arXiv:2305.01915v1 [cs.IR] 3 May 2023

基于对比学习的去噪多模态序列推荐系统

Dong Yao 浙江大学 杭州，中国 yaodongai @zju.edu.cn

Shengyu Zhang 浙江大学 中国 杭州 sy_ zhang@zju.edu.cn

Zhou Zhao 浙江大学 中国杭州 zhaozhou @zju.edu.cn

Jieming Zhu 华为 中国深圳 jie mingzhu@ieee.org

Wenqiao Zhang 浙江大学 中国杭州 wenqiaozhang @zju.edu.cn

Rui Zhang ruizhang.info 深圳，中国 rayteam@yeah.net

Xiaofei He 浙江大学 杭州，中国 xiaofei_ h@qq.com

Fei Wu 浙江大学 杭州，中国 w ufei@zju.edu.cn

摘要

近年来，如何利用多模态数据提升推荐系统中的用户建模准确性，已成为一个快速发展的研究方向。现有的多媒体推荐系统通过整合各种模态并设计精细的模块，取得了显著的性能提升。然而，当用户决定与物品进行交互时，他们中的大多数并不会完整阅读所有模态的内容。我们将直接引发用户行为的模态称为兴趣点，这是捕捉用户兴趣的重要方面。相比之下，未引发用户行为的模态则属于潜在噪声，可能会误导推荐模型的学习。毫不意外的是，由于无法获取用户对其兴趣点的显式反馈，现有文献中鲜有研究致力于消除这类潜在噪声。为弥补这一研究空白，我们提出了一种基于对比学习的弱监督框架，用于多模态推荐系统的去噪（简称Demure）。通过弱监督方式，Demure 无需用户提供显式反馈，而是通过分析给定用户所有已交互物品的模态来识别噪声。具体而言，受可解释机器学习技术的启发，Demure 通过评估每种模态对最终推荐预测的贡献，识别潜在噪声和兴趣点。在完成识别后，Demure 通过强化兴趣点并削弱潜在噪声，对原始用户行为序列进行增强，分别获得增强后的正样本和负样本多模态行为序列。利用这些自增强序列，应用对比学习目标来学习更加准确且稳健的用户表示。通过大量实验，我们证明，Demure框架不仅显著提升了推荐效果，还能解释用户兴趣的关键模态。

CCS 概念

• **计算机系统组织** → **嵌入式系统**；冗余；机器人学； • **网络** → 网络可靠性。

关键词

数据集、神经网络、注视检测、文本标注

ACM 参考文献格式：

Dong Yao、Shengyu Zhang、Zhou Zhao、Jieming Zhu、Wenqiao Zhang、Rui Zhang、Xiaofei He 和 Fei Wu。2018 年。使用对比学习对多模态序列推荐系统进行去噪。在 请从您的权利确认电子邮件中输入正确的会议标题（会议缩写 '*XX*）。ACM，纽约州纽约市，美国，11 页。https: //doi.org/XXXXXXX.XXXXXXX

允许为个人或课堂使用而免费制作本作品的全部或部分内容的数字或纸质副本，但不得为盈利或商业利益制作或分发副本，且副本必须保留本声明和首页上的完整引文。对于本作品中不归 ACM 所有的组成部分，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、发布到服务器或分发至列表，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。会议缩写 '*XX*，*2018*年*6*月*3*日至*5*日，纽约州伍德斯托克© 2018 计算机协会。ACM ISBN 978-1-4503-XXXX-X/18/06...$15.00
https://doi.org/XXXXXXX.XXXXXXX

1 引言

对于许多线上内容共享网站而言，个性化推荐是一项至关重要的服务，旨在帮助用户发现内容并提升用户体验。内容共享网站上的大多数物品都关联着多模态特征。例如，微视频包含视觉帧和音轨，而文本描述[16, 22, 36, 41], 和新闻文章则包含图像、视频和文本标题 [26, 31, 34]。与其他机器学习任务中观察到的收益类似，内容推荐领域的最新进展 [10, 14, 36, 40, 47] 表明，将多模态数据融入用户建模，能够显著提升推荐性能。

尽管以往研究取得了丰硕进展，但大多数研究都只处理单一模态。由于以下挑战，将多模态信息融入推荐仍是一个正在发展的研究领域：1）如何弥合语义鸿沟



图 1：不同用户的点击行为可能由不同的模态驱动。未直接导致用户行为的模态可能会误导用户表示学习。

不同模态之间如何进行交互？2）如何设计以用户偏好为导向的多模态建模模块？近期，[5] 采用多层注意力机制，对物品的多模态特征以及行为序列中的多个物品进行聚合。[41]构建了模态特定的用户-物品二部图，并分别捕获每种模态中的用户偏好。尽管现有研究取得了显著进展，但大多数研究都忽略了多模态推荐中的噪声。具体而言，我们并不是指多模态特征本身存在噪声或受到损坏。然而，大多数用户在决定点击或购买物品时，并不会阅读所有模态的完整内容。实际上，在大多数情况下，用户会受到某些特定内容的吸引，而某些特定模态会驱动他们的行为。本文将这些直接促使用户产生行为的模态称为**兴趣点**。例如，在图1中，用户A在点击某部电影时，可能只想寻找一部适合儿童和家庭观看的电影（即，类别属性），而不关心其具体内容（例如，视频或图像）。在这种情况下，那些不会直接促使用户产生点击行为的模态，对于准确建模用户A的偏好而言，就属于潜在噪声。此外，考虑到点击等隐式反馈以及无处不在的干扰因素（例如，折扣信息），历史交互物品中也可能包含噪声物品，这些物品可能无法充分代表用户的潜在偏好。[37] 证明，与常规训练（包含噪声样本）相比，干净训练（不包含噪声样本）的性能要好得多，这凸显了在推荐系统中探索去噪方法的必要性。尽管我们可以收集一些显式反馈，例如物品层面的点赞或关注信号，但在模态层面，我们能够获取的用于指示直接驱动用户行为的兴趣点的信号却要少得多。

本文提出了一种统一的弱监督框架，用于在物品层面和模态层面对多模态推荐系统进行去噪，简称为**Demure**。Demure 属于弱监督方法，因为它不需要任何显式反馈作为监督信号。相反，它通过评估物品和模态对预测用户接下来会交互的物品所做的贡献，来识别其中的兴趣点和噪声；这是一种典型的序列推荐任务形式。受可解释机器学习技术的启发[30]，，我们将这些贡献度量为在下一物品预测任务中估计出的梯度。在直观上，我们将训练期间对预测下一物品贡献较高的模态/物品视为兴趣点。确定局部兴趣点和潜在噪声后，我们采用对比学习来消除噪声对最终用户嵌入的影响。具体而言，我们通过强化兴趣点或减弱噪声，在物品层面和模态层面增强原始多模态行为序列。鉴于作为兴趣点的物品/模态经过了修改，多模态用户序列和用户编码器应共同生成用户嵌入，作为原始用户嵌入的负视图。这类负用户嵌入应主要包含带有噪声的特征。将这些负样本推离原始嵌入，有助于消除噪声对最终用户嵌入的影响。同样，我们可以从已修改潜在噪声的多模态行为序列中获得正用户嵌入。因此，我们设计了对比学习目标，基于这些正样本和负样本，利用多模态特征实现准确的用户偏好建模。

我们在多个公开可用的多模态推荐基准上进行了广泛实验。实验结果表明，Demure 在多模态推荐方面取得了当前最先进的结果，并在模态级别和物品级别提供了对用户偏好的解释。值得注意的是，公开的新闻推荐数据集 MIND [43] 最初仅包含一种模态；因此，我们抓取了视觉图像，以获得多模态版本[1]。总的来说，本文的贡献有以下三点：

- 我们指出了用户表示学习中的多模态噪声问题，并进一步初步探讨了在没有明确指示哪些模态直接导致用户行为的信号时，如何对推荐进行去噪。
- 我们提出了一种名为 Demure 的弱监督框架，通过衡量各模态和物品在预测用户接下来会与哪个物品交互时所作的贡献，识别兴趣点和潜在噪声。我们设计了对比学习目标，以在模态层面和物品层面学习不受噪声影响的鲁棒用户表示。
- 我们在公开的多模态推荐基准数据集上进行了大量实验，并取得了最先进的性能。我们还基于开放的新闻推荐数据集 MIND，爬取每条新闻的封面图像，构建了一个多模态数据集。[2]。

2 相关工作

2.1 多模态推荐

基于内容的推荐具有重要意义且颇具挑战性，因此引发了浓厚的研究兴趣和多样化的探索 [3, 9, 15, 16, 24, 33, 45, 46, 50]。与仅基于行为的偏好建模相比，基于内容的推荐具有以下优势。首先，内容包含有关物品的多方面、细粒度

[1]我们将发布该数据集，以推动后续研究并使我们的结果可复现。[2]所有代码和数据集都将公开，以促进进一步改进。

详细信息，能够增强物品的表征，并支持对用户进行深入的偏好建模。其次，内容特征有助于缓解新物品上的流行度偏差，也适用于冷启动场景 [5]。

然而，许多研究仅处理单模态，例如新闻推荐中的新闻标题 [42, 43]以及微视频推荐中的视频特征 [16, 22]。如何在推荐中封装多模态信息，在研究领域仍鲜有探索。近期，ACF [5] 利用注意力机制，逐步聚合物品内的多模态信息，以及用户序列中的多个物品。MMGCN [41] 按模态构建用户-物品二部图，并进行特定于模态的信息传播。GRCN [40]基于 MMGCN，通过端到端学习进一步优化特定于模态的图，从而对其加以改进。与这些研究不同，我们明确学习如何定位用户的兴趣点，并提出采用对比学习框架对多模态学习进行去噪。

2.2 可解释机器学习

机器学习的解释日益受到研究人员的关注，尤其是在计算机视觉领域。[48] 提出了一种可解释的卷积神经网络，其中高层卷积层中的每个滤波器都表示一个特定的物体部件，这不同于传统的离线可视化方法。Grad-CAM [30] 利用梯度生成类别判别定位图，从而突出显示图像中与预测类别相关的关键区域。

2.3 对比学习

近年来，对比学习在广泛的研究领域中取得了巨大成功 [6, 11, 17, 20]。这一显著进展也引起了推荐研究领域的关注。最近，Zhou 等， [51] 将对比学习目标应用于现实世界的推荐系统中。他们惊讶地发现，这一目标有助于缓解流行度偏差，并从理论上证明，对比学习目标与更先进的去偏目标函数（如逆倾向评分）密切相关 [28]。Xie 等， [44] 提出了用于从推荐中预训练用户编码器的自监督学习框架。Demure 与这些工作的不同之处在于其目标，即*i.e*，去噪多模态推荐；以及其技术，即*i.e*，基于在模态层面和物品层面对兴趣点进行定位的对比学习。

3 方法

3.1 符号与问题

推荐数据集通常包含两类实体集合，即用户集合 $\mathrm{U} = \{u_1, u_2, \ldots, u_{|\mathrm{U}|}\}$ 和物品集合 $\mathrm{I} = \{i_1, i_2, \ldots, i_{|\mathrm{I}|}\}$。$|\mathrm{U}|$ 和 $|\mathrm{I}|$ 分别表示用户数和物品数。设 t 表示交互 $< u, i, t >$ 发生的时间戳，则用户 u 的交互序列可表示为 $\mathrm{S} = \{i_{u,t_1}, i_{u,t_2}, \ldots, i_{u,t_{|S|}}\}$，其中 i_{u,t_1} 表示用户 u 在时间戳 t_1 交互的物品，$|\mathrm{S}|$ 表示行为序列的长度。$t_1, t_2, \ldots, t_{|\mathrm{S}|}$ 通常按时间顺序排列。在多模态设置中，每个物品 i 关联若干模态 $\{i^m\}_{m=1,\ldots,M_i}$。M_i 表示物品 i 的模态数量。我们使用物品具有完整模态的公开数据集，因此，属于同一数据集的所有物品的 M_i 均相同。为方便起见，我们将其简化为 M。具体而言，我们采用序列推荐模型训练的通用设置 [1, 27, 35, 39, 49]。

3.2 多模态用户编码器

我们首先使用预训练模态专属特征提取器，为每个物品 i 提取不同模态的向量表示 $\{\mathbf{i}^m\}_{m=1,\ldots,|M|}$ $\{i^m\}_{m=1,\ldots,|M|}$[3]。给定用户序列 $\mathrm{S} = \{i_{t_1}, i_{t_2}, \ldots, i_{t_{|S|}}\}$，用户编码器旨在从历史行为中的多模态特征提取偏好模式，并据此生成用户嵌入。注意力机制是建模不同模态 [4, 23] 与物品 [4]之间相关性和交互的成熟工具。我们提出在 Demure 的骨干网络中利用这一优势。如图 2 所示，我们将用户编码器分解为两部分：

模态级编码器 为弥合模态差异并捕捉不同模态之间的异质依赖关系，我们利用自注意力机制 [32]，因为它效果显著。具体而言，给定不同模态的特征 $\{\mathbf{i}^m\}_{m=1,\ldots,|M|}$，我们将第 m 个模态的增强表示计算如下：

$$\beta_{m,n} = \frac{\exp\left(s_{m,n}\right)}{\sum_{k=1}^{M} \exp\left(s_{m,k}\right)}, \text{where } s_{m,n} = f_m\left(\mathbf{i}^m\right)^T g_m\left(\mathbf{i}^n\right) \tag{1}$$

$$\tilde{\mathbf{i}}^m = \sum_{k=1}^{M} \beta_{m,k} h_m\left(\mathbf{i}^k\right) \tag{2}$$

其中，f_m、g_m和 h_m 分别表示由权重矩阵$\mathbf{W}f$、$\mathbf{W}g$和$\mathbf{W}_h$参数化的线性变换。$s_{m,n}$ 表示模型在表示第 m个模态时对第 n个模态的关注程度。 $\tilde{}\mathbf{i}^m$ 表示第 m个模态的增强表示。为获得聚合后的物品嵌入，我们重复使用公式 1 中计算出的自注意力权重。具体而言，模态 m 的注意力权重可按如下方式计算：

$$\beta_m = \frac{\sum_{n=1}^{M} \exp\left(s_{m,n}\right)}{\sum_{m=1}^{M}\sum_{n=1}^{M} \exp\left(s_{m,n}\right)} \tag{3}$$

其中，$\sum_{n=1}^{M}\exp\left(s_{m,n}\right)$ 表示模型在表示所有模态时对第 m个模态的关注程度。由于物品由所有模态表示，因此可以将模态 m 的累积注意力除以$\sum_{m=1}^{M}\sum_{n=1}^{M}\exp\left(s_{m,n}\right)$ 后，视为表示该物品的注意力权重。这种设计也很轻量，无需引入额外的可学习参数。因此，聚合后的物品嵌入可以表示为$\mathbf{i}_t = \sum_{m=1}^{M} \beta_m \tilde{}\mathbf{i}^m$。

物品级编码器。 同样，为了捕捉不同物品之间的异质依赖关系，我们提出在表示用户时应用自注意力机制。具体而言，给定用户序列 $\mathbf{S} = \{\mathbf{i}_{t_1}, \mathbf{i}_{t_2}, \ldots, \mathbf{i}_{t_{|S|}}\}$ 的向量表示，其中每个物品 $\mathbf{i}_t$ 的表示均由模态级编码器生成。我们不再简单地对它们取平均，而是得到序列中所有物品的增强表示

[3]特征提取器的详细信息请参见第 4.1 节

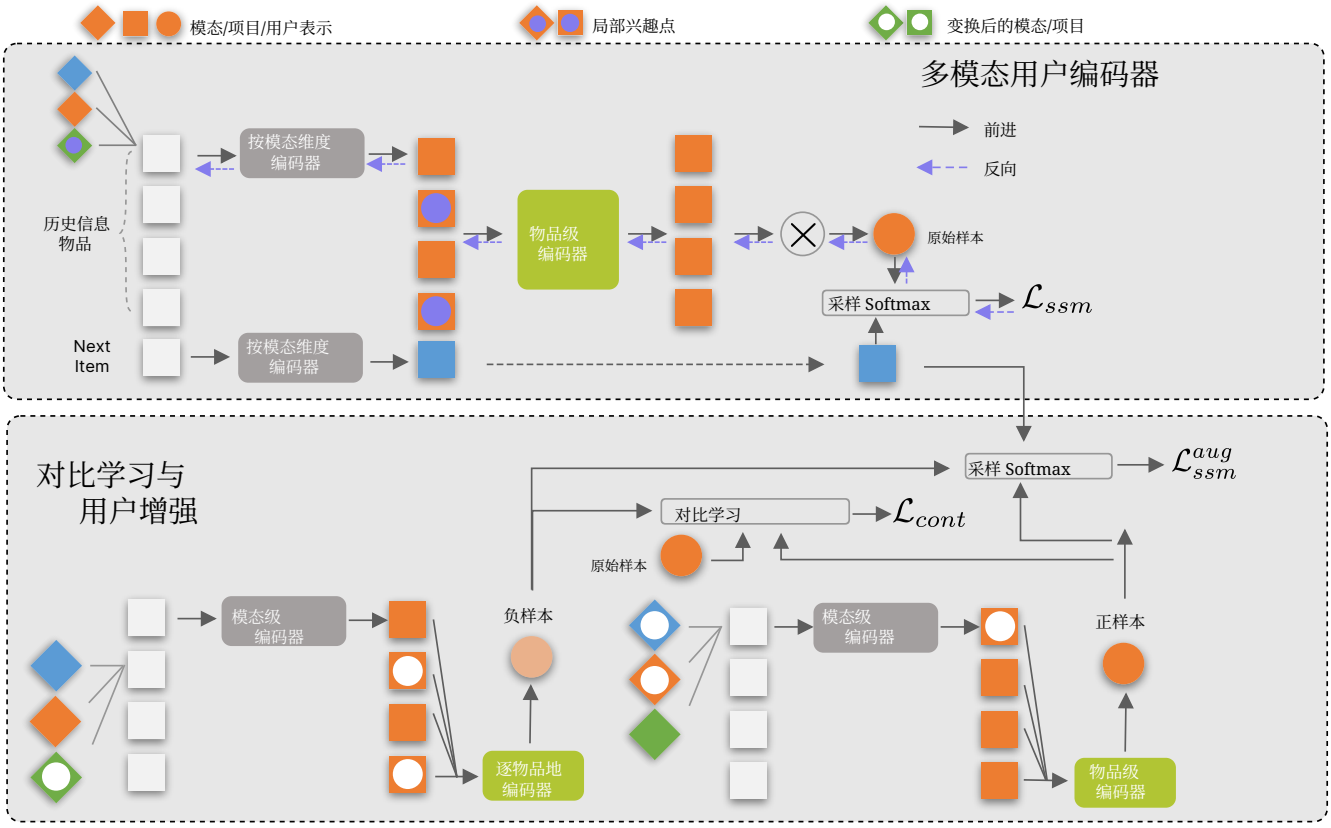


图 2：所提出的 Demure框架示意图。

并以与模态级编码器相同的方式，重用注意力权重将它们聚合为 $\mathbf{i}$。随后，我们通过前馈神经网络（FFN）对用户 u 进行编码，从而获得最终表示 $\mathbf{u}$：

$$\mathbf{u} = \mathrm{FFN}(\mathbf{i}) \tag{4}$$

3.3 预测与优化

与封装了复杂且可学习的预测层、并处理数量有限的用户-物品对的排序模型不同，匹配模型需要从庞大的物品库中检索前K个候选项，通常采用非参数且高效的技术。本文使用点积作为评分函数，并使用采样Softmax进行优化。这是一种广泛使用的训练目标，可表示如下：

$$\arg\min_{\theta} \mathcal{L}_{ssm} = \arg\min_{\theta} \frac{1}{|\mathcal{D}|} \sum_{(u,i^+)\in\mathcal{D}} -\log p'_\theta(i^+ \mid u), \tag{5}$$

$$p'_\theta(i^+ \mid u) = \frac{\exp\left(\psi\left(u, i^+\right)\right)}{\exp\left(\psi\left(u, i^+\right)\right) + \sum_{i^-\in\mathcal{I}^-} \exp\left(\psi\left(x, i^-\right)\right)} \tag{6}$$

其中，$^-$表示采样得到的负样本物品集合。i^+表示目标物品。$\psi(u,i)$表示 u和 i的表示向量的内积，可以将其视为 u和 i之间的相似度。

3.4 兴趣点定位

上述流程遵循了匹配阶段的典型训练框架。然而，我们认为，多模态推荐容易受到多种模态内部噪声的影响，而去噪的关键在于定位用户的兴趣点。我们借助可解释机器学习 [30, 48]，提出一种基于梯度在模态级和物品级定位兴趣点的方法。具体而言，我们分别计算 Softmax 之前正样本物品 i^+的分数 $p_\theta(i^+ \mid u)$ 相对于物品特征激活$\mathbf{i}_t \in \mathbb{R}^{d_i}$和模态特征激活$\mathbf{i}^m \in \mathbb{R}^{d_m}$的梯度。我们对激活值应用全局平均池化，以获得兴趣分数：

$$\alpha_{i_t} = \frac{1}{d_i} \sum_{k}^{d_i} \frac{\partial p'_\theta(i^+ \mid u)}{\partial \mathbf{i}_{t,k}}, \tag{7}$$

$$\alpha_{i_t^m} = \frac{1}{d_m} \sum_{k}^{d_m} \frac{\partial p'_\theta(i^+ \mid u)}{\partial \mathbf{i}^m_{t,k}} \tag{8}$$

在图像分类任务中，Grad-CAM [30] 利用目标类别预测向量相对于卷积层特征图激活值的梯度，粗略突出图像中的重要区域。受此启发，α_{i_t}和$\alpha_{i_t^m}$可以粗略评估用户 u对 i_t和 i_t^m的兴趣程度，以便

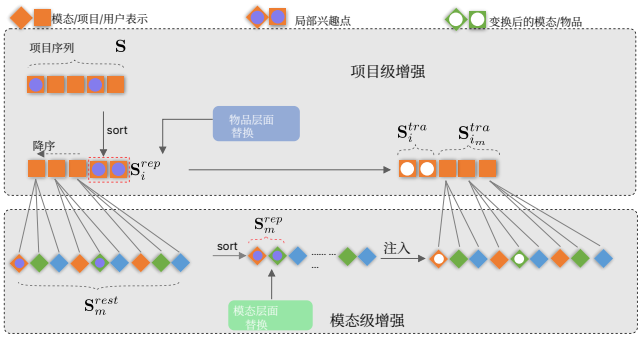


图 3：用户增强的基本流程。

选择下一个交互物品。在第 4.4.1 节中，我们通过实验验证了这一方法的合理性。

3.5 用户增强

我们提出了一种新颖的方法，用于学习对噪声模态/物品不那么敏感的用户表示。我们根据通过公式 7 和 8 计算出的兴趣分数，在模态层面和物品层面进行用户增强。由于篇幅限制，我们仅详细阐述负用户嵌入的生成过程：

预备知识。 $\mathbf{S} = \{\mathbf{i}_{u,t_1}, \mathbf{i}_{u,t_2}, \ldots\ldots, \mathbf{i}_{u,t_{|S|}}\}$ 是用户 u 的交互物品序列。为便于描述，我们将其简记为 $\mathbf{S} = \{\mathbf{i}_1, \mathbf{i}_2, \ldots\ldots, \mathbf{i}_l, \ldots\ldots, \mathbf{i}_L\} \in \mathbb{R}^{L\times d_i}$，其中 $\mathbf{i}_l$ 是物品嵌入向量。用户 u 未交互过的物品构成集合 $\mathbf{U} \in \mathbb{R}^{N\times d_i}$，其模态集合为 $\mathbf{U}_m \in \mathbb{R}^{NM\times d_m}$，其中 M 表示一个物品的模态数量。

项目级增强。 我们引入一个比率 γ_i，表示被转换物品所占的比例。如图 3 所示，给定用户的行为序列 S，我们根据公式 7 计算出的兴趣分数按降序对其排序。然后，从排序后的集合中选取排名前 K_i 个物品，构成被替换物品集合 $\mathbf{S}_i^{rep}$，并计算它们与属于 U 的其他物品之间的相似度分布，

$$\mathbf{S} = \{\mathbf{i}_1, \mathbf{i}_2, \ldots, \mathbf{i}_L\} \xrightarrow[select]{sort} \mathbf{S}_i^{rep} = \{\mathbf{i}_1', \mathbf{i}_2', \ldots, \mathbf{i}_{K_i}'\} \tag{9}$$

$$\mathbf{P}_i = \frac{\mathbf{D}_i - \min}{\max - \min}, \mathbf{D}_i = \mathbf{S}_i^{rep} \cdot \mathbf{U}^\top \tag{10}$$

其中，$\lfloor L \times \gamma_i \rfloor = K_i$，$\mathbf{S}_i^{rep} \in \mathbb{R}^{K_i\times d_i}$，$\mathbf{D}_i \in \mathbb{R}^{K_i\times N}$，$\max \in \mathbb{R}^{K_i}$ 和 $\min \in \mathbb{R}^{K_i}$ 分别包含矩阵 $\mathbf{D}_i$ 中每一行的最大值和最小值。为了从 U 中选出与 $\mathbf{S}_i^{rep}$ 中的物品不相似的物品进行替换，我们将分布 $1 - \mathbf{P}_i$ 视为其他物品替换对应物品的概率。基于分布 $1 - \mathbf{P}_i$，我们得到转换后的物品集合 $\mathbf{S}_i^{tra}$。

模态级增强。 经过物品层面增强后，剩余的物品集合为：

$$\mathbf{S}_i^{rest} = \mathbf{S} \setminus \mathbf{S}_i^{rep} \tag{11}$$

其中 $\mathbf{S}_i^{rest} \in \mathbb{R}^{K\times d_i}$，$K = L - K_i$。其模态集合记为 $\mathbf{S}_m^{rest} = \{\mathbf{i}_{11}, \mathbf{i}_{21}, \ldots\ldots, \mathbf{i}_1^m, \ldots\ldots, \mathbf{i}_1^M, \ldots\ldots, \mathbf{i}_K^1, \mathbf{i}_K^2, \ldots\ldots, \mathbf{i}_K^m, \ldots\ldots, \mathbf{i}_K^M\} \in \mathbb{R}^{KM\times d_m}$，其中 $\mathbf{i}_k^m$ 是模态嵌入向量。与物品层面的增强类似，我们根据公式 8 计算出的模态兴趣分数，按降序对 $\mathbf{S}_m^{rest}$ 进行排序。然后，我们选择排名前 K_m 的模态进行修改，

$$\mathbf{S}_m^{rest} = \{\mathbf{i}_1^1, \mathbf{i}_1^2, \ldots, \mathbf{i}_K^M\} \xrightarrow[select]{sort} \mathbf{S}_m^{rep} = \{\mathbf{i}_1^{m'}, \mathbf{i}_2^{m'}, \ldots, \mathbf{i}_{K_m}^{m'}\} \tag{12}$$

$$\mathbf{P}_m = \frac{\mathbf{D}_m - \min}{\max - \min}, \mathbf{D}_m = \mathbf{S}_m^{rep} \cdot \mathbf{U}_m^\top \tag{13}$$

其中，$\lfloor KM \times \gamma_m \rfloor = K_m$，$\mathbf{S}_m^{sep} \in \mathbb{R}^{K_m\times d_m}$，$\mathbf{D}_m \in \mathbb{R}^{K_m\times NM}$，$\max \in \mathbb{R}^{K_m}$ 和 $\min \in \mathbb{R}^{K_m}$ 分别包含矩阵 $\mathbf{D}_m$ 中每一行的最大值和最小值。为了从 $\mathbf{U}_m$ 中选出与 $\mathbf{S}_m^{sep}$ 中的模态不相似的模态进行替换，我们将分布 $1 - \mathbf{P}_m$ 视为其他模态替换对应模态的概率。基于分布 $1 - \mathbf{P}_m$，我们生成转换后的模态集合 $\mathbf{S}_m^{tra}$。

用户增强。 我们将变换后的模态集合 $\mathbf{S}_m^{tra}$ 注入 $\mathbf{S}_m^{rest}$，并替换相应的原始模态。随后，我们使用**模态级编码器** 对结果集合进行编码，得到模态层面的增强物品 $\mathbf{S}_{i_m}^{tra}$。最后，我们使用**物品级编码器** 对变换后的序列进行编码。该序列由物品层面和模态层面的增强物品汇集而成，即 *i.e.*，$\mathbf{S}_i^{tra}$ 和 $\mathbf{S}_{i_m}^{tra}$，从而得到负用户表示。我们也可以将兴趣分数排名前 γ_m 和 γ_i **最低** 的模态和物品进行替换，以生成正用户表示。值得一提的是，用于选择模态/物品的相似度分布无需通过从 1 中减去它们来进行转换。

3.6 对比学习

利用增强后的正/负用户嵌入 $\mathbf{u}^+/\mathbf{u}^-$，我们以原始用户嵌入 $\mathbf{u}$ 进行对比学习。数学上，我们有：

$$\mathcal{L}_{cont} = -\log \frac{\sum_{j=1}^{J} \exp\left(\phi\left(\mathbf{u}_j^+, \mathbf{u}\right)\right)}{\sum_{j=1}^{J} \exp\left(\phi\left(\mathbf{u}_j^+, \mathbf{u}\right)\right) + \sum_{k=1}^{J} \exp\left(\phi\left(\mathbf{u}_k^-, \mathbf{u}\right)\right)} \tag{14}$$

其中，J 是增强后的正/负用户数量。ϕ 是一个相似度函数，实验中我们采用点积。简而言之，从多模态推荐去噪的角度来看，该目标具有以下优势：

- 在增强后的正用户序列中，噪声模态/物品会被随机采样的模态/物品特征替换。将用户表示 $\mathbf{u}$ 与 $\mathbf{u}^+$ 拉近，可以减少对噪声模态的过拟合。
- 此外，这种拉近会使用户编码器学会更多地信任兴趣点，因为它们在正用户序列中未经过变换。
- 将原始用户嵌入 $\mathbf{u}$ 与负用户表示 $\mathbf{u}^-$ 推远，会使用户编码器学会减少受噪声模态/物品的影响。

对于增强后的正样本用户嵌入，我们还采用了第 3.3 节所述的推荐专用目标，进一步提升用户编码器学习更优表示的能力。我们

将相应的目标记为 L_{ssm}^{aug}。因此，Demure 的最终损失函数可写为：

$$\mathcal{L} = \mathcal{L}_{ssm} + \lambda_1 \mathcal{L}_{ssm}^{aug} + \lambda_2 \mathcal{L}_{cont} \quad (15)$$

其中，λ_1 和 λ_2 是用于控制各损失项权重的超参数。

4 实验

本节在三个公开数据集上进行实验，以展示 Demure 的有效性。我们从以下角度分析 Demure：

• **RQ1**：在多模态设置下，与当前最先进的多模态推荐模型及其他最先进的序列推荐系统相比，Demure 的表现如何？ • **RQ2**：不同设计（例如，移除对比学习目标、移除模态级/物品级增强，以及采用不同的变换率 γ_i/γ_m）会如何影响 Demure 的表现?

• **RQ3**：与注意力基础模型相比，Demure 是否能对多模态建模进行去噪?

4.1 实验设置

数据集。 根据 [40, 41], , 我们在两个公开的多模态推荐基准数据集上进行实验，即*i.e.*，TikTok 和 MovieLens [4]。为在更广泛的领域评估 Demure，我们纳入了新闻推荐数据集 MIND-NEWS[43], , 并通过爬取每条新闻对应的图像，构建其多模态版本 MIND-MM。表 1 列出了这些数据集的特征。

• **TikTok**[5]。 一个微视频推荐数据集，其中每个微视频都关联了预先提取的视频、文本和声学特征。根据 [41], , 我们使用一个采样版本来加快评估。

• **MovieLens**[6]。为便于进行多模态推荐评估，我们选用包含视频封面图像、视频标题/描述和类别的 MovieLens-20M 版本。我们采用预训练的 100 维 GloVe 词嵌入 [25]来表示视频标题/描述中的每个词。我们使用预训练的 ResNet50 [7] [12] 模型，为视频封面图像提取一个 2048 维的视觉特征向量。我们使用可训练的嵌入层，将每个类别映射为一个 64 维特征向量。

• **MIND-MM**。 MIND-NEWS [8] 是一种新发布的新闻推荐数据集，仅包含文本和类别。我们通过从给定的新闻 URL 抓取图像来扩充原始数据集。我们舍弃不包含图像的新闻，并使用剩余新闻构建 MIND-MM。我们采用与 MovieLens 数据集相同的流程进行特征提取。

[4]论文中使用的快手数据集无法从官方网站 https://www.kuaishou.com/activity/uimc 下载。[5] http://ai-lab-challenge.bytedance.com/tce/vc/。[6] https://grouplens.org/datasets/movielens/。[7] https://pytorch.org/vision/stable/models.html [8]原始 MIND 数据集可在 https://msnews.github.io/#about-mind 找到。我们将发布多模态版本，以推动多模态推荐研究。

表 1：数据集统计信息。V、A/C、T 分别表示视频、音频/类别和文本模态。

数据集	#用户数	#物品数	#交互数	#密度	V	A/C	T
TikTok	40,357	89,335	1,208,213	0.033%	128	128	128
MovieLens	138,493	25,381	19,882,894	0.56%	2,048	64	100
MIND-新闻	58,868	21,525	903,139	0.071%	2,048	64	100

评估协议。 我们按照 8:1:1 的比例，将数据集中的用户随机划分为训练集、验证集和测试集。我们注意到，与划分用户行为序列相比，划分用户更具挑战性，因为模型需要应对分布外序列，而这对于评估去噪效果至关重要。在测试期间，我们将前 80% 的行为视为历史行为，并使用剩余 20% 的行为进行评估。与 MMGCN 为每位用户配对 1000 个其此前未交互过的未观测物品不同，我们建议从整个物品库中召回正样本，这根据 [19] 可能更符合实际情况。

基线方法。 为了评估 Demure 的效果，我们将其与以下最先进的序列推荐模型进行比较。如第 3.1 节所述，我们聚焦于推荐的匹配阶段，特别选择匹配类对比方法。我们不将 Demure 与基于图的方法进行比较，因为它们依赖用户 ID 嵌入进行协同过滤，无法处理训练集之外的用户。我们采用两种广泛使用的数值指标，即 Recall@K 和 NDCG@K，并报告所有模型在测试集上推荐的前 K=20 个和前 50 个结果的表现。

• **YoutubeDNN [8]**。YoutubeDNN 通过平均池化整合历史交互物品。• **GRU4Rec [13]**。GRU4Rec 使用门控循环单元 [7] 替代简单的平均池化操作，以捕获物品之间的序列依赖关系。• **MIND [21]**。MIND 利用胶囊网络 [29] 从给定的行为序列中提取多个兴趣向量。• **ComiRec-DR [2]**。除了为单个用户生成多兴趣表示外，ComiRec-DR 还引入了一个可控因子，以平衡推荐质量与推荐多样性。• **ComiRec-SA [2]**。ComiRec-SA 与 MIND 的不同之处在于，它利用注意力机制提取多种兴趣，并在实证研究中表现出良好的效果。此外，它还引入了一个可控因子，以平衡推荐质量与推荐多样性。• **MTIN [16]**。MTIN 是一种先进的基于内容的序列推荐系统，它将多尺度时间效应融入用户建模，并利用群组路由模块根据行为序列更新兴趣向量。

基线方法的实现。我们参考了 https://github.com/THUDM/ComiRec 中基线方法的实现，并将其扩展为多模态版本。具体而言，在将物品嵌入编码为用户表示之前，我们通过注意力机制聚合各模态的嵌入，以获得物品嵌入。

表2：不同多模态推荐数据集和比较方法下的性能比较。基础模型是基于 YoutubeDNN 修改得到的模型。具体而言，我们将 YoutubeDNN 扩展到多模态设置，并额外采用注意力机制代替均值池化层，将模态/物品聚合为物品/用户表示。

模型	TikTok				MovieLens-20M				MIND-MM			
	@20指标		@50指标		@20指标		@50指标		@20指标		@50指标	
	召回率	NDCG	召回率	NDCG	召回率	NDCG	召回率	NDCG	召回率	NDCG	召回率	NDCG
YoutubeDNN	0.025	0.0492	0.0451	0.0864	0.1251	0.3617	0.249	0.5432	0.2057	0.3373	0.3226	0.4851
GRU4Rec	0.0266	0.0577	0.0447	0.0897	0.1255	0.3734	0.2368	0.5391	0.2014	0.3469	0.323	0.5069
MIND	0.0254	0.0493	0.0494	0.0947	0.1308	0.3655	0.2449	0.5366	0.2063	0.3393	0.3244	0.4873
ComiRec-DR	0.0223	0.0473	0.0472	0.0942	0.1201	0.3562	0.2346	0.5302	0.1913	0.3212	0.3182	0.4856
ComiRec-SA	0.0362	0.0708	0.059	0.1145	0.1191	0.3843	0.2436	0.5726	0.2072	0.3383	0.316	0.496
MTIN	0.0352	0.0715	0.0697	0.1369	0.1328	0.3804	0.2601	0.5631	0.2092	0.3195	0.3212	0.4626
基础模型	0.026	0.056	0.0526	0.1069	0.1303	0.373	0.2572	0.5581	0.1906	0.3237	0.3115	0.4795
Demure	**0.0428**	**0.0848**	**0.0746**	**0.1478**	**0.1413**	**0.4037**	**0.2694**	**0.5805**	**0.2193**	**0.3633**	**0.3509**	**0.5249**
%improv.	18.23%	18.60%	7.03%	7.96%	6.4%	5.05%	3.58%	1.38%	4.83%	4.73%	8.17%	3.55%

参数设置。对于所有数据集，我们使用大小为 1024 的小批次训练 Demure；在 TikTok/MovieLens 和 MIND-MM 上，学习率分别设为 0.001/0.003。我们使用 Adam优化器 [18] $\beta_2 =$, 其中 $\beta_1 = 0.9$, 。$^{0.99}$ $^{1e-8}$ ^{2}我们以速率 $1e-7$ 对可训练参数进行 L 归一化。我们分别训练 20/20/100 个训练轮次，并在 TikTok/MovieLens/MIND-MM 数据集中，为每位用户最多保留 20/50/20 个历史项。

4.2 性能比较（RQ1）

我们在表2中报告了实证结果，其中 %improv. 表示相对于最强最先进方法的相对提升，该方法以<u>下划线</u>标出。通过分析性能比较，我们得出以下几点关键观察：

• Demure在所有指标上均以较大幅度持续优于所有对比模型。例如，在Tiktok、MovieLens-20M和MIND-MM数据集上，相对于最强基线方法，在Recall@20指标上的相对提升分别为18.23%、6.4%和4.83%。此外，不进行去噪的基础模型（ w.o.）不具备竞争力，在大多数情况下无法超过最强的S OTA方法。我们将Demure的实证有效性归因于以下两点：1）借助可解释机器学习中的技术，Demure能够有效识别模态层面和物品层面的兴趣点/噪声。相比之下，大多数比较方法忽略了多模态和用户行为中的隐式反馈所引入的噪声，因此会导致含噪的用户兴趣建模。2）对比学习框架通过将从噪声以及随机采样的模态/物品中获得的用户表示，与用于预测的用户表示推离，显式消除了已识别噪声点的影响。

• 具体而言，传统方法 YoutubeDNN 仅对所有物品的表示进行均值池化，且深受多模态和物品中噪声的影响，在大多数情况下会导致较差的结果。GRU4Rec 中的更新门和重置门可能通过差异化地处理多模态/物品。MIND、ComiRec-DR、ComiRec-SA 等更先进的方法使用胶囊网络和注意力机制构建多个兴趣向量。这些架构可能会将兴趣点/噪声聚合到部分多兴趣向量中，因此受噪声兴趣点影响的可能性较低，进而提升性能。MTIN 架构通过根据每个物品的兴趣-物品匹配分数与所有物品的平均匹配分数之间的差距，为每个物品计算一个折扣因子，从而隐式处理隐式反馈问题，即物品层面的噪声。这一设计使 MTIN 在所有比较方法中取得最佳性能，进一步证明了推荐去噪的必要性。Demure 通过可解释的机器学习，显式缓解模态层面和物品层面的噪声，取得了大幅提升和最佳结果。

4.3 性能比较（RQ1）

我们在表2中报告了实证结果，其中 %improv. 表示相较于以<u>下划线</u>标出的最先进方法所取得的相对提升。通过分析性能比较，我们得出以下几点重要观察：

• Demure 在所有指标上始终以较大优势优于所有对比模型。例如，在 TikTok、MovieLens-20M 和 MIND-MM 上，针对 Recall@20 指标，相较于最强基线方法，分别提升了 18.23% 、6.4% 和 4.83%（就...而言）。

• 具体而言，传统方法 YoutubeDNN 对所有物品的表示进行平均池化，并且会严重受到多模态和物品中噪声的影响，导致在大多数情况下效果不佳。更先进的方法，如 MIND、ComiRec-DR 和 ComiRec-SA，使用胶囊网络或注意力机制构建多个兴趣向量。这些架构可能会聚合兴趣点/噪声

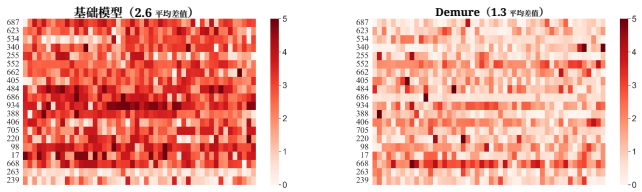
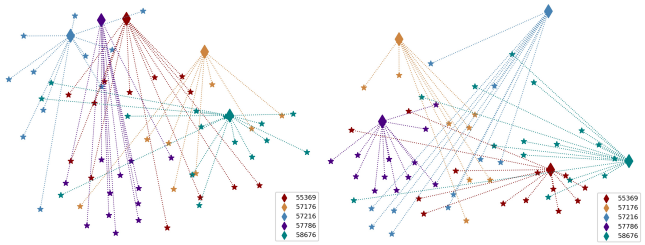


图 4：预测兴趣分数（梯度）与用户对物品的相应真实评分之间一致性的可视化。每个像素的颜色深浅表示绝对差异。我们绘制了基础模型和 Demure 的图。



（a）不使用去噪 （b）Demure

图 5：MIND-MM 测试集上随机抽样用户及其相关物品经 t-SNE 转换后的表示可视化。这些表示由多模态特征获得。

属于多兴趣向量的一部分，因此受点噪声影响的可能性较小，从而提升性能。MTIN 架构通过根据每个物品的兴趣-物品匹配分数与所有物品的平均匹配分数之间的距离，为每个物品计算折扣因子，从而隐式处理隐式反馈问题，即物品层面的噪声。这种设计使 MTIN 在所有比较方法中取得了最佳性能，证明了推荐去噪的必要性。Demure 通过利用可解释机器学习，显式缓解模态层面和物品层面的噪声，取得了最佳结果，并实现了显著的性能提升。

• 与缺少用户增强和对比学习目标的基础模型相比，Demure 的效果要好得多。我们将 Demure 在实证中的有效性归因于：**1**）借助可解释机器学习中的技术，Demure 能够有效识别模态层面和物品层面的兴趣点/噪声。相比之下，大多数比较方法忽视了多模态和用户行为中的隐式反馈所引入的噪声，因此用户兴趣建模会受到噪声影响；**2**）正如我们在上文第 3.6 节中所分析的，对比学习框架通过将增强后的负向用户序列所获得的用户表示与原始用户表示推远，明确消除了已识别噪声的影响。

4.4 案例研究（RQ2）

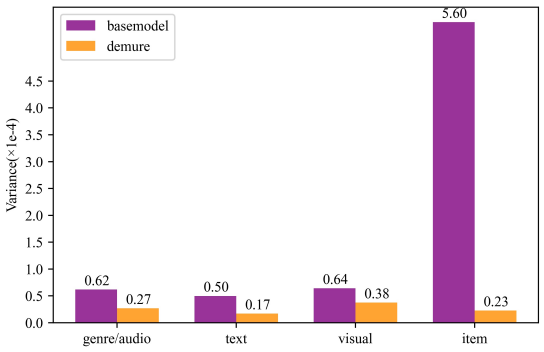
4.4.1 梯度可视化。如果没有数据集中的（*w.o.*）显式反馈，就很难验证将梯度视为兴趣分数以定位兴趣点是否合理。因此，我们选择 ml-1m 数据集进行实验，该数据集包含用户对其交互物品的评分。我们沿用上述三个数据集相同的实验设置，使用 ml-1m 训练 Demure 和基础模型。随后，我们对用户序列内的梯度进行归一化，并将其乘以 5，使其尺度与实际评分保持一致。考虑到图表的大小和清晰度，我们随机选取 20 位用户及其行为序列中的 50 个物品。我们用实际评分减去归一化并缩放后的梯度，并取绝对值作为差异分数。随后，我们将差异分数可视化。颜色越深，表示差异越大。从图 4 可以看出，与实际评分分布相比，Demure 训练后的梯度分布更趋近于实际评分分布。实验结果在一定程度上证明，梯度较高的物品/模态可以近似视为兴趣点。用户增强和对比学习引导兴趣点定位朝着更准确的方向发展。

4.4.2 嵌入空间分析。本节旨在分析所提出框架在学习准确的用户/物品表示方面的去噪效果。我们采用 [38] 提出的流程，对 MIND-MM 测试集中随机抽取的五名用户及其相关物品的嵌入进行 t-SNE 转换。基础模型和 Demure 的结果分别绘于图 5a 和图 5b。值得一提的是，这些结果是在运行 5 次后获得的。我们有以下两点观察：

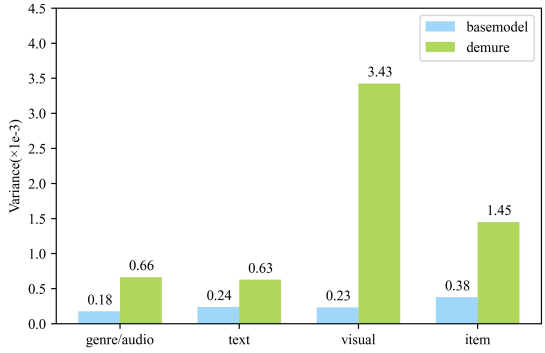
• 在图 5b 中，与模型不使用去噪相比，物品明显聚集在对应用户（例如，57786）周围，且混淆程度更低。这意味着 Demure 能够从多模态特征中学习到更准确的用户/物品表示，并在训练期间于模态/物品层面成功捕捉兴趣点。

• 图 5a 显示了不同用户之间明显的纠缠现象（*e.g.*，57216、57786、55369），导致物品被错误分配，并最终产生带噪声的推荐结果。相比之下，图 5b 中的用户彼此之间具有更高的可区分性，证明了多模态推荐去噪在学习准确的用户表示方面的有效性。

4.4.3 注意力权重分析。上述可视化通过描绘物品与用户之间的相关性，间接反映了多模态特征的去噪效果。为了明确展示用户与不同模态之间的相关性，我们提出可视化同一用户在不同多模态下所学注意力权重的稳定性。具体而言，我们分别对基线模型和 Demure 模型进行 10 次实验，直至收敛，并计算不同运行结果中注意力权重在模态层面／物品层面的方差。我们对随机抽取的 1024 名用户的方差取平均值，结果如图 6a 所示。我们观察到，基础模型在模态层面和物品层面的方差都显著高于 Demure。直观而言，数据中的噪声会导致对关键分配的置信度较低，



(a) 物品/模态注意力权重相对于10 次运行的方差。



(b) 每个用户内的物品/模态注意力权重方差。

图 6：不同场景下物品/模态注意力权重的方差。

从而使其随时间不断变化。这些结果可能表明，Demure 有助于为多模态推荐去噪，并学习能够感知兴趣点／噪声的用户表示。

为避免上述分析受到 Demure 学习到普遍无意义的注意力权重这一问题的影响（例如，对于同一用户的所有模态/物品，注意力权重几乎相同），我们计算每个用户内部注意力权重的方差，并取 1024 个用户的平均值。结果如图 6 所示。可以看出，Demure 会为同一用户内的不同模态/物品分配显著不同的注意力权重。图 6a 和图 6 共同表明，Demure 能够识别出置信度较高的兴趣点/噪声。

4.5 定量结果（RQ3）

4.5.1 增强率研究 γ_i/γ_m。 增强率 γ_i 和 γ_m 之和表示总共被转换的物品数量。 γ_i 和 γ_m 反映了物品层面增强与细粒度模态层面增强之间的权衡。我们改变 $\gamma_i+\gamma_m$、 γ_i 和 γ_m，以深入分析不同层次增强的影响。结果如表 3 所示：

- 总体而言，Demure 在所有配置下的表现均优于基础模型。换句话说，Demure 是对这些超参数不敏感，并且能够实现改进-无需繁复花哨的设计。
- 当 $\gamma_i + \gamma_m = 0.6$ 时效果最佳。过高或过低的 $\gamma_i + \gamma_m$ 都会导致效果不佳。由于 $\gamma_i + \gamma_m$ 表示经过转换的物品数量（包括物品层面和模态层面的转换），转换过多的物品（$\gamma_i + \gamma_m = 0.8$）可能会导致正样本坍塌，其中许多兴趣点也可能被转换；而转换的物品不足（$\gamma_i + \gamma_m = 0.4$）则可能降低增强效果，因为增强后的样本大多与原始样本相同。
- 在大多数情况下，较大的 γ_i 会比更大的 γ_m 表现更好。例如，在 $\gamma_i+\gamma_m = 0.8$ 和 0.6 时，大多数情况下 $\gamma_m = 0.2$ 能取得最佳效果。不过， γ_m 过低（*e.g.*， $\gamma_m = 0.1$， $\gamma_i = 0.3$）仍会导致性能下降。
- 三个不同领域的数据集上的结果（微视频、视频、新闻推荐）存在细微差异，这意味着不存在适用于所有应用场景的通用最优配置。

表 3：增强率对模态层面 γ_m 和物品层面 γ_i 的影响研究。

		TikTok		MovieLens		MIND-MM	
$\gamma_m + \gamma_i$	γ_m	召回率	NDCG	召回率	NDCG	召回率	NDCG
0.8	0.6	0.0417	0.0867	0.1404	0.3991	0.2119	0.3569
	0.4	0.04	0.0797	0.1373	0.3985	0.2102	0.3564
	0.2	0.0396	0.0824	0.1411	0.4000	0.2132	0.3559
0.6	0.4	0.041	0.0825	0.1385	0.3935	0.2106	0.3537
	0.3	0.0376	0.0781	0.1406	0.4005	0.2145	0.3591
	0.2	0.0428	0.0848	0.1413	0.4037	0.2193	0.3633
0.4	0.3	0.0409	0.0842	0.1438	0.4058	0.2097	0.3556
	0.2	0.0387	0.0789	0.1407	0.3992	0.2164	0.3618
	0.1	0.0403	0.082	0.1406	0.3984	0.2126	0.3583

4.5.2 增强样本数量研究。我们评估增强正/负样本对的数量， J。我们将 J设为 {1、2、4}。如表4所示，我们得出以下观察结果：

- Demure 在 $J = 1$和 $J = 2$ 上的表现，在三个数据集之间没有显著差异，这意味着 Demure 对增强样本的数量不太敏感。我们进行了相应的显著性检验，TikTok、MovieLens、MIND-MM 的 p 值分别为 0.153、0.281、0.397。它们均大于 0.05。
- 进一步增加 J会导致性能略有下降。我们认为这一现象是因为同时将查询用户嵌入拉近多个正样本，并使其远离多个负样本，可能会增加优化难度，最终导致优化陷入平凡解。

5 结论

本文提出对多模态推荐进行去噪，因为引入多种模态既能带来有用特征，也会引入噪声信息。为此，我们设计了一种新颖的框架，能够明确定位兴趣点，并将原始用户行为序列扩展到模态／物品

表4：增强后的正样本-负样本用户对数量研究。

	TikTok		MovieLens		MIND-MM	
	召回率	NDCG	召回率	NDCG	召回率	NDCG
J = 1	**0.0428**	**0.0848**	0.1413	0.4037	**0.2193**	0.3633
J = 2	0.0393	0.0813	**0.1432**	**0.4046**	0.2172	**0.365**
J = 4	0.0366	0.0748	0.1398	0.3985	0.2126	0.3572

级别。Demure 还引入了对比学习目标，帮助用户编码器学习对兴趣点更敏感、对其他模态／物品不那么敏感的偏好表示，从而实现多模态推荐去噪。实验结果表明，Demure 在两个公开的基准数据集以及一个基于新闻推荐数据集构建的多模态数据集上均有效。案例研究定性地表明，Demure 有助于对多模态推荐进行去噪。

参考文献

[1] 2020. Make It a Chorus：面向序列推荐的知识与时间感知物品建模。*SIGIR 2020*——第*43*届国际 *ACM SIGIR* 信息检索研究与开发会议论文集（2020），109 118 10 1145 3397271 3401131 – 。https://doi.org/ 。/ 。

[2] Yukuo Cen、Jianwei Zhang、Xu Zou、Chang Zhou、Hongxia Yang 和 Jie Tang。2020。用于推荐的可控多兴趣框架。。载于*KDD '20*：第*26*届 *ACM SIGKDD* 知识发现与数据挖掘会议，线上会议，美国加利福尼亚州，*2020*年*8*月*23*日至*27*日。[3] Muthusamy Chelliah、Soma Biswas 和 Lucky Dhakad。2019。原理到程序：采用多模态输入的神经时尚推荐。。载于第*27*届 *ACM* 国际多媒体会议论文集，*MM 2019*，法国尼斯，*2019*年*10*月*21*日至*25*日。[4] Jingyuan Chen 和 Xiangnan He。2017。注意力协同过滤：基于项目级和组件级注意力的多媒体推荐。（2017），335–344。[5] Jingyuan Chen、Hanwang Zhang、Xiangnan He、Liqiang Nie、Wei Liu 和 Tat-Seng Chua。2017。注意力协同过滤：基于项目级和组件级注意力的多媒体推荐。。载于第*40*届国际 *ACM SIGIR* 信息检索研究与开发会议论文集，日本东京新宿，*2017*年*8*月*7*日至*11*日。[6] Ting Chen、Simon Kornblith、Mohammad Norouzi 和 Geoffrey E. Hinton。2020。视觉表示对比学习的简单框架。。载于第*37*届国际机器学习会议论文集，*ICML*——*2020*年*7*月*13*日至*18*日，线上会议。[7] Kyunghyun Cho、Bart van Merrienboer、Caglar Gulcehre、Dzmitry Bahdanau、Fethi Bougares、Holger Schwenk 和 Yoshua Bengio。2014。使用 RNN 编码器-解码器学习用于统计机器翻译的短语表示。。载于*2014*年自然语言处理实证方法会议论文集，*EMNLP 2014*，*2014*年*10*月*25*日至*29*日，多哈，卡塔尔；*ACL* 特别兴趣小组 *SIGDAT* 会议。[8] Paul Covington、Jay Adams 和 Emre Sargin。2016。用于 YouTube 推荐的深度神经网络。。载于第*10*届 *ACM* 推荐系统会议论文集，美国马萨诸塞州波士顿，*2016*年*9*月*15*日至*19*日。[9] Xiaoyu Du、Xiang Wang、Xiangnan He、Zechao Li、Jinhui Tang 和 Tat-Seng Chua。2020。如何学习冷启动多媒体推荐的物品表示? 载于*MM '20*：第*28*届 *ACM* 国际多媒体会议，线上会议／美国华盛顿州西雅图，*2020*年*10*月*12*日至*16*日。[10] Wenjing Fu、Zhaohui Peng、Senzhang Wang、Yang Xu 和 Jin Li。2019。通过深度融合评论和内容，为跨域推荐系统中的冷启动用户提供推荐。第*33*届 *AAAI* 人工智能会议，*AAAI 2019*；第*31*届人工智能创新应用会议，*IAAI 2019*；以及第*9*届 *AAAI* 人工智能教育进展研讨会，*EAAI 2019*（2019），94 101 10 1609 33 01 330194 – 。https://doi.org/ 。/aaai.v i 。[11] Kaiming He、Haoqi Fan、Yuxin Wu、Saining Xie 和 Ross B. Girshick。2020。用于无监督视觉表示学习的动量对比。。载于*2020*年 *IEEE/CVF* 计算机视觉与模式识别会议，*CVPR 2020*，美国华盛顿州西雅图，*2020*年*6*月*13*日至*19*日。[12] Kaiming He、Xiangyu Zhang、Shaoqing Ren 和 Jian Sun。2016。用于图像识别的深度残差学习。载于*IEEE* 计算机视觉会议论文集视觉与模式识别. 770–778. [13] Balázs Hidasi、Alexandros Karatzoglou、Linas Baltrunas 和 Domonkos Tikk。2016年。基于循环神经网络的会话推荐。. 载于 第*4*届国际学习表征会议，*ICLR 2016*，波多黎各圣胡安，*2016*年*5*月*2*日至*4*日，会议论文集。[14] Xiaowen Huang、Quan Fang、Shengsheng Qian、Jitao Sang、Yan Li 和 Chang- sheng Xu。2019年。基于知识图谱的可解释交互驱动序列推荐用户建模。. 载于 第*27*届 *ACM* 国际多媒体会议论文集，*MM 2019*，法国尼斯，*2019*年*10*月*21*日至*25*日。[15] Hao Jiang、Wenjie Wang、Meng Liu、Liqiang Nie、Ling-Yu Duan 和 Changsheng Xu。2019年。Market2Dish：健康感知食品推荐系统。. 载于第*27*届 *ACM* 国际多媒体会议论文集，*MM 2019*，法国尼斯，*2019*年*10*月*21*日至*25*日。[16] Hao Jiang、Wenjie Wang、Yinwei Wei、Zan Gao、Yinglong Wang 和 Liqiang Nie。2020年。你喜欢哪个方面：面向微视频推荐的多尺度时间感知用户兴趣建模。. 载于 *MM '20*：第*28*届 *ACM* 国际多媒体会议，线上会议／美国华盛顿州西雅图，*2020*年*10*月*12*日至*16*日。[17] Minguk Kang 和 Jaesik Park。2020年。ContraGAN：用于条件图像生成的对比学习。. 载于 神经信息处理系统进展第*33*卷：*2020*年神经信息处理系统年度会议，*NeurIPS 2020*，*2020*年*12*月*6*日至*12*日，线上会议。[18] Diederik P. Kingma 和 Jimmy Ba。2015年。Adam：一种随机优化方法。. 载于 第*3*届国际学习表征会议，*ICLR 2015*，美国加利福尼亚州圣迭戈，*2015*年*5*月*7*日至*9*日，会议论文集。[19] Walid Krichene 和 Steffen Rendle。2020年。关于物品推荐的采样指标。. 载于 *KDD '20*：第*26*届 *ACM SIGKDD* 知识发现与数据挖掘会议，线上会议，美国加利福尼亚州，*2020*年*8*月*23*日至*27*日。[20] Michael Laskin、Aravind Srinivas 和 Pieter Abbeel。2020年。CURL：用于强化学习的对比式无监督表征。. 载于 第*37*届国际机器学习会议论文集，*ICML 2020*，*2020*年*7*月*13*日至*18*日，线上会议。[21] Chao Li、Zhiyuan Liu、Mengmeng Wu、Yuchi Xu、Huan Zhao、Pipei Huang、Guoliang Kang、Qiwei Chen、Wei Li 和 Dik Lun Lee。2019年。采用动态路由的多兴趣网络及其在天猫推荐中的应用。. 载于 第*28*届 *ACM* 国际信息与知识管理会议论文集，*CIKM 2019*，中国北京，*2019*年*11*月*3*日至*7*日。[22] Yongqi Li、Meng Liu、Jianhua Yin、Chaoran Cui、Xin-Shun Xu 和 Liqiang Nie。2019年。通过时间图引导的推荐系统实现微视频路由。. 载于 第*27*届 *ACM* 国际多媒体会议论文集，*MM 2019*，法国尼斯，*2019*年*10*月*21*日至*25*日。[23] Jiasen Lu、Dhruv Batra、Devi Parikh 和 Stefan Lee。2019年。ViLBERT：面向视觉与语言任务的任务无关视觉语言表征预训练。神经信息处理系统进展 32 2019 1 11 1908 02265 ( ), –。2020 [24] arXiv：。Lei Meng、Fuli Feng、Xiangnan He、Xiaoyan Gao 和 Tat-Seng Chua。异构融合语义信息与协同信息的视觉感知食品推荐。. 载于 *MM '20*：第*28*届 *ACM* 国际多媒体会议，线上会议／美国华盛顿州西雅图，*2020*年*10*月*12*日至*16*日。[25] Jeffrey Pennington、Richard Socher 和 Christopher D. Manning。2014年。GloVe：用于词表示的全局词向量。. 载于 *2014*年自然语言处理实证方法会议论文集，*EMNLP 2014*，*2014*年*10*月*25*日至*29*日，卡塔尔多哈，*ACL* 特别兴趣小组 *SIGDAT* 会议。[26] Tao Qi、Fangzhao Wu、Chuhan Wu 和 Yongfeng Huang。2021年。基于知识感知交互匹配的个性化新闻推荐。（2021年）。arXiv：arXiv:2104.10083v3[27] Ruiyang Ren、Zhaoyang Liu、Yaliang Li、Wayne Xin Zhao、Hui Wang、Bolin Ding 和 Ji Rong Wen。2020年。采用自注意力多对抗网络的序列推荐。*SIGIR 2020*——第*43*届国际 *ACM SIGIR* 信息检索研究与开发会议论文集（2020年），89 98 10 1145 3397271 3401111 2005 10602 –。1983 [28] https://doi.org/。/。arXiv：。Paul R Rosenbaum 和 Donald B Rubin。因果效应观察性研究中倾向得分的核心作用。*Biometrika* 70，1（1983年），41–55。[29] Sara Sabour、Nicholas Frosst 和 Geoffrey E. Hinton。2017年。胶囊间动态路由。. 载于 神经信息处理系统第*30*卷：*2017*年神经信息处理系统年度会议，*2017*年*12*月*4*日至*9*日，美国加利福尼亚州长滩。[30] Ramprasaath R. Selvaraju、Michael Cogswell、Abhishek Das、Ramakrishna Vedan- tam、Devi Parikh 和 Dhruv Batra。2020年。Grad-CAM：通过基于梯度的定位实现深度网络的视觉解释。国际计算机视觉期刊（2020年）。[31] Yu Tian、Yuhao Yang、Xudong Ren、Pengfei Wang、Fangzhao Wu、Qian Wang 和 Chenliang Li。2021年。面向新闻推荐的联合知识剪枝与循环图卷积。*SIGIR 2021*——第*44*届国际 *ACM SIGIR* 信息检索研究与开发会议论文集（2021年），51–60。https://doi.org/10.1145/3404835.3462912[32] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N. Gomez、Lukasz Kaiser 和 Illia Polosukhin。2017年。注意力就是一切。. 载于 神经信息处理系统第*30*卷：*2017*年神经信息处理系统年度会议，*2017*年*12*月*4*日至*9*日，美国加利福尼亚州长滩。

[33] Dhruv Verma、Kshitij Gulati、Vasu Goel 和 Rajiv Ratn Shah。2020。Fashionist：面向冷启动场景的个性化服装搭配推荐。。发表于*MM '20*：第*28*届*ACM* 国际多媒体会议，线上会议 / 美国华盛顿州西雅图，*2020*年*10*月*12*日至*16*日。[34] Chong Wang、Lisa Kim、Grace Bang、Himani Singh、Russell Kociuba、Steven Pomerville 和 Xiaomo Liu。2020。Discovery news：财经新闻推荐的通用框架。*AAAI 2020 - 第34届 AAAI* 人工智能会议 (2020 13390 13395 10 1609 34 08 7054)，–。[35] https://doi.org/ 。/aaai.v i。Jianling Wang、Kaize Ding、Liangjie Hong、Huan Liu 和 James Caverlee。2020。基于序列超图的下一项目推荐。*SIGIR 2020 - 第43届* 国际 *ACM SIGIR* 信息检索研究与开发会议论文集 (2020)，1101–1110。https://doi.org/10.1145/3397271. 3401133[36] Peng Wang、Yunsheng Jiang、Chunxu Xu 和 Xiaohui Xie。2019。视频推荐内容点击率预测挑战赛概述。。发表于第*27*届 *ACM* 国际多媒体会议论文集，*MM 2019*，法国尼斯，*2019*年*10*月*21*日至*25*日。[37] Wenjie Wang、Fuli Feng、Xiangnan He、Liqiang Nie 和 Tat Seng Chua。2021。推荐中的隐式反馈去噪。*WSDM 2021 - 第14届 ACM* 网络搜索与数据挖掘国际会议论文集 (2021), 373 381 10 1145 3437963 3441800 2006 04153 –。[38] https://doi.org/ 。/ 。arXiv：。Xiang Wang、Xiangnan He、Meng Wang、Fuli Feng 和 Tat-Seng Chua。2019。神经图协同过滤。。发表于第*42*届国际 *ACM SIGIR* 信息检索研究与开发会议论文集，*SIGIR 2019*，法国巴黎，*2019*年7月*21*日至*25*日。[39] Zhenlei Wang、Jingsen Zhang、Hongteng Xu、Xu Chen、Yongfeng Zhang、Wayne Xin Zhao 和 Ji Rong Wen。2021。反事实数据增强序列推荐。*SIGIR 2021 - 第44届国际 ACM SIGIR* 信息检索研究与开发会议论文集 (2021 347 356 10 1145 3404835 3462855)，–。https://doi.org/ 。/ 。[40] Yinwei Wei、Xiang Wang、Liqiang Nie、Xiangnan He 和 Tat-Seng Chua。2020。用于多媒体推荐的图精炼卷积网络与隐式反馈。。发表于*MM '20*：第*28*届 *ACM* 国际多媒体会议，线上会议 / 美国华盛顿州西雅图，*2020*年*10*月*12*日至*16*日。[41] 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。。发表于第*27*届 *ACM* 国际多媒体会议论文集，*MM 2019*，法国尼斯，*2019*年*10*月*21*日至*25*日。[42] Chuhan Wu、Fangzhao Wu、Mingxiao An、Jianqiang Huang、Yongfeng Huang 和 Xing Xie。2019。NPA：基于个性化的神经新闻推荐final

注意力机制。在第*25*届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集，*KDD 2019*，*2019*年*8*月，阿拉斯加州安克雷奇，美国。[43] Fangzhao Wu、Ying Qiao、Jiun-Hung Chen、Chuhan Wu、Tao Qi、Jianxun Lian、Danyang Liu、Xing Xie、Jianfeng Gao、Winnie Wu等。2020年。MIND：用于新闻推荐的大规模数据集。在第*58*届计算语言学协会年会论文集，*ACL 2020*，线上，*2020*年*7*月*5*日至*10*日。[44] Xu Xie、Fei Sun、Zhaoyang Liu、Jinyang Gao、Bolin Ding和Bin Cui。2020年。面向序列推荐的对比式预训练。*CoRR*（2020年）。[45] Xuewen Yang、Dongliang Xie、Xin Wang、Jiangbo Yuan、Wanying Ding和Pengyun Yan。2020年。学习用于条件式穿搭推荐的元组兼容性。在*MM '20*：第*28*届*ACM*国际多媒体会议，线上会议/美国华盛顿州西雅图，*2020*年*10*月*12*日至*16*日。[46] Xuzheng Yu、Tian Gan、Yinwei Wei、Zhiyong Cheng和Liqiang Nie。2020年。面向二手交易平台的个性化物品推荐。在*MM '20*：第*28*届*ACM*国际多媒体会议，线上会议/美国华盛顿州西雅图，*2020*年*10*月*12*日至*16*日。[47] Ningyu Zhang、Qianghuai Jia、Shumin Deng、Xiang Chen、Hongbin Ye、Hui Chen、Huaixiao Tou、Gang Huang、Zhao Wang、Nengwei Hua和Huajun Chen。2021年。AliCG：阿里巴巴语义搜索的细粒度、可演进概念图构建。在第*27*届*ACM SIGKDD*知识发现与数据挖掘会议，线上会议，新加坡，*2021*年*8*月*14*日至*18*日。[48] Quanshi Zhang、Ying Nian Wu和Song-Chun Zhu。2018。可解释的卷积神经网络。在*2018 IEEE*计算机视觉与模式识别会议，*CVPR 2018*，美国犹他州盐湖城，*2018*年*6*月*18*日至*22*日。[49] Shengyu Zhang、Dong Yao、Zhou Zhao、Tat Seng Chua和Fei Wu。2021年。CauseRec：用于序列推荐的反事实用户序列合成。*SIGIR 2021*——第*44*届*ACM*信息检索研究与开发国际会议论文集（2021年），367–377。https://doi.org/ 10 1145 3404835 3462908 2109 05261 。/。arXiv：。Xinran Zhang、Xin Yuan、Yunwei Li和Yanru Zhang。冷启动表征学习：一种结合Bert4Movie和Movie2Vec的推荐方法。在第*27*届*ACM*国际多媒体会议论文集，*MM 2019*，法国尼斯，*2019*年*10*月*21*日至*25*日。[51] Chang Zhou、Jianxin Ma、Jianwei Zhang、Jingren Zhou和Hongxia Yang。2020年。面向大规模推荐系统去偏候选生成的对比学习。*arXiv*预印本 *cs.IR/2005.12964*（2020年）。[50] 2019