111	0.0124
CAMMSR	**0.0477***	**0.0542***	**0.0421***	**0.0449***	**0.0560***	**0.0701***	**0.0474***	**0.0533***	**0.0429***	**0.0517***	**0.0374***	**0.0404***	**0.0168***	**0.0199***	**0.0149***	**0.0159***
改进	14.66%	10.84%	15.98%	14.25%	6.67%	4.47%	9.72%	8.55%	15.32%	11.42%	21.04%	16.43%	10.53%	6.42%	15.50%	11.97%
负向比例	0.16%	0.16%	0.24%	0.24%	0.33%	0.33%	0.61%	0.61%	0.24%	0.24%	0.40%	0.40%	0.28%	0.28%	0.54%	0.54%

表IV：CAMMSR 在不同模态配对选择下的有效性。○ 表示所选配对。

数据集	$\langle v, t\rangle$	$\langle id,\ t\rangle$	$\langle id,\ v\rangle$	NDCG@5	NDCG@10	MRR@5	MRR@10
Toys	○	○	○	0.0415	0.0504	0.0371	0.0415
	●	○	○	0.0435	0.0530	0.0396	0.0439
	○	●	○	0.0424	0.0516	0.0382	0.0422
	○	○	●	0.0421	0.0513	0.0380	0.0421
	●	●	○	0.0439	0.0536	0.0406	0.0444
	●	○	●	0.0439	0.0538	0.0407	0.0446
	○	●	●	0.0430	0.0524	0.0387	0.0427
	●	●	●	**0.0447**	**0.0542**	**0.0421**	**0.0449**
游戏	○	○	○	0.0529	0.0679	0.0436	0.0501
	●	○	○	0.0541	0.0690	0.0455	0.0525
	○	●	○	0.0535	0.0686	0.0442	0.0508
	○	○	●	0.0533	0.0684	0.0440	0.0505
	●	●	○	0.0550	0.0693	0.0462	0.0527
	●	○	●	0.0552	0.0696	0.0465	0.0529
	○	●	●	0.0545	0.0689	0.0449	0.0515
	●	●	●	**0.0560**	**0.0701**	**0.0474**	**0.0533**
美妆	○	○	○	0.0374	0.0469	0.0312	0.0350
	●	○	○	0.0414	0.0503	0.0367	0.0392
	○	●	○	0.0383	0.0479	0.0330	0.0365
	○	○	●	0.0381	0.0479	0.0328	0.0363
	●	●	○	0.0417	0.0506	0.0367	0.0394
	●	○	●	0.0418	0.0509	0.0369	0.0396
	○	●	●	0.0388	0.0492	0.0343	0.0378
	●	●	●	**0.0429**	**0.0517**	**0.0374**	**0.0404**
Home	○	○	○	0.0152	0.0185	0.0128	0.0142
	●	○	○	0.0162	0.0191	0.0141	0.0152
	○	●	○	0.0155	0.0188	0.0131	0.0146
	○	○	●	0.0155	0.0187	0.0133	0.0145
	●	●	○	0.0163	0.0193	0.0142	0.0153
	●	○	●	0.0165	0.0195	0.0147	0.0156
	○	●	●	0.0160	0.0190	0.0140	0.0148
	●	●	●	**0.0168**	**0.0199**	**0.0149**	**0.0159**

B. 总体比较（*RQ1*）

表 III 展示了所有基线模型在四个公开可用数据集上的全面推荐性能。最佳结果以粗体标出，次佳结果以下划线标出。我们还报告了所提出的 CAMMSR 所取得的相对提升（%Improv.），并通过 t 检验实验验证了统计显著性（$p < 0.05$），如上标 $*$所示。基于这些实验结果，我们给出以下详细观察：

• CAMMSR 在所有评估指标和数据集上均展现出稳定且具有统计显著性的优势。具体而言，在 Toys 数据集上，与最强基线相比，其 NDCG@5 相对提升了 14.66%，MRR@5 提升了 15.98%。在其他数据集上也观察到类似的显著提升：Games 数据集上的 NDCG@5 提升了 6.67%，Beauty 数据集上提升了 15.32%，Home 数据集上提升了 10.53%。这一稳定的性能优势验证了我们的类别引导注意力混合专家框架结合模态交换对比学习，能够有效挖掘具有区分性的模态特征，并利用模态间的协同效应进行序列推荐。

• 比较分析表明，在大多数实验设置下，多模态序列方法通常优于传统序列方法。例如，在 Beauty数据集上，所有多模态基线方法的表现都大幅超过传统方法，其中表现最强的多模态基线方法 IISAN 的 NDCG@5 达到 0.0372，而表现最佳的传统方法 FEARec 为 0.0342。值得注意的是，CAMMSR 进一步突破了这一性能上限，充分证明了我们专门设计的类别引导注意力MoE组件至关重要：它能够根据上下文信号，为不同模态动态分配适当的权重。

• 结果表明，有效建模模态间关系对推荐性能至关重要。IISAN 和 M3SRec 通过各自专门的模态间建模方法取得了有竞争力的结果（例如，IISAN 在 Games数据集上的 NDCG@5 达到 0.0525），而 CAMMSR 在所有数据集和指标上始终优于它们。其性能优势在 Beauty数据集中尤为明显：CAMMSR 的 NDCG@5 比 IISAN 提高了 15.32 %，这表明我们的类别引导注意力MoE 能够通过自适应专家组合和对比对齐，更有效地利用不同模态之间的协同效应。

•尽管类别-模态相关性与推荐效用不一致的极端情况可能偶然发生，但我们通过分析测试集中移除

CAMoE 后推荐准确率得到提升的比例，以实证方式回应了这一问题。实验结果表明，CAMoE 在不到 1% 的情况下会产生负面影响。这证明了其较强的泛化能力和稳定性。从性能提升的角度来看，与功能需求较高的场景（如游戏和家居）相比，CAMMSR 在功能需求较低的场景（如玩具和美妆）中展现出更显著的性能优势。

总体而言，全面的实验结果证实，CAMMSR 在推荐性能方面取得了统计显著且具有实际意义的提升。

C. 消融研究（*RQ2*）

为了分析 CAMMSR 的有效性，我们进行了全面的消融研究，以评估模型中各个组件的必要性及其贡献。具体而言，我们将模型与以下变体进行比较：a) **w/o 注意力** 将公式 5 中的注意力门控路由器替换为线性门控路由器。b) **w/o 类别引导** 将公式 10 中的类别引导融合替换为等权融合。c) **w/o DyT**将公式 14 中的 DyT层替换为 LayerNorm层。d) **r CMoE** 将 CAMoE 替换为先前工作 [21] 中的跨模态 MoE。e) **r MoE** 将 CAMoE 替换为原始MoE。评估结果如图2所示。所有组件均有助于提升 CAMMSR 的性能。具体而言，我们得到以下观察结果：1) CAMMSR 的性能优于 **w/o 注意力**，这表明为混合专家中的不同专家赋予适当权重是必要的。2) **w/o 类别引导**的性能显著低于 CAMMSR，这表明根据类别预测任务中不同模态的表现分配权重，能够有效指导如何利用多模态特征。3) CAMMSR 的性能显著优于 **w/o DyT**，这表明与 Transformer 中的 LayerNorm层相比，DyT层能够取得相当甚至更优的性能。这一发现与先前研究 [53] 报告的观察结果一致。4) CAMMSR 的表现始终优于**r CMoE**和**r MoE**，证明了 CAMoE 的有效性。此外，我们进一步研究了模态交换对比学习组件中不同模态对的选择对 CAMMSR 整体性能的影响。表IV展示了研究结果。我们观察到，纳入更多模态对能够带来更高的性能。仅选择一个模态对时，视觉模态与文本模态的组合效果最佳。我们推测，这是因为视觉模态与文本模态之间存在显著的语义差距，而模态交换对比学习组件能够有效缩小这一差距，从而更好地利用多模态特征。

D. 噪声研究（*RQ3*）

为评估我们的类别引导机制在实际数据质量场景下的鲁棒性，我们开展了一项

表V：类别噪声的深度分析。

Type	级别	Toys		游戏	
		NDCG@10	MRR@10	NDCG@10	MRR@10
CAMMSR	-	0.0542	0.0449	0.0701	0.0533
移除	10%	0.0530	0.0439	0.0693	0.0524
	20%	0.0520	0.0428	0.0689	0.0519
	30%	0.0513	0.0420	0.0684	0.0515
误导	10%	0.0525	0.0433	0.0688	0.0519
	20%	0.0505	0.0411	0.0678	0.0504
	30%	0.0482	0.0391	0.0667	0.0490
w/o 类别引导	-	0.0493	0.0400	0.0675	0.0499

Type	级别	美妆		Home	
		NDCG@10	MRR@10	NDCG@10	MRR@10
CAMMSR	-	0.0517	0.0404	0.0199	0.0159
移除	10%	0.0505	0.0391	0.0195	0.0156
	20%	0.0499	0.0384	0.0193	0.0152
	30%	0.0492	0.0377	0.0190	0.0148
误导	10%	0.0500	0.0386	0.0191	0.0150
	20%	0.0480	0.0368	0.0186	0.0146
	30%	0.0470	0.0357	0.0182	0.0142
w/o 类别引导	-	0.0476	0.0362	0.0185	0.0143

系统性噪声研究，模拟两种常见的类别标签损坏类型：(1) **移除**，即删除损坏的标签，并将相同权重分配给所有模态作为回退方案；(2) **误导**，即将损坏的标签替换为错误标签。作为参考，我们纳入了完全移除类别引导权重分配的w/o 类别引导变体。超参数配置与主要实验保持一致（V-A节）。四个数据集上的实验结果汇总于表V，由此我们得出以下关键观察：

- 在所有噪声级别下，**移除** 条件造成的性能下降始终小于**误导**情景。例如，在含有 30% 噪声的玩具数据集上，移除会使 NDCG@10 下降 5.35%，而误导则会导致其下降 11.07%。这一规律凸显了准确类别信号对于合理分配模态权重的关键作用，也说明多模态系统容易受到误导性监督信号的影响。

- 即使在噪声较大的情况下，我们的类别引导机制仍能保持有效。在 20% 的误导噪声下，CAMMSR 仍取得了优异的性能（玩具数据集上的 NDCG@10 为 0.0505，游戏数据集上为 0.0678），优于 w/o CG 变体（玩具数据集上的 NDCG@10 为 0.0493，游戏数据集上为 0.0675），展现出对标签不准确的显著鲁棒性。这种韧性对于实际部署尤其重要，因为无法保证标签质量完美无缺。

基于这些实证发现，我们对实际影响进行如下分析：

- 在现实世界的数据管理场景中，在维护良好的系统里，类别标签错误超过 10-20% 在统计上十分罕见。我们的结果证实，在这一现实噪声范围内，类别引导组件仍能带来净收益，其性能优于不采用类别引导的消融版本。这

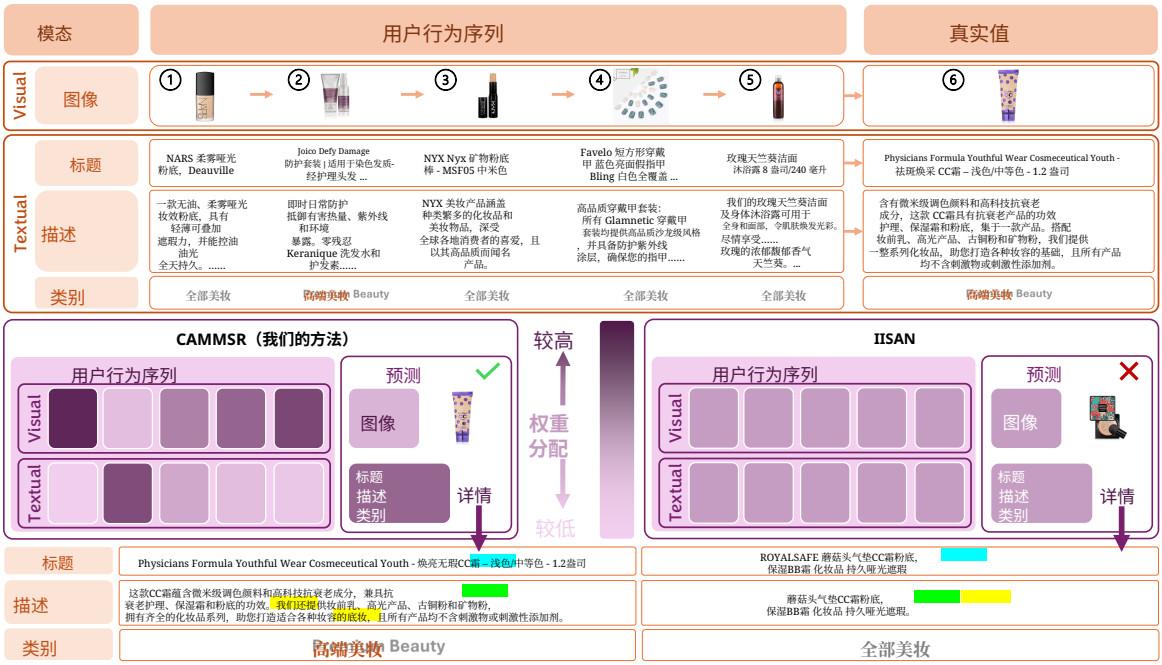
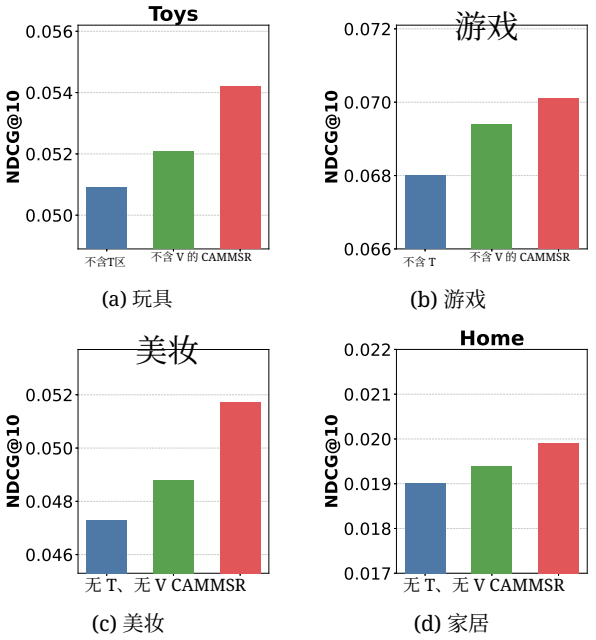


图 3：案例研究 fo r 购买序列以及 CAMMSR 和 IISAN 在美妆数据集上的预测结果 et.



(a) 玩具 (b) 游戏

(c) 美妆 (d) 家居

图 4：移除各个模态的性能比较。

使 CAMMSR 特别适用于实用的被动数据管理系统，因为在这些系统中，完美的标签整理在经济上难以实现。

- 移除与误导性条件之间显著的性能差距（例如，在 30% 噪声的 Toys 数据集上，NDCG@10 分别为 0.0513 和 0.0482）表明，开发有效的标签噪声过滤机制是一个很有前景的研究方向。未来的工作可以着重于整合置信度感知类别估计或自校正标签精炼，以进一步提升噪声环境下的鲁棒性，从而提高被动网络规模应用中数据管理系统的可靠性。

E. 深入分析（*RQ4*）

我们进行了深入分析，以验证利用多模态是否有助于提升推荐的准确性。具体而言，我们设计了以下两个变体：a) **w/o T** 移除文本模态。b) **w/o V**移除视觉模态。

如图 4 所示，CAMMSR 的表现优于所有变体。此外，变体 **w/o V** 优于**w/o T**，这表明文本数据相较于视觉数据具有更高的重要性。这一现象可归因于多模态能够提供更丰富、更全面的信息。此外，与视觉信息相比，文本信息往往与物品特征具有更强的语义对齐，因此与推荐任务的相关性更直接。

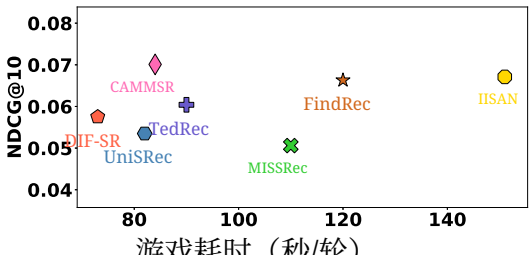
F. 效率研究（*RQ5*）

为进一步分析 CAMMSR 的训练效率，我们在 Games 和 Beauty 数据集上绘制了效率（每轮训练时间）与效果（NDCG@10）的散点图。如图 5 所示，CAMMSR 在效率和性能方面均表现出色；在效率方面，除 DIF-SR 外，它优于所有基线方法；在性能方面，它则显著优于所有基线方法。此外，考虑到总训练时间与模型的收敛速度密切相关，我们还分析了收敛所需训练轮数。在表VI中，我们从每轮平均训练时间、收敛所需轮数和总训练时间三个方面展示了 CAMMSR 与先进基线方法的效率。结果表明，CAMMSR

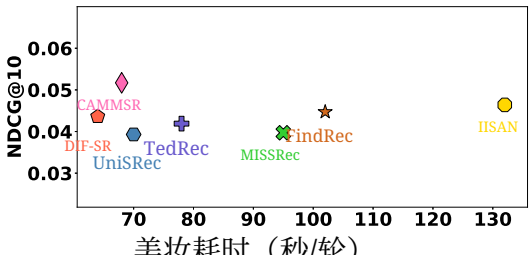
表VI：两个数据集上不同基线方法的效率比较，包括每轮平均训练时间、收敛所需轮数、总训练时间和每轮推理时间（TT/E：每轮训练时间，#E：轮数，TTT：总训练时间，IT/E：每轮推理时间，N@10：NDCG@10；s：秒，m：分钟，h：小时）。

模型	游戏					美妆				
	每轮训练时间↓	#E↓	TTT↓	每轮推理时间↓	N@10↑	每轮训练时间↓	#E↓	TTT↓	每轮推理时间↓	N@10↑
DIF-SR	73s	119	2小时24分47秒	9.2s	0.0575	64s	103	1小时49分52秒	7.9s	0.0436
UniSRec	82s	107	2小时26分14秒	11.7秒	0.0535	70s	95	1小时50分50秒	9.4s	0.0393
MISSRec	110s	89	2小时43分10秒	15.2秒	0.0506	95s	74	1小时47分10秒	10.6秒	0.0397
IISAN	151s	114	4小时46分54秒	23.1秒	0.0671	132s	99	3小时37分48秒	19.8秒	0.0464
TedRec	90s	103	2小时34分30秒	11.5秒	0.0604	78s	88	1小时54分24秒	10.1秒	0.0419
FindRec	120s	90	3小时0分0秒	19.8秒	0.0663	102s	71	2小时0分42秒	16.0秒	0.0447
CAMMSR	84s	65	**1小时31分0秒**	10.4秒	**0.0701**	68s	53	**1小时0分4秒**	8.8s	**0.0517**

每轮训练效率具有竞争力，且在所有模型中收敛速度最快。这表明 CAMMSR 不仅能减少计算时间，还能迅速达到最佳性能，因此尤其适用于时间或计算资源有限的场景，例如快速部署和频繁更新。此外，在推理时间方面，CAMMSR 的效率优于其他基线方法，仅次于 DIF-SR。



(a) 游戏



(b) 美妆

图5：在游戏和美妆数据集上的效率研究。

G. 案例研究（*RQ6*）

为实证验证我们提出的类别引导注意机制的有效性及其在被动信息管理中的实际意义，我们使用美妆数据集中的一个代表性序列开展了详细的案例研究。如图3所示，我们将 CAMMSR 生成的排名第一的推荐结果与 IISAN 的结果进行比较；在该数据集的所有基线方法中，IISAN 的性能最强。虽然两种模型都正确识别出大致的产品类别（CC霜），但只有 CAMMSR 成功推荐了与用户内在偏好模式相符的准确真实物品。

这一出色表现得益于我们的类别引导式注意力混合专家框架，它能够实现更智能的被动信息管理，并通过上下文感知的模态加权加以实现。我们的分析表明，真实物品和历史序列中的第二个物品都属于“高端美妆”类别。通过辅助类别预测任务，CAMMSR 学会了，对于高价值化妆品，文本描述——其中包含详细的成分列表、使用说明和质量保证——比单纯的视觉外观更具区分力。这一发现与现实中的消费者行为相呼应：高端美妆产品的购买者会在做决定前仔细查看文本细节。

因此，在处理目标序列时，CAMMSR 通过其注意力门控机制动态增强文本信号的贡献，从而有效过滤视觉上相似但语义不匹配的备选项。这种自适应模态加权展示了我们的框架如何作为一种先进的被动数据管理系统运行：它会根据物品类别和上下文模式，自动辨别并优先处理最相关的信息维度，从而减轻用户浏览复杂多模态内容时的认知负担。相比之下，IISAN 的静态融合策略无法捕捉这些细微的类别相关模态偏好，因此推荐了一个视觉上相似、但与上下文不符的备选项。

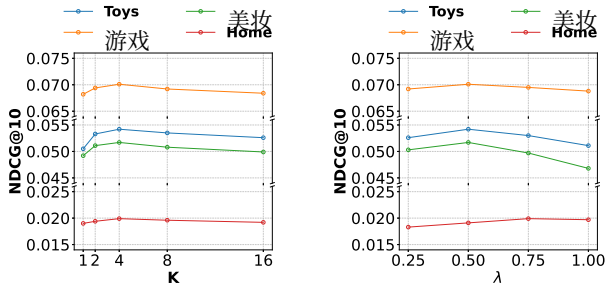
该案例研究具体展示了 CAMMSR 如何超越传统融合策略，推进多模态序列推荐。通过动态调整模态重点，使其与类别特定的决策模式相匹配，CAMMSR 为被动信息管理提供了更为 sophisticated 的方法——自动呈现最相关的内容，同时过滤外围信号，最终提供更加精准且符合上下文的推荐。

H. 超参数分析（*RQ7*）

我们进一步探究了关键超参数对 CAMMSR 性能的影响。以 NDCG@10 衡量的评估结果如图 6 所示。

I. 关于 $w.r.t.\ K$的性能比较：

我们详细研究了专家数量对模型性能的影响，如图 6(b) 所示。



(a) 超参数 K (b) 超参数 λ

图 6：关键超参数（k 和 λ）对 NDCG@10 的性能比较 $w.r.t.$。

拥有 4 个专家的模型在所有评估指标上均取得最佳结果，优于拥有 1 个或 2 个专家的变体。这一提升表明，1 个或 2 个专家不足以捕捉跨模态交互中复杂多样的模式，而 4 个专家则能在提供充足的专门处理能力与避免显著增加计算开销之间取得平衡。然而，将专家数量增加到 8 个或 16 个会导致性能下降。这种性能退化可能是由于过度专门化所致，它会引入冗余并增加路由机制的复杂性。

J. 性能比较 $w.r.t.$ λ:

我们通过实证分析了权重超参数 λ 对 CAMMSR 模型的影响。如图6(a)所示，大多数数据集的推荐值为 $\lambda = 0.25$，但 Home数据集除外，其最优设置为 $\lambda = 0.5$。我们认为，这种差异源于 Home数据集极高的稀疏度（99.97%），这使得在数据稀缺环境中，更需要通过更强烈的对比学习信号来加强跨模态对齐，以弥合模态间的差距。

这种对超参数选择的鲁棒性显著提升了我们框架在现实数据管理场景中的实际部署能力，因为它减少了多模态系统通常所需的大量参数调优工作，同时保持可靠的推荐质量。观察到的稳定性符合工业级推荐系统的要求；对于这类系统而言，运行一致性和最低限度的维护开销是确保长期可持续部署的关键考量。

K. 类别分析

我们进一步探讨类别引导融合（CG）在多标签类别、层级类别和无类别场景中的应用。对于多标签类别和层级类别场景，在多个差异 d 上采用不同的聚合策略（均值/最大值/最小值池化），是一种简单而有效的解决方案。在无类别场景中，利用大语言模型构建虚拟类别展现出良好的应用前景。具体而言，对于涉及多标签类别和层级类别的场景，我们采用最大池化。此外，我们使用 GPT-4o 生成虚拟类别

表VII：CAMMSR 与**不使用 CG** 变体在多标签类别和无类别场景下的性能。

Baby	NDCG@5	NDCG@10	MRR@5	MRR@10
多标签类别	64.09	70.28	55.29	61.03
无类别	62.88	68.05	53.11	58.66
不使用 CG	57.13	62.20	44.13	48.70
Microlens	**NDCG@5**	**NDCG@10**	**MRR@5**	**MRR@10**
无类别（10）	61.40	65.51	52.59	56.91
无类别（20）	61.52	65.48	52.45	57.12
不使用 CG	57.83	62.21	46.00	49.45

，并根据所有物品的文本和视觉信息生成相应类别。为评估这一方法，我们在包含多标签类别的 Amazon Baby 数据集 [62], [63]，上开展实验。对于不存在现有类别的场景，我们移除所有原始类别，并使用与原始数据集相同的词汇表生成虚拟类别。此外，为验证 CG 的泛化能力，我们进一步在 Microlens 数据集 [64] 上开展评估。具体而言，我们使用 GPT-4o 生成 10 个或 20 个虚拟类别。表 VII 显示，CG 在所有场景中均能持续提升性能。对于 Microlens 数据集，与**不使用 CG**的版本相比，使用 10 个和 20 个虚拟类别均带来了稳定且相近的性能提升。

VI. 结论

本文提出了 CAMMSR，这是一种面向多模态序列推荐的新型类别引导注意力混合专家框架。为了解决现有多模态模型中静态融合策略固有的局限性，CAMMSR 全面整合了用户行为序列中的多模态内容及其对应的时间戳，从而生成更加细致且具备时间感知能力的推荐结果。该框架的核心是类别引导注意力混合专家（CAMoE）模块，它不仅能从多个角度提取专门化物品表示，还能显式建模不同模态之间的协同交互。通过辅助类别预测任务动态分配模态权重，我们的模型实现了多模态信号的自适应融合，使其同时与物品特征和用户情境偏好相契合。此外，我们提出的模态交换对比学习任务通过序列级增强策略，有效加强了跨模态对齐并提升了表示一致性。在四个真实世界数据集上开展的大量实验，证明了所提出的 CAMMSR 的有效性和效率。

展望未来，CAMMSR 为推动自适应多模态推荐系统的发展奠定了灵活的基础。它能够动态整合多模态信号，同时保持跨模态一致性，为多项有意义的拓展提供了支持，例如将类别引导推广至其他情境维度、开发更精细的协同建模方法，以及将该框架应用于传统推荐场景之外的序列决策任务。

七、人工智能的使用

我们仅有限度地使用大语言模型（LLMs）辅助写作，包括语法纠正、文风润色以及表格布局与格式调整。所有建议的修改均由作者人工审阅，并有选择地采纳。所有科学内容、观点、分析和结论均完全由我们提出。作者对本文全部内容承担全部责任，包括可能仍然存在的任何错误或不准确之处。

参考文献

[1] J. Xu、Z. Chen、W. Wang、X. Hu、S.-W. Kim 和 E. C. Ngai，“Cohesion：用于多模态推荐的双阶段融合复合图卷积网络”，载于第*48*届 *ACM* 信息检索研究与发展国际会议论文集，2025年，第1830–1839页。[2] X. He、K. Deng、X. Wang、Y. Li、Y. Zhang 和 M. Wang，“Lightgcn：简化并增强用于推荐的图卷积网络”，载于第*43*届 *ACM* 信息检索研究与发展国际会议论文集，2020年，第639–648页。[3] J. Xu、Z. Chen、S. Yang、J. Li 和 E. C. Ngai，“最好的尚未来临：用于多模态推荐的测试阶段图卷积”，载于第*33*届 *ACM* 多媒体国际会议论文集，2025年，第6325–6334页。[4] Z. Guo、J. Li、G. Li、C. Wang、S. Shi 和 B. Ruan，“Lgmrec：用于多模态推荐的局部与全局图学习”，载于*AAAI* 人工智能大会论文集，第38卷，第8期，2024年，第8454–8462页。[5] J. Xu、Z. Chen、S. Yang、J. Li、Z. Wan、H. Wang、W. Liu、Y. Li 和 E. C. Ngai，“Vi-mmrec：用于多模态推荐的相似性感知、无需训练成本的虚拟用户-物品交互”，*arXiv* 预印本 *arXiv:2512.08702*，2025年。[6] Z. Chen、J. Xu、H. Wang、S. Yang、Z. Wan 和 H. Hu，“超复数提示感知多模态推荐”，载于第*34*届 *ACM* 信息与知识管理国际会议论文集，2025年，第403–414页。[7] Z. Chen、J. Xu、Y. Wei 和 Z. Peng，“挤压与激励：一种用于协同过滤的加权图对比学习方法”，载于第*48*届 *ACM* 信息检索研究与发展国际会议论文集，2025年，第2769–2773页。[8] J. Xu、Z. Chen、S. Yang、J. Li、H. Wang、W. Wang、X. Hu 和 E. Ngai，“Nlgcl：利用自然存在的邻居层进行推荐图对比学习”，载于第十九届 *ACM* 推荐系统会议论文集，2025年，第319–329页。[9] J. Xu、Z. Chen、J. Li、S. Yang、W. Wang、H. Wang、Y. Li、X. Hu 和 E. Ngai，“Dggvae：用于群体推荐的双粒度图变分自编码器”，*ACM* 信息系统汇刊，第44卷，第2期，第1–30页，2026年。[10] F. Sun、J. Liu、J. Wu、C. Pei、X. Lin、W. Ou 和 P. Jiang，“Bert4rec：基于来自 Transformer 的双向编码器表示的序列推荐”，载于第*28*届 *ACM* 信息与知识管理国际会议论文集，2019年，第1441–1450页。[11] W.-C. Kang 和 J. McAuley，“自注意力序列推荐”，载于*2018*年 *IEEE* 国际数据挖掘会议（*ICDM*）。IEEE，2018年，第197–206页。[12] Y. Dang、E. Yang、G. Guo、L. Jiang、X. Wang、X. Xu、Q. Sun 和 H. Liu，“TiCoSeRec：通过时间间隔扩充数据以生成均匀序列，从而实现有效推荐”，*IEEE*知识与数据工程汇刊，第36卷，第6期，第2686–2700页，2023年。[13] X. Du、H. Yuan、P. Zhao、J. Qu、F. Zhuang、G. Liu、Y. Liu 和 V. S. Sheng，“用于序列推荐的频率增强混合注意力网络”，载于第*46*届 *ACM* 信息检索研究与发展国际会议论文集，2023年，第78–88页。[14] J. Xu、Z. Chen、S. Yang、J. Li、H. Wang 和 E. C. Ngai，“Mentor：用于多模态推荐的多层次自监督学习”，载于*AAAI* 人工智能大会论文集，第39卷，第12期，2025年，第12 908–12 917页。

[15] Z. Chen、J. Xu 和 H. Hu，“Don’t lose yourself：通过减少图卷积网络中的节点-邻居差异提升多模态推荐”，载于 *ICASSP 2025-2025 IEEE* 声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2025，第 1–5 页。[16] J. Xu、Z. Chen、J. Li、S. Yang、W. Wang、X. Hu、R. C.-W. Wong 和 E. C. Ngai，“通过锐度感知最小化提升多模态推荐系统的鲁棒性和泛化能力”，*IEEE*知识与数据工程汇刊，2025。[17] C. Liu、X. Li、G. Cai、Z. Dong、H. Zhu 和 L. Shang，“用于序列推荐中侧信息融合的非侵入式自注意力”，载于 *AAAI* 人工智能大会论文集，第 35 卷，第 5 期，2021，第 4249–4256 页。[18] W. Wei、C. Huang、L. Xia 和 C. Zhang，“用于推荐的多模态自监督学习”，载于 *ACM*万维网会议论文集 *2023*，2023，第 790–800 页。[19] J. Xu、Z. Chen、W. Wang、X. Hu、J. Liu 和 E. C. Ngai，“Lobster：用于多媒体推荐的双向全局语义增强”，信息融合，第 103778 页，2025。[20] J. Wang、Z. Zeng、Y. Wang、Y. Wang、X. Lu、T. Li、J. Yuan、R. Zhang、H.-T. Zheng 和 S.-T. Xia，“Missrec：用于推荐的多模态兴趣感知序列表示的预训练与迁移”，载于第 *31* 届 *ACM* 多媒体国际会议论文集，2023，第 6548–6557 页。[21] S. Bian、X. Pan、W. X. Zhao、J. Wang、C. Wang 和 J.-R. Wen，“用于序列推荐的多模态混合专家表示学习”，载于 *ACM*信息与知识管理国际会议论文集，2023，第 110–119 页。[22] J. Liang、X. Zhao、M. Li、Z. Zhang、W. Wang、H. Liu 和 Z. Liu，“MMMLP：用于序列推荐的多模态多层感知机”，载于 *ACM*万维网会议论文集 *2023*，2023，第 1109–1117 页。[23] L. Xu、Z. Tian、B. Li、J. Zhang、D. Wang、H. Wang、J. Wang、S. Chen 和 W. X. Zhao，“面向推荐系统的序列级语义表示融合”，载于 *ACM*信息与知识管理国际会议论文集，2024，第 5015–5022 页。[24] Z. Zhang、S. Liu、J. Yu、Q. Cai、X. Zhao、C. Zhang、Z. Liu、Q. Liu、H. Zhao、L. Hu 等，“M3oE：多领域多任务混合专家推荐框架”，载于*ACM*信息检索研究与发展国际会议论文集，2024，第 893–902 页。[25] J. Fu、X. Ge、X. Xin、A. Karatzoglou、I. Arapakis、J. Wang 和 J. M. Jose，“IISAN：通过解耦参数高效微调高效适配用于序列推荐的多模态表示”，载于 *ACM*信息检索研究与发展国际会议论文集，2024，第 687–697 页。[26] B. Hidasi、A. Karatzoglou、L. Baltrunas 和 D. Tikk，“基于会话的循环神经网络推荐”，*arXiv* 预印本 *arXiv:1511.06939*，2015。[27] J. Tang 和 K. Wang，“通过卷积序列嵌入实现个性化Top-N序列推荐”，载于第十一届 *ACM* 网络搜索与数据挖掘国际会议论文集，2018，第 565–573 页。[28] J. Li、Y. Wang 和 J. McAuley，“面向序列推荐的时间间隔感知自注意力”，载于 第十三届网络搜索与数据挖掘国际会议论文集，2020，第 322–330 页。[29] S. M. Cho、E. Park 和 S. Yoo，“MEANTIME：融合多时间嵌入的注意力机制混合模型，用于序列推荐”，载于 第十四届 *ACM* 推荐系统会议论文集，2020，第 515–520 页。[30] S. Zhang、L. Chen、C. Wang、S. Li 和 H. Xiong，“用于序列推荐的时序图对比学习”，载于 *AAAI* 人工智能大会论文集，第 38 卷，第 8 期，2024，第 9359–9367 页。[31] Z. Yuan、F. Yuan、Y. Song、Y. Li、J. Fu、F. Yang、Y. Pan 和 Y. Ni，“推荐系统下一步走向何方？重新审视基于 ID 与基于模态的推荐模型”，载于 *ACM*信息检索研究与发展国际会议论文集，2023，第 2639–2649 页。[32] Y. Ye、Z. Zheng、Y. Shen、T. Wang、H. Zhang、P. Zhu、R. Yu、K. Zhang 和 H. Xiong，“利用多模态大语言模型进行多模态序列推荐”，*arXiv* 预印本 *arXiv:2408 09698*，2024。

[33] J. Li、M. Wang、J. Li、J. Fu、X. Shen、J. Shang 和 J. McAuley，“文本即所需：为序列推荐学习语言表征”，载于 第 *29* 届 *ACM*知识发现与数据挖掘特别兴趣小组会议论文集，2023，第 1258–1267 页。[34] W. Ji、X. Liu、A. Zhang、Y. Wei、Y. Ni 和 X. Wang，“在线蒸馏增强的多模态 Transformer，用于序列推荐”，载于第 *31* 届 *ACM* 多媒体国际会议论文集，2023，第 955–965 页。[35] Q. Cui、S. Wu、Q. Liu、W. Zhong 和 L. Wang，“MV-RNN：用于序列推荐的多视图循环神经网络”，*IEEE*知识与数据工程汇刊，第 32 卷，第 2 期，第 317–331 页，2018。[36] Y. Hou、S. Mu、W. X. Zhao、Y. Li、B. Ding 和 J.-R. Wen，“面向推荐系统的通用序列表征学习”，载于 第 *28* 届 *ACM*知识发现与数据挖掘特别兴趣小组会议论文集，2022，第 585–593 页。[37] H. Hu、W. Guo、Y. Liu 和 M.-Y. Kan，“序列推荐系统中的自适应多模态融合”，载于*ACM*信息与知识管理国际会议论文集，2023，第 843–853 页。[38] M. Wang、Y. Xiao、B. Wang、S. Zhang、S. Ye、W. Wang、H. Yin、R. Guo 和 Z. Xu，“FindRec：用于多模态序列推荐的 Stein引导的熵流”，载于 第 *31* 届 *ACM*知识发现与数据挖掘特别兴趣小组会议第 *2* 卷论文集，2025，第 3008–3018 页。[39] X. Zhou 和 Z. Shen，“两个图的故事：冻结并去噪用于多模态推荐的图结构”，载于 第 *31* 届 *ACM* 多媒体国际会议论文集，2023，第 935–943 页。[40] J. Xu、Z. Chen、S. Yang、J. Li、W. Wang、X. Hu、S. Hoi 和 E. Ngai，“多模态推荐系统综述：近期进展与未来方向”，*IEEE*多媒体汇刊，2026。[41] S. Zhang、L. Chen、D. Shen、C. Wang 和 H. Xiong，“用于多模态序列推荐的分层时间感知混合专家”，*arXiv* 预印本 *arXiv:2501.14269*，2025。[42] J. Devlin、M.-W. Chang、K. Lee 和 K. Toutanova，“BERT：用于语言理解的深度双向Transformer预训练”，载于 *2019* 年计算语言学协会北美分会会议论文集：人类语言技术，第 *1* 卷（长文与短文），2019，第 4171–4186 页。[43] A. Dosovitskiy、L. Beyer、A. Kolesnikov、D. Weissenborn、X. Zhai、T. Unterthiner、M. Dehghani、M. Minderer、G. Heigold、S. Gelly 等，“一张图像胜过 16×16 个词：用于大规模图像识别的 Transformer”，载于 国际学习表征会议，2020。[44] J. Xu、L. Sun 和 D. Zhao，“Mome：用于高效多任务推荐的掩码专家混合模型”，载于 第 *47* 届 *ACM*信息检索研究与发展国际会议论文集，2024，第 2527–2531 页。[45] J. Ma、Z. Zhao、X. Yi、J. Chen、L. Hong 和 E. H. Chi，“利用多门控混合专家模型对多任务学习中的任务关系进行建模”，载于 第 *24* 届 *ACM*知识发现与数据挖掘特别兴趣小组会议论文集，2018，第 1930–1939 页。[46] W. Cai、J. Jiang、F. Wang、J. Tang、S. Kim 和 J. Huang，“混合专家综述”，*arXiv* 预印本 *arXiv:2407.06204*，2024。[47] M. Singh，“推荐数据集中的可扩展性与稀疏度问题：综述”，知识与信息系统，第 62 卷，43，2020 1，1第 页 – 期。[48] C. Wang 、 H. Zhu 、 P. Wang 、 C. Zhu 、 X. Zhang 、 E. Chen 和 H. Xiong，“个性化且可解释的员工培训课程推荐：贝叶斯变分方法”，*ACM* 信息系统汇刊（*TOIS*），第 40 卷，第 4 期，第 1–32 页，2021。[49] C. Wang、H. Zhu、C. Zhu、C. Qin、E. Chen 和 H. Xiong，“Setrank：用于推荐系统协同排序的集合式贝叶斯方法”，*ACM* 信息系统汇刊，第 42 卷，第 2 期，第 1–32 页，2023。[50] H. Du、H. Yuan、P. Zhao、F. Zhuang、G. Liu、L. Zhao、Y. Liu 和 V. S. Sheng，“结合对比知识蒸馏的序列推荐集成建模”，载于 *ACM*信息检索研究与发展国际会议论文集，2023，第 58–67 页。[51] Y. Xie、P. Zhou 和 S. Kim，“用于序列推荐的解耦侧信息融合”，载于第 *45* 届 *ACM*信息检索研究与发展国际会议论文集，2022，第 1611–1621 页。

[52] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, 《注意力就是你所需要的一切》，神经信息处理系统进展，第 30 卷，2017 年。[53] J. Zhu, X. Chen, K. He, Y. LeCun, and Z. Liu, 《无归一化变换器》，*arXiv* 预印本 *arXiv:2503.10622*，2025 年。[54] B. Dufumier, J. Castillo-Navarro, D. Tuia, and J.-P. Thiran, 《多模态对比学习中应对齐什么？》*arXiv* 预印本 *arXiv:2409.07402*，2024 年。[55] A. v. d. Oord, Y. Li, and O. Vinyals, 《通过对比预测编码进行表示学习》，*arXiv* 预印本 *arXiv:1807.03748*，2018 年。[56] R. He and J. McAuley, 《起起落落：利用单类协同过滤建模时尚趋势的视觉演化》，载于第 *25* 届万维网国际会议论文集，2016 年，第 507–517 页。[57] J. Xu, Z. Chen, J. Li, S. Yang, H. Wang, and E. C. Ngai, 《Aligngroup：在群体推荐中学习并协调群体共识与成员偏好》，载于第 *33* 届 *ACM*信息与知识管理国际会议论文集，2024 年，第 2682–2691 页。[58] Z. Yue, Y. Wang, Z. He, H. Zeng, J. McAuley, and D. Wang, 《用于序列推荐的线性循环单元》，载于第 *17* 届 *ACM* 网络搜索与数据挖掘国际会议论文集，2024 年，第 930–938 页。[59] W. X. Zhao, S. Mu, Y. Hou, Z. Lin, Y. Chen, X. Pan, K. Li, Y. Lu, H. Wang, C. Tian 等，《Recbole：面向推荐算法的统一、全面且高效的框架》，载于第 *30* 届 *ACM* 信息与知识管理国际会议论文集，2021 年，第 4653–4664 页。[60] D. P. Kingma and J. Ba, 《Adam：一种随机优化方法》，*arXiv* 预印本 *arXiv:1412.6980*，2014 年。[61] X. Glorot and Y. Bengio, 《理解深度前馈神经网络训练的难点》，载于第 *13* 届人工智能与统计国际会议论文集。JMLR 研讨会与会议论文集，2010 年，第 249–256 页。[62] X. Zhou, 《Mmrec：简化多模态推荐》，载于第 *5* 届 *ACM* 亚洲多媒体国际会议研讨会论文集，2023 年，第 1–2 页。[63] Y. Ye, Z. Zheng, Y. Shen, T. Wang, H. Zhang, P. Zhu, R. Yu, K. Zhang, and H. Xiong, 《运用多模态大语言模型进行多模态序列推荐》，载于*AAAI* 人工智能大会论文集，第 39 卷，第 12 期，2025 年，第 13 069–13 077 页。[64] Y. Ni, Y. Cheng, X. Liu, J. Fu, Y. Li, X. He, Y. Zhang, and F. Yuan, 《大规模内容驱动的短视频推荐数据集》，载于第 *34* 届 *ACM*信息与知识管理国际会议论文集，2025 年，第 6486–6491 页。