arXiv:2603.00854v1 [cs.LG] 1 Mar 2026

GeMi：一种基于图的、多模态长卷叙事画推荐系统

Haimonti Dutta[1,2*], Pruthvi Moluguri[3], Jin Dai[1], Saurabh AmarnathMahindre[4]

[1]管理科学与系统系，纽约州立大学，Jacobs管理中心，布法罗，14260，纽约州，美国。[2]人工智能与数据科学研究所，纽约州立大学，Lockwood Hall，布法罗，14260，纽约州，美国。[3]计算机科学与工程系，纽约州立大学，Davis Hall，布法罗，14260，纽约州，美国。[4]科技、信息、互联网，eBay*，Embassy科技园，班加罗尔，560103，卡纳塔克邦，印度。

*通讯作者。电子邮箱：haimonti@buffalo.edu；贡献作者：pmolugur@buffalo.edu；jindai@buffalo.edu；smahindr@buffalo.edu；

摘要

推荐系统能够有效管理当今不断增长的多模态数据，并帮助用户发现有趣的新物品。这类系统可以处理图像、文本、音频和视频数据等各种媒体类型，因此能够利用从物品中提取的特征进行基于内容的推荐，同时纳入用户偏好。基于图神经网络（GNN）的推荐系统是一类特殊的推荐系统，能够处理物品与用户之间的关系，因此尤其适用于基于内容的推荐。它们之所以广受欢迎，也在于其运用图结构数据上的深度学习等先进机器学习技术，充分利用用户与物品的交互。图中的节点能够获取高阶邻居信息，并结合最先进的视觉语言模型处理多模态内容；此外，也已有精心设计的算法用于

*本研究是在作者加入其现任机构之前完成的。

嵌入、消息传递和传播。在本研究中，我们介绍了一种基于图神经网络的推荐系统的设计，该系统采用了通过实地研究收集的新型数据集。该推荐系统面向一种濒危的 *performing* 艺术形式，利用多模态内容（文本和图像数据）推荐相似的绘画作品，供人们观赏和购买。据我们所知，目前尚无专为叙事性手卷画设计的推荐系统，因此我们的工作具有多重意义，包括艺术保护、为濒危艺术品建立数据存储系统，以及构建一种最先进的推荐系统，充分利用数据的新颖特征和对叙事性手卷画感兴趣的用户群体的偏好。

关键词： 图神经网络、基于内容的推荐、多模态视觉语言模型、民间艺术推荐

1 引言

推荐系统（RS）（Deldjoo 等人（2020），Zhou 等人（2023a））帮助组织用户可获取的大量物品（如书籍、音乐、家电）。它们通过预测用户对候选物品的偏好，并利用历史数据推荐最相关的物品，从而找到用户感兴趣的物品。数据驱动的推荐系统主要分为三类：协同过滤（CF）、基于内容的过滤（CBF）和混合式。在协同过滤模型中，重点在于用户群体及其行为模式和偏好。基于内容的方法依赖用户的行为数据以及物品信息，例如文本文件中的词语、图像中的颜色和音频内容。一些推荐系统将协同过滤方法与基于内容的方法结合，以获得额外优势。这三类推荐系统都广泛使用多模态数据（图像、文本、音频、视频）。

推荐系统中整合不同模态数据的传统方法包括拼接或求和（He and McAuley（2016）；Liu 等人（2017））。然而，随着具备非线性建模能力的深度学习技术兴起，各模态可以分别表示，随后再进行组合，以生成合适的表示。近年来，基于图的推荐系统日益受到关注，催生了一系列新研究，利用图神经网络的强大能力探索用户与物品之间的高阶交互（He等人（2020）；Wu等人（2023）；Zhang 等人（2021a，2022）；Zhou等人（2023c，b）；Wei 等人（2019，2020）；Wang 等人（2023a）；Zhou等人（2023b））。例如，MMGCN（Wei 等人（2019））通过构建具有拓扑特征节点的图卷积网络，并聚合邻居的特征，利用用户—物品交互引导各模态的表征学习。GRCN（Wei 等人（2020））在此基础上改进，设计了图精炼层，以高置信度识别与用户偏好相关的噪声边。然而，由于用户—物品交互稀疏，通过图聚合学习到的用户和物品表示往往不可避免地带有噪声。这些系统还面临冷启动问题，即缺乏足够的数据（如点击、评分或购买记录）来学习用户偏好。

解决用户-物品稀疏问题的一种方法是强化节点（用户和物品）特征。然而，节点中存在的噪声往往使得难以提取对推荐有用的信息。物品特征中的噪声（例如模糊的图像、文本中的拼写错误、受损的音轨或误导性的描述）会导致表示存在不确定性。物品特征还可能包含无关信息，例如背景颜色或图像中分散注意力的物体。在多模态场景中，人们利用不同模态的互补信息来应对稀疏问题，但噪声特征可能导致不理想的结果。如果某些模态缺失或不存在，多模态系统必须对此加以适应。为应对多模态学习中的这些挑战，变分自编码器（Yi 和 Chen (2022); Liang 等人 (2018); Ma 等人 (2019); Li 和 She (2017)) 已广受欢迎。在这类方法中，模态特定编码器会为每个节点学习一个由均值向量和方差向量参数化的高斯变量。均值向量用于表示语义信息，方差向量则捕捉相应模态的不确定性。

本文介绍了一种结合用户偏好信息的多模态基于内容的推荐系统，专为一种独特的社会文化应用而设计，旨在保护并研究一种源自印度东部的濒危表演艺术。歌唱画师是一群流动说书人，他们绘制卷轴画，并通过歌曲或诗歌讲述艺术作品。其艺术创作通常以史诗、地方民间故事以及画师日常生活中发生的事件为灵感（McCutchion 和 Bhowmick (1999年)）。歌曲或诗歌通常通过口耳相传的方式习得。人们曾零星地、临时性地尝试以非数字化形式保存艺术作品样本或歌曲文本，但据我们所知，尚无人对这些由图像和文本构成的多模态数据进行细致研究，也无人设计推荐系统对其进行整理。在近期的研究中（Dutta 和 Tiwari(2024)），我们制作了一个定制数据集，其中包含数字化图像，将歌曲歌词与图像相对应，并在相当程度上记录了风格、主题、颜色选择等相关细节的地区差异。我们使用这个数据集的增强版本[1] 来设计推荐引擎，从而推荐与爱好者喜欢的艺术作品相似的作品。在此过程中我们发现，尽管已有许多技术用于构建针对已知基准数据集（如亚马逊（Ni 等人(2019年)）、MovieLens（Harper 和 Kon- stan (2015年)）或 Netflix（Bennett 和 Lanning (2007年)）的多模态推荐系统，但如果不进行审慎分析和有原则的实验，将这些技术应用于全新的数据集很快就会变得繁琐。我们报告了多模态推荐系统 GeMi 的设计成果，并将其与文献中的先进系统进行了审慎的基准比较。本研究所用的数据集可从以下地址获取：https://github.com/Haimonti/Folk-Art-Recommendation-System。

本文的主要贡献如下：

- 通过实地调查收集并整理用于研究的独特多模态数据集
- GeMi 推荐系统——采用先进视觉语言模型并进行复杂多模态融合的基于内容的推荐，图

[1]该州其他地区后续实地研究收集了更多数据。

结构学习，并通过协同过滤风格的即插即用模块融入用户偏好。

本文结构安排如下：第2 节介绍相关工作；第3节介绍数据收集流程以及非数字化来源；第4节介绍推荐系统；第5节展示 GeMi 与最先进的传导式和归纳式同质图及异构图方法的性能对比实证结果；第6节讨论可能的改进；第7 节总结全文。

2 相关工作

本节介绍艺术推荐领域的先前研究，以及叙事性手卷画的简要历史。附录A 回顾了多模态数据、变分自编码器（VAEs）和大语言模型（LLMs）在增强图神经网络（GNNs）以构建推荐系统中的应用。

2.1 艺术推荐

博物馆已认识到，游客参观馆藏时需要获得个性化体验，因此采用推荐系统来满足这一需求。例如，游客档案可能会记录其喜爱的艺术家姓名或绘画技法，这些信息提取自与艺术品相关的简短文本描述（Wang等人（2008, 2009））。随后，系统会将用户档案与物品特征相匹配，提供个性化推荐。文化遗产信息个性化（CHIP）项目（Aroyo等人（2007））为荷兰的荷兰国立博物馆构建了一个推荐引擎。在基于大众分类法的物品推荐系统（FIRSt）（Semeraro等人（2012））中，社交标签产生的用户生成内容被整合到基于内容的推荐系统中，用户既可以输入数值评分来表达对物品的偏好，也可以为物品添加自由标签。在主动文化遗产个性化体验（PEACH）项目（Stock等人（2007）；Messina等人（2019）；Kuflik等人（2011））中，开发了一款基于个人数字助理（PDA）的博物馆导览应用，该应用具备位置感知能力，其内容也会根据游客的需求进行调整。根据博物馆游客在物理环境中的活动历史，以及与其历史记录中各项目相关的文本信息来提供推荐的方法，已相当普遍（Bohnert和Zukerman（2009）；Grieser等人（2011））。Yilma等人（Yilma和Leiva（2023））着重研究如何高效地捕捉视觉艺术中的潜在语义关系，以实现个性化推荐。他们表明，文本特征与视觉特征同样有用，但融合两者才能捕捉到大部分潜在语义信息。Kuflik等人（Kuflik等人（2011））展示了采用基于图的推荐的优势。一种名为“我不喜欢的艺术”（Frost等人（2019））的反推荐系统，会推荐与用户喜爱的艺术作品不同的作品，这表明，限制人们接触相反观点可能会导致过滤气泡（Pariser（2011））和思想孤立。然而，据我们所知，迄今尚不存在专为叙事性手卷画设计的艺术推荐系统。

2.2 南亚叙事性手卷画传播简史

叙事性手卷画是印度东部一种流行的表演艺术形式。它融合了多种传播方式——视觉信息、口述传统和音乐——借此讲述社会变迁、迁徙故事，以及社会政治与宗教反思（Korom (2006); Bajpai (2015)）。尽管这一行当确切的起源时间尚不清楚，但古代传说与传统表明，它可追溯至5至6世纪（甚至可能更早）（Mair (2019）。

这些来自印度东部的绘画已传播至该国其他地区，包括西部和南部地区（如德干地区说画艺人、马哈拉施特拉邦 Chitrakathis 的“派坦”绘画、古吉拉特邦“加罗达”说画艺人、特伦甘纳邦的叙事绘画以及拉贾斯坦邦的 Bhopas）(Jain(1998))。此外，面对东部的外来入侵，许多长卷画传播到了尼泊尔、西藏以及丝绸之路沿线的其他邻近地区。这些传统也被用于传播宗教教导，并传入了中国。据悉，说画技艺最迟于16世纪末传至土耳其语世界的西部腹地，也可能更早。Evliy¯a Efendi（约1611—1660年）在其关于伊斯坦布尔生活的记述中，对一群被称为“画家占卜师”的人作了引人入胜的描述。还应指出，在印度尼西亚，戏剧化表演广为人知，称为哇扬——一种用于宗教教导的转化性体验。Wayang 意为“影子”，因此，手卷画、木偶、影子和舞者在技术上都属于哇扬。引用 Mair（Mair(2019年)）的话：“印度元素对于（中国）变文的形成至关重要，而哇扬显然也深受印度叙事和戏剧的影响。”遗憾的是，这种艺术形式鲜有记载或数字化保存。

3 数据

在近期的研究中，第一作者在印度东部进行了广泛的实地调查[2]，从而得以将卷轴画、表演视频和音频录制数字化，随后对歌曲歌词进行转写、转写和翻译（原歌曲为孟加拉语）。这项实地研究收集的数据被用于设计首个面向艺术爱好者和策展人的基于内容的推荐系统版本（Dutta 和 Tiwari(2024）。第二次实地调查于一年后开展，帮助扩充了资料库中的数据。需要注意的是，第一次调查收集的图像和文本彼此匹配良好，而第二次调查收集的内容则并非如此。数据缺失（歌曲文本不可用或缺少画作）给多模态数据的对齐带来了诸多挑战。所收集的数据版权仍归艺术家和歌曲创作者所有，但位于纽约州立大学布法罗分校的机构知识库已获得许可和同意，可以对其进行归档和存储。

[2]这涉及调查比尔布姆、班库拉和普等县的村庄鲁利亚，位于西孟加拉邦。

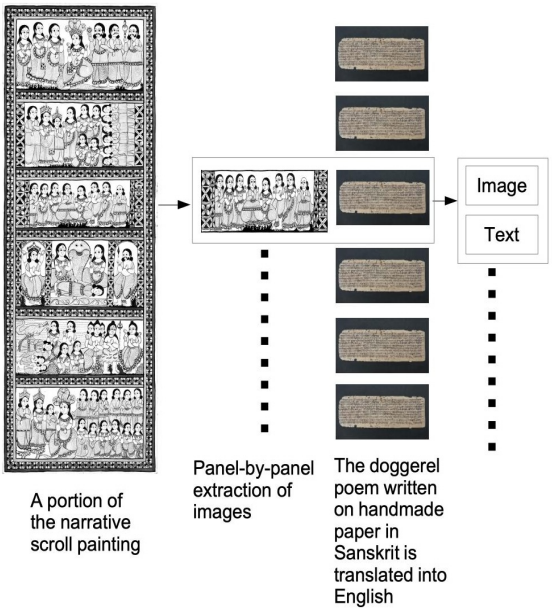


图 1：逐面板提取叙事卷轴画中的图像，并与写在手工纸上的古代打油诗文本进行对齐。该诗原始手稿以黑墨书写，许多部分已缺失或遗失。

3.1 现有的未数字化数据来源

• **文本：** 目前没有已知的歌曲数据库可用于设计推荐系统。因此，歌曲及其歌词必须通过实地研究手动收集。不过，我们知道有若干来源曾初步尝试收集这些歌曲：（1）书籍（McCutchion 和 Bhowmick (1999年)、Dutta (1939)、Dimock (1963)）以及学位论文（Yazijian (2007年)；（2）个人收藏，Korom博士（Korom (2006)）——这些资料以磁带形式保存，目前尚未数字化。

• **图像：** 现存尚未数字化的叙事性手卷画收藏于大英博物馆、国际民间艺术博物馆（新墨西哥州圣菲）、维多利亚与艾尔伯特博物馆（英国）以及Gurusaday博物馆（印度加尔各答）。然而，需要指出的是，在大多数情况下，将图像与歌曲进行对齐并非易事，因为零星开展的保护工作目标各异，且可能未包含本研究所需的全部相关数据。

4 推荐系统

图 1展示了多模态数据提取流程，图 2 呈现了推荐系统的整体架构。该系统包含以下分量：（a）经过微调、用于多模态特征提取的视觉语言模型；（b）图结构学习；（c）节点分类；以及（d）下游基于内容的推荐。以下各节将对这些内容进行更详细的介绍。

预备知识 我们的基本前提是，面板和歌曲的推荐应反映用户感兴趣的概念。我们研究用户-面板偏好矩阵，以开发推荐算法。设 $\mathcal{U}$，$\mathcal{P}$ 分别表示用户和面板的集合。每个用户 $u_i \in \mathcal{U}$ 都关联一组其喜欢的面板 $p_j \in \mathcal{P}$，$i \in 1, \cdots, |\mathcal{U}|$；$j \in 1, \cdots, |\mathcal{P}|$。图1 展示了逐画格图像提取的示意图。偏好分数 $y_{ij} = 1$表示，如果用户 u_i喜欢面板 p_j，则取相应值，否则为 0。每个面板 p_j 都有一个多模态表示，即 $p_j \in R^{D_m}$，其中 D_m 表示模态 $m \in \mathcal{M}$ 的维度，$\mathcal{M}$ 是模态集合。多模态推荐旨在根据用户已知的偏好，准确找出用户可能喜欢的面板。

在以下小节中，我们将介绍 GeMi 推荐系统的各个组成部分。

4.1 用于多模态特征提取的微调视觉语言模型

我们介绍一种多模态特征提取框架，该框架将叙事性手卷画相关的图像和文本转换为语义结构化的高维表示，以用于下游的基于图的推荐。

鉴于叙事性手卷画固有的符号密度、叙事连贯性和文化特殊性，单模态表示不足以捕捉数据的完整语义频谱。因此，我们采用模块化的多模态编码策略，融合语言建模、跨模态对比学习和概率潜变量建模所带来的互补归纳偏置。

形式化地，令每个面板由元组 $\mathcal{P}_i = (x_i^{\text{img}}, x_i^{\text{text}})$, 表示，其中 $x_i^{\text{img}} \in \mathcal{X}_{\text{img}}$ 表示面板的视觉描绘，$x_i^{\text{text}} \in \mathcal{X}_{\text{text}}$ 表示其相关的文本描述。我们的目标是构建一组嵌入函数

$$\Phi^{(k)} : \mathcal{X}_{\text{img}} \times \mathcal{X}_{\text{text}} \to \mathbb{R}^{d_k}, \tag{1}$$

每个函数捕捉同一面板内容的不同语义视图，这些视图随后通过基于图的传播进行整合。为此，我们采用三种概念上各不相同但彼此互补的表示机制：

1. 一种大语言模型（大语言模型）（LLaMA，Touvron 等人 （2023）），用于文本描述的语义抽象和规范化。2. 一种 Sigmoid 对比语言—图像预训练模型（Sig-CLIP，Radford 等人 （2021）），用于联合图文对齐，

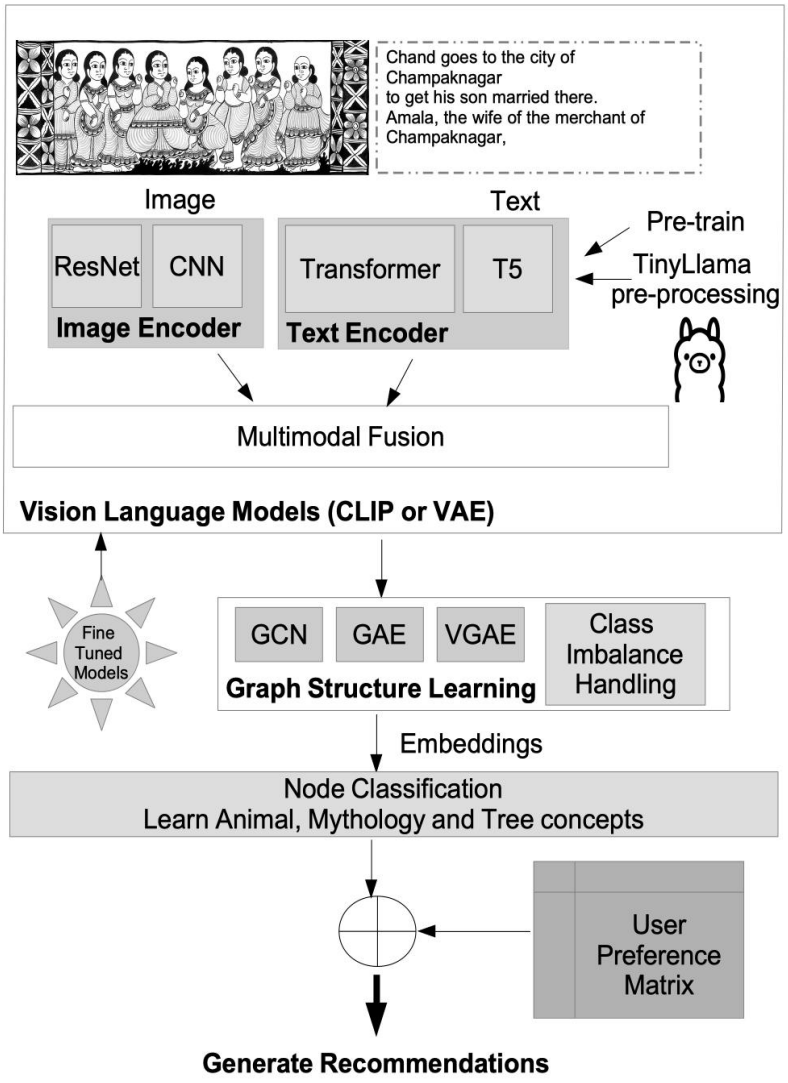


图 2：GeMi 推荐系统。使用微调后的视觉语言模型提取特征、学习潜在图结构和嵌入，并通过图卷积进行处理。用户-面板偏好矩阵用于增强推荐效果。

3. 一种多模态变分自编码器（变分自编码器），用于视觉与语言信息的概率融合。

4.1.1 基于大语言模型的语义文本规范化

与叙事性手卷画相关的文本描述往往含有噪声、风格各异且语义信息不充分。为减少词汇差异，同时保留高层语义意图，我们采用大语言模型作为确定性的语义规范化算子，在进行多模态融合之前，将原始文本改写为简洁、以概念为中心的表示。

我们采用 LLaMA 系列的自回归仅解码器 Transformer，通过因果语言建模对 token 序列 $\mathbf{x} = (x_1, \ldots, x_T)$ 进行建模：$P(\mathbf{x}) = \prod_{t=1}^{T} P(x_t \mid x_1, \ldots, x_{t-1})$，并采用多头因果自注意力实现。LLaMA 采用预归一化 Transformer 架构，包含 RMSNorm、旋转位置嵌入（RoPE，Su 等人(2024)) 和 SwiGLU 前馈模块（Shazeer (2020)），从而在受限的推理预算下提升稳定性和语义表达能力。我们选择 TinyLlama（Zhang 等人(2024)），原因在于其精简的参数规模（1.1B）使其无需量化即可在我们的计算环境的内存限制下进行推理，同时其 LLaMA-2 架构保留了足够的语义表达能力，能够胜任规范化任务；该任务要求进行实体识别和词汇规范化，而非开放式生成。令 $\mathbf{h}^{(0)}$ 表示经 RoPE 增强的输入 token 嵌入。每个 Transformer 层 $\ell \in \{1, \ldots, L\}$ 计算 $\mathbf{h}^{(\ell)} = \mathrm{Block}^{(\ell)}\big(\mathbf{h}^{(\ell-1)}\big)$，最终隐藏状态则通过线性投影和 Softmax，确定词汇表上的分类分布。

在我们的流程中，语言模型并不用作直接的特征编码器。相反，~我们将其形式化为一个确定性改写函数 $\mathcal{R}_{\mathrm{LLaMA}} : \mathcal{X}_{\mathrm{text}} \to \mathcal{X}_{\mathrm{text}}$，该函数将原始画面描述映射为语义规范化文本形式，强调实体、动作和象征性母题，同时抑制表层语言噪声。对于每个面板 i，改写后的文本由 $\tilde{x}_i^{\mathrm{text}} = \mathcal{R}_{\mathrm{LLaMA}}(x_i^{\mathrm{text}})$ 给出。

规范化文本 $\tilde{x}_i^{\mathrm{text}}$随后被传递给下游多模态编码器（SigCLIP 和多模态变分自编码器），在其中作为对比对齐与潜在空间融合的语言输入。通过在文本层面强制语义一致性，这一改写阶段降低了跨模态噪声，并在不改变下游编码器架构的情况下提升了所学图文表征的质量。

4.1.2 对比语言—图像预训练（CLIP）模型及其变体

在对文本描述进行语义规范化后，我们通过对比式视觉—语言预训练学习联合图像—文本嵌入空间。目标是将视觉面板及其关联文本嵌入共享的潜在空间，使语义对齐的图像—文本对在几何上彼此接近，而不匹配的配对则相互远离。令 $\mathcal{D} = \{(I_i, T_i)\}_{i=1}^{N}$ 表示由配对图像 $I_i \in \mathcal{I}$ 和文本$T_i \in \mathcal{T}$构成的语料库。我们学习双编码器 $f_\theta : \mathcal{I} \to \mathbb{R}^d$，$g_\phi : \mathcal{T} \to \mathbb{R}^d$，将每种模态映射到共同的 d维嵌入空间。该架构采用双塔设计：编码器联合训练，但在推理时独立运行。

给定一个小批量 $\mathcal{B} = \{(I_i, T_i)\}_{i=1}^{B}$，我们得到经过 ℓ_2归一化的嵌入 $\mathbf{x}_i = f_\theta(I_i) \| f_\theta(I_i) \|_2$、$\mathbf{y}_i = g_\phi(T_i) \| g_\phi(T_i) \|_2$，其中 $\mathbf{x}_i, \mathbf{y}_i \in \mathbb{R}^d$。图像—文本之间的成对相似度通过余弦相似度衡量，即 $s_{ij} = \mathbf{x}_i^\top \mathbf{y}_j$，并由温度参数 $t > 0$, $\ell_{ij} = t\, s_{ij}$缩放。CLIP目标将图像到文本和文本到图像的匹配视为对称的分类问题。定义 $p_{ij}^{I \to T} = \exp(\ell_{ij}) / \sum_{k=1}^{B} \exp(\ell_{ik})$、$p_{ij}^{T \to I} = \exp(\ell_{ij}) / \sum_{k=1}^{B} \exp(\ell_{kj})$。由此得到的对比损失为

$$\mathcal{L}_{\mathrm{CLIP}} = -\frac{1}{2B} \sum_{i=1}^{B} \left(\log p_{ii}^{I \to T} + \log p_{ii}^{T \to I} \right),$$

该方法最大化匹配图文对之间的一致性，同时隐式地将批次中的所有其他样本作为负样本。

多模态嵌入构建： 设 $\mathcal{M}$表示一个经过训练的 CLIP系列编码器，其参数为 (θ, ϕ)。对于面板 (I_i, T_i)，我们定义统一的多模态嵌入算子 $\mathcal{E}_{\mathcal{M}} : (\mathcal{I}, \mathcal{T}) \to \mathbb{R}^d$，其形式为对归一化图像和文本嵌入进行简单的特征融合：$\mathbf{e}_i^{(\mathcal{M})} = \mathcal{E}_{\mathcal{M}}(I_i, T_i) = \frac{1}{2}(\mathbf{x}_i + \mathbf{y}_i)$。

文本描述在长度和叙事结构方面差异很大。为获得对局部措辞效应具有鲁棒性的表示，我们将文本编码建模为多个子上下文上的期望。给定一个长度为 $|T_i|$ 的文本序列 T_i，我们从所有有效子串中均匀采样 C 个固定最大长度为 s 的连续子串 $\{T_i^{(c)}\}_{c=1}^{C}$。最终的多模态表示定义为

$$\mathbf{e}_i^{(\mathcal{M})} = \frac{1}{C} \sum_{c=1}^{C} \mathcal{E}_{\mathcal{M}}(I_i, T_i^{(c)}),$$

该表示可以解释为对文本子上下文上的嵌入进行边缘化的蒙特卡洛近似。

CLIP变体：我们提出了三种 CLIP 风格编码器变体。它们的架构骨干网络和嵌入构建流程完全相同，区别仅在于训练目标以及用于获得编码器权重的参数化方式。**预训练 CLIP** 第一种变体采用了一个模型，该模型使用基于 softmax 的对比目标，在大规模网络图文对上进行训练。令 (θ_0, ϕ_0) 表示通过这一预训练过程获得的参数。由此得到的嵌入 $\mathbf{e}_i^{\text{CLIP-pre}} = \mathcal{E}_{(\theta_0, \phi_0)}(I_i, T_i)$ 提供了强大的通用多模态对齐先验，并作为领域自适应的基线。

数据集特定softmax对比微调CLIP 此变体对应于一个 CLIP 模型，其参数（$\theta_{\mathrm{ft}}, \phi_{\mathrm{ft}}$）通过仅使用从我们的数据集中抽取的图像-文本对进行对比训练而获得。这种数据集特定微调明确地使表示空间与叙事性手卷画特有的风格、图像志和语言分布相对齐，同时保留 CLIP 原始跨模态对齐机制的归纳偏置。由此得到的多模态表示，$\mathbf{e}_i^{\text{CLIP-ft}} = \mathcal{E}_{(\theta_{\mathrm{ft}}, \phi_{\mathrm{ft}})}(I_i, T_i)$ 相较于预训练基线，对数据集特定语义表现出更强的敏感性，

从而为后续的基于图的传播和标签条件推荐奠定了更坚实的基础。

数据集特定的 Sigmoid 对比微调 CLIP（SigCLIP）该变体采用与前述模型相同的双编码器 CLIP 架构，但在微调过程中以基于 sigmoid 的成对损失取代基于 softmax 的对比目标。与基于 softmax 的对比学习不同，后者会在整个批次范围内对相似度进行归一化，而 sigmoid 对比形式将每个图像–文本配对视为一个独立的二元分类问题。设$\mathbf{x}_i$ 和$\mathbf{y}_j$分别表示编码器生成的经 ℓ_2归一化的图像和文本嵌入。配对(i, j)的相似度 logit 定义为 $\ell_{ij} = t\,\mathbf{x}_i^\top \mathbf{y}_j + b$，其中 $t > 0$是可学习的温度参数， b 是可选的可学习偏置项，用于在类别严重不平衡时稳定优化。成对标签定义为

$$z_{ij} = \begin{cases} +1, & \text{if } i = j \quad \text{(matched image–text pair)}, \\ -1, & \text{if } i \neq j \quad \text{(mismatched pair)}. \end{cases}$$

微调过程中优化的 sigmoid 对比目标为

$$\mathcal{L}_{\text{SigCLIP}} = -\frac{1}{B} \sum_{i=1}^{B} \sum_{j=1}^{B} \log \sigma(z_{ij}\, \ell_{ij}),$$

其中， $\sigma(\cdot)$表示逻辑 Sigmoid 函数。该损失函数取消了批次归一化，转而通过独立的成对决策来实现对齐，从而使批次内各样本的梯度贡献相互解耦。由于架构和嵌入算子保持不变，最终得到的多模态表示$\mathbf{e}_i^{\text{SigCLIP-ft}} = \mathcal{E}_{(\theta_{\text{sig}}, \phi_{\text{sig}})}(I_i,\ T_i)$与预训练变体和经 softmax 微调的变体之间的差异，仅源于对比学习信号的改变。

4.1.3 多模态变分自编码器

我们提出一种多模态变分自编码器（VAE），它学习基于视觉和文本模态的概率潜在表示。我们的公式遵循自编码变分贝叶斯框架（AEVB，Zhi-Han(2022)），并将其扩展到异构多模态输入；在下游评估中，该模型用作冻结特征提取器。

对于每个叙事面板 i，令 x_i^{img} 表示图像模态， x_i^{text} 表示其关联的文本描述。我们假设一个连续潜变量 $\mathbf{z}_i \in \mathbb{R}^d$，它能够生成这两种模态。联合分布分解如下：

$$p(\mathbf{z}_i) = \mathcal{N}(\mathbf{0}, \mathbf{I}), \qquad p_\theta(x_i^{\text{img}}, x_i^{\text{text}} \mid \mathbf{z}_i) = p_\theta(x_i^{\text{img}} \mid \mathbf{z}_i)\, p_\theta(x_i^{\text{text}} \mid \mathbf{z}_i),$$

其中，给定$\mathbf{z}_i$ 时的条件独立性使多模态建模易于处理，同时允许每个解码器专注于各自的模态。精确的后验推断$p(\mathbf{z}_i \mid x_i^{\text{img}}, x_i^{\text{text}})$难以实现。因此，我们引入模态特定的变分

近似，每个近似都参数化为对角协方差高斯分布：

$$q_\phi^{\text{img}}(\mathbf{z}_i \mid x_i^{\text{img}}) = \mathcal{N}\Big(\boldsymbol{\mu}_i^{\text{img}}, \operatorname{diag}\big(\boldsymbol{\sigma}_i^{\text{img}}\big)^2\Big),$$
$$q_\phi^{\text{text}}(\mathbf{z}_i \mid x_i^{\text{text}}) = \mathcal{N}\Big(\boldsymbol{\mu}_i^{\text{text}}, \operatorname{diag}\big(\boldsymbol{\sigma}_i^{\text{text}}\big)^2\Big).$$

这些编码器分别提取关于潜在语义状态$\mathbf{z}_i$的模态条件证据。

4.2 模态融合

为整合来自不同模态的信息，我们采用专家乘积（PoE）形式。融合后的变分后验定义为

$$q_\phi(\mathbf{z}_i \mid x_i^{\text{img}}, x_i^{\text{text}}) \;\propto\; q_\phi^{\text{img}}(\mathbf{z}_i \mid x_i^{\text{img}})\; q_\phi^{\text{text}}(\mathbf{z}_i \mid x_i^{\text{text}}),$$

当所有专家均为高斯分布时，该分布仍为高斯分布。令 $\boldsymbol{\Lambda}_i^{(m)} = \operatorname{diag}\big(\boldsymbol{\sigma}_i^{(m)}\big)^{-2}$ 表示模态 $m \in \{\text{img},\text{text}\}$ 的精度矩阵。融合后的后验参数可通过以下闭式解求得

$$\begin{aligned}
\boldsymbol{\Lambda}_i &= \boldsymbol{\Lambda}_i^{\text{img}} + \boldsymbol{\Lambda}_i^{\text{text}}, \\
\boldsymbol{\mu}_i &= \boldsymbol{\Lambda}_i^{-1}\Big(\boldsymbol{\Lambda}_i^{\text{img}}\boldsymbol{\mu}_i^{\text{img}} + \boldsymbol{\Lambda}_i^{\text{text}}\boldsymbol{\mu}_i^{\text{text}}\Big), \\
\boldsymbol{\Sigma}_i &= \boldsymbol{\Lambda}_i^{-1}
\end{aligned}$$

这种融合机制实现了不确定性感知整合：后验方差较小的模态对共享潜在表示的影响更大。对于每个面板，我们使用融合分布的后验均值作为确定性特征表示：

$$\mathbf{e}_i^{\text{VAE}} = \boldsymbol{\mu}_i$$

这种选择消除了采样方差，同时保留了训练过程中学到的不确定性感知融合能力。

4.3 概念（或目标标签）学习

要求20名经过培训的标注者识别一个面板中是否存在以下三种不同的概念：(a)动物；(b)树木；(c)神话人物。其目的是了解人类识别这些概念的能力。使用克里彭多夫 α 系数(Krippendorff (2004))来衡量标注者之间的一致性。不同概念的 α 系数值如下：(a) 神话人物：0.69；(b) 树木：0.85；(c) 动物：0.69。

4.4 图结构学习（图结构学习）

4.4.1 学习同质图

为了从面板中发现图的潜在结构，我们首先使用多模态特征来学习 $|\mathcal{P} \times \mathcal{P}|$ 相似度矩阵。基于相似面板相比不相似面板更可能引起终端用户兴趣这一假设 McPherson 等（2003），我们使用余弦相似度（Wang 等（2020）、Zhu 等（2021）、Zhang 和 Zitnik（2020）、Chen 等（2020））、基于核的函数（Li 等（2018））或注意力机制（Chen 等（2020））等相似度度量来量化面板之间的语义关系。在本研究中，我们采用无需参数的余弦相似度，使相似度 $S_m = (e_i^m{}^T e_j^m)$，其中 e_i^m、e_j^m表示多模态特征向量。由于图的邻接矩阵包含正值，而完全连通的完全图通常并不理想，因此我们抑制 S_m中的负值。此外，鉴于真实世界图具有噪声且可能包含许多与任务无关的边，我们施加了稀疏性约束。借鉴 ϵ-邻近阈值图（ϵ-图 Bentley（1975））的思想，当两个节点之间的相似度小于预先设定的、由数据驱动的阈值 ϵ时，我们就在它们之间创建一条边。

4.4.2 传导式与归纳式图学习

在图上训练神经网络可以在传导式设置或归纳式设置中进行（Xu 等人 (2020)；Lachaud 等人 (2023)）。在传导式设置中，模型在训练阶段可以访问测试特征。在归纳式设置中，测试数据保持不可见。我们主要探讨归纳学习与直推式学习之间的差异，因为在不同时间、不同地理位置开展的实地研究持续不断地向我们的语料库添加数据，因此需要重新训练模型，并依赖归纳学习。

4.5 节点分类

设 $\mathcal{G} = (A, X)$ 表示一个图，其中 $A \in R^{N \times N}$ 是邻接矩阵，$X \in R^{N \times F}$ 是节点特征矩阵，其第 i 项 $x_i \in R^F$ 表示节点 v_i 的特征。给定 $\mathcal{G}$，图结构学习的目标是同时学习邻接矩阵A^* 及其对应的节点表示，以用于下游任务。图结构学习模型有两个可训练分量：（1）图神经网络编码器 f_θ，它以图作为输入并生成嵌入。（2）结构学习模块 g_ϕ，其中包含图的边连接信息。

模型参数（θ，ϕ）通过以下学习目标进行训练：

$$\mathcal{L} = \mathcal{L}_{tsk}(Z^*, Y) + \lambda \mathcal{L}_{reg}(A^*, Z^*, \mathcal{G}) \tag{2}$$

其中，右侧第一项表示原始任务，第二项表示对学习所得图的正则项，而 λ 是需要调节的超参数。

因此，节点分类可根据图中节点的属性及其关系，为节点分配标签（例如某个概念是否存在）。

分配给节点（面板）的标签用于推荐系统中的相关性估计。

公式 2 中使用的表述依赖于这样一个假设：图中相连的边很可能具有相同的标签。由于图的边不一定编码相似性，也可能包含其他信息，因此这一假设可能会限制模型的建模能力。为解决这一问题，研究者提出了图自动编码器（GAEs Kipf 和 Welling （2017年）），通过神经网络对图结构进行建模$f(X, A)$。在我们的设计和实证分析中，将其称为 GeMi GAE。

结合深度神经网络与近似贝叶斯推断的思想，可以使用变分图自动编码器（ VGAE，Kipf 和 Welling （2016）进一步改进图结构的编码框架。VGAE 能够为无向图学习具有可解释性的潜在表示。使用图卷积网络（GCN）编码器和内积解码器即可说明这一点。假设随机潜变量 z_i 由一个 $N \times D$矩阵 Z概括。由双层 GCN 参数化的推断模型定义为

$$q(z|X,A) = \prod_{i=1}^{N} q(z_i|X,A)$$

其中 $q(z_i|X, A) = \mathcal{N}(z_i|\mu_i, diag(\sigma_i^2))$ 此外， $\mu = GCN_\mu(X, A)$ 是 μ_i 和 $log\sigma = GCN_\sigma(X, A)$ 的均值向量矩阵。双层图卷积网络定义如下：

$$GCN(X,A) = \tilde{A} \quad ReLU(\tilde{A}XW_0)W_1$$

其中 W_i 为权重矩阵； GCN_μ 和 GCN_σ 共享第一层参数 ˜ters W_0 和 A 是对称归一化的邻接矩阵。生成模型如下：

$$p(A|Z) = \prod_{i=1}^{N}\prod_{j=1}^{N} p(A_{ij}|z_i, z_j)$$

其中 $p(A_{ij} = 1|z_i z_j) = \sigma(z_i^T z_j)$， $\sigma(\text{。})$ 是逻辑 Sigmoid 函数。最后，学习步骤包括针对变分参数 W_i 优化变分下界 $\mathcal{L}$ ：

$$\mathcal{L} = E_{q(Z|X,A)}[logp(A|Z)] - KL[q(Z|X,A)||p(Z)]$$

其中 $KL[q||p]$是 $q(\text{。})$ 与 $p(\text{。})$ 之间的 Kullback-Leibler 散度。在我们的架构中，该模型称为 GeMi VGAE。

GeMi 中的三种图编码器变体——GCN、GAE 和 VGAE——经过选择，旨在涵盖一系列具有明确原则的归纳偏置。GCN 提供了一种判别式基线，可直接通过图传播有监督的标签信号。GAE 引入了无监督重构目标，使模型能够利用未标注的结构信息，而不必仅仅依赖稀疏的概念标签。VGAE 则在此基础上扩展出概率潜在空间，这尤其适合

我们的研究场景：卷轴画数据集规模较小且含有噪声，而变分形式能够显式地对节点表示中的不确定性进行建模。这三种架构使我们能够研究监督信号强度、结构自监督与潜在不确定性建模之间的权衡——对于这类新数据集而言，这些问题很难预先给出答案。

4.6 处理类别不平衡的策略

通过节点分类学习的概念可能存在不平衡（例如，树）。在基于图的推荐系统中，这种偏斜可能导致：（i）梯度被多数类标签主导；（ii）决策阈值偏向于将稀有标签预测为负类；（iii）图邻域反映的是密集的多数类簇，而非少数类结构（Markwald andDemidova (2024)；Zhao 等人(2021)）。为抵消这些影响，我们从以下三个层面采用平衡策略：

- 图结构
- 损失重加权与困难样本强调
- 优化正则化/稳定化

有关这些方法的详细信息，请参见第5节。

4.7 纳入协同过滤（CF）方法

在推荐任务中，GeMi 将多模态特征提取与用户面板交互信息分离。学习面板表示后，再将用户面板交互纳入下游协同过滤任务。这使协同过滤能够作为即插即用模块融入整体架构。

5 实证结果

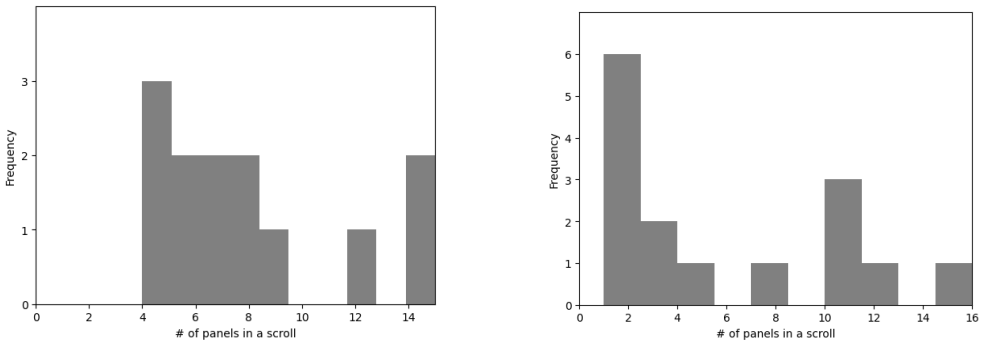
5.1 目标

我们的目标是回答以下研究问题：

•**研究问题1：（基线比较）**与采用同构物品-物品图构建的最先进的基于图的推荐系统，以及协同过滤场景中使用的其他异构用户-物品图相比，我们的系统（GeMi）表现如何？

- **研究问题2：（特征生成）**大语言模型增强的视觉语言特征是否有助于提升推荐系统的模型性能？

- **研究问题3：（图结构学习——归纳式与传导式训练）**归纳式还是传导式训练过程能为推荐提供更好的结果？



(a) 第一阶段　　(b) 第二阶段

图3：田野调查数据收集两个阶段中每幅卷轴面板数量的分布。

5.2 材料

5.2.1 数据

我们针对在两个阶段实地研究中收集并精心整理的面板及相关歌曲数据集开展了一系列实验。第一阶段共有来自13幅卷轴画的106个面板，并与13首不同的歌曲配对。第二阶段收集了15幅不同的卷轴画，共计83个面板。图3a 与图3b展示了这两个阶段所收集卷轴画中每幅卷轴画面板数量的分布。尽管第一阶段的数据对齐良好（每个面板都描绘了某首歌曲的一个片段），但第二阶段的卷轴画中只有26.67%配有相关歌曲。数据集中的每个面板均标注了三个关键概念的有无（标签 =0或1）：树、动物和神话。

5.2.2 机器

所有归纳实验均在由大学管理的高性能计算（HPC）计算节点上，以单GPU配置进行。除非另有说明，每次运行的作业均预留8个CPU核心、64 GB内存和1个GPU。系统支持Jupyter、PyTorch 3和CUDA软件栈，计算则在支持AVX-512的CPU上运行。所有传导实验均使用Google Colab作为执行环境，采用Python 3，并以单个NVIDIA Tesla T4 GPU作为硬件加速器，提供16 GB GPU内存。

5.2.3 基线算法

我们将 GeMi 的性能与最先进的传导式和归纳式基于图的推荐系统进行比较。本节将简要介绍这些算法。

- 传导式同质基线

  1. LATTICE（Zhang 等人（2021b））为每种模态构建 k近邻物品图，并将其细化为跳跃连接潜在邻接矩阵，再通过 Softmax 在不同模态之间进行融合。随后，在融合图上应用 LightGCN 式传播，以增强经过贝叶斯个性化排序优化的协同过滤骨干网络。2. PMGT（Liu 等人（2021b））通过 Transformer编码器在物品多模态图上进行预训练，并利用图结构重建和掩码节点特征重建进行优化。3. HUIGN（Wei 等人（2022））通过从共同交互物品图中学习的层次图卷积，对多层级用户意图进行建模，形成由细到粗的潜在意图层次结构，从而共同细化用户和物品嵌入。4. DualGNN（Wang等人（2023b））在模态特定的用户–物品二部图上应用简化的 GNN传播，以学习每种模态下的用户偏好；随后通过注意力进行融合，并在用户共现图上传播，以处理用户表示中的模态缺失问题。5. FREEDOM（Zhou 和 Shen（2023））通过冻结从原始多模态特征构建的基于 k近邻的物品–物品图，并通过度敏感边剪枝对用户–物品图进行去噪，从而抑制噪声交互，提升多模态推荐效果。6. MICRO（Zhang 等人（2023））从多模态特征中学习模态特定的物品–物品图，通过图卷积细化嵌入，并通过对比融合目标对齐不同模态。7. DGVAE（Zhou 和 Miao（2024））使用解耦图变分自编码器学习可解释的多模态用户偏好，通过在固定的物品–物品图上同时编码用户–物品交互和用户–词语偏好。

  由于上述所有算法都构建同构物品-物品图，它们在理念上与 GeMi 系统最为接近。

- 传导式异质基线

  1. MambaRec（Ren 等人（2025））通过结合多尺度卷积和双重注意力的膨胀细化注意力模块（DREAM）实现局部对齐，并通过最大均值差异（MMD）和对比目标进行全局模态正则化，以确保跨模态语义一致性。2. PGL（Yu 等人（2025））提取主要用户–物品子图（通过全局感知截断或局部感知采样），以保留局部结构信息，并采用基于 InfoNCE 的特征掩码，同时利用多模态物品–物品图。3. HPMRec（Chen 等人（2025a））通过多分量超复数嵌入编码用户–物品表示。该方法利用超复数乘法进行非线性跨模态融合，并通过提示感知补偿动态对齐各分量。4. COHESION（Xu 等人（2025b））是一种复合GCN，整合用户-物品图、用户–用户图和物品-物品图，通过行为模态细化嵌入，

然后谨慎融合各模态的表示，并采用自适应贝叶斯个性化排序目标来平衡学习。5. SMORE（Ong 和 Khong （2025））在频域中融合视觉和文本特征，并通过自适应谱滤波在融合前抑制模态特定噪声。该方法通过多视图图学习和模态感知偏好模块，对单模态偏好和融合偏好进行建模。6. CMDL（Lin 等人（2025b））采用一种新颖的、基于互信息的解耦正则化，将多模态表示解耦为模态不变分量和模态特定组件。

- 归纳式基线：所有归纳式基线均在同质（物品-物品）和异质（用户-物品）设置下进行研究。

  1. GraphSAGE（Hamilton等人 （2017年））：在同质设置下，该算法首先使用物品嵌入的余弦相似度和固定的百分位数阈值构建物品-物品图；然后训练一个两层GraphSAGE编码器，以学习节点特征的聚合函数。在异质设置下，该算法通过邻域聚合构建用户-物品二部图。2. PinSAGE[3]（Cao等人 （2025）；Ying等人（2018））它通过预先计算共同访问次数，并用访问频率最高的前k个邻居重构物品-物品图，以基于随机游走的邻域取代纯粹基于特征的节点聚合邻域。3. GATNE-I（Cen等人 （2019年））在单模态k近邻图上，使用带负采样的Skip-gram目标函数学习物品嵌入。

5.2.4 用户偏好矩阵生成

我们通过蒙特卡洛采样，从训练语料中构建合成用户偏好矩阵，以模拟用户偏好。设训练集由 N_{tr} 个面板组成。每个面板 i 都标注有一个二元多标签指示向量 $\mathbf{y}_i$，其形式为 $\mathbf{y}_i = (y_{i1}, \ldots, y_{i|\mathcal{L}|}) \in \{0,1\}^{|\mathcal{L}|}$，对应固定的语义标签集 $\mathcal{L}$，其中 $|\mathcal{L}| = 3$ 对应于 动物、神话和树类别。

合成用户构造：我们生成一组由 U 个合成用户组成的群体，每个用户都旨在表示标签空间中的一种合理潜在偏好画像。对于每个合成用户 $u \in \{1, \ldots, U\}$，我们从训练物品中采样一个由 K 个不同项目组成的子集，$\mathcal{I}_u \subset \{1, \ldots, N_{\text{tr}}\}, \quad |\mathcal{I}_u| = K,$ $1 \le K \le N_{\text{tr}}$ 交互集合按照训练索引集中所有 K 元子集上的均匀分布独立采样：

$$\mathcal{I}_u \sim \text{Unif}(\{\mathcal{S} \subseteq \{1, \ldots, N_{\text{tr}}\} \mid |\mathcal{S}| = K\})$$

该抽样方案将每个训练集项目视为等可能的候选交互，并通过禁止重复项目来确保用户档案中的多样性。当训练集大小小于 K 时，有效交互数量会隐式限制为 N_{tr}，以确保方案可行。

[3]我们的 PinSAGE 算法实现可从 https://github.com/Haimonti/Folk-Art-Recommendation-System/tree/Inductive Jin/Inductive settings 获取，是对 Cao 等人 (2025) 的改进版本。

偏好估计： 给定抽样交互集合 $\mathcal{I}_u$，我们通过汇总与所选项目相关联的语义标签来估计用户 u 的潜在兴趣。具体而言，我们计算经验标签频率向量 $\mathbf{f}$ $\frac{1}{K}\sum_{i\in\mathcal{I}_u} \; i \; \in \; [0,1]^{|\mathcal{L}|}$ $f_{u\ell}$ $_u \; =$ $\mathbf{y}$，其中每个分量 $\hat{}$ 表示呈现标签 ℓ 的抽样交互所占的比例。为获得离散且易于解释的偏好表示，我们将固定阈值 $\tau \in (0,1)$ 应用于经验频率，从而得到二元偏好向量 $\mathbf{p}_u = \mathbb{I}\left[\hat{\mathbf{f}}_u \geq \tau\right] \in \{0,1\}^{|\mathcal{L}|}$，其中指示函数逐元素应用。等价地，如果某个标签出现在至少占抽样交互 τ 的比例中，则认为该标签受到用户 u 的偏好。等价地，对于每个标签 $\ell \in \{1,\ldots,|\mathcal{L}|\}$，

$$p_{u\ell} = \begin{cases} 1, & \text{if } \frac{1}{K}\sum_{i\in\mathcal{I}_u} y_{i\ell} \geq \tau, \\ 0, & \text{otherwise} \end{cases}$$

这种表述能够自然适应交互预算 K，确保偏好激活反映的是重复证据，而非孤立出现。**用户偏好矩阵：** 汇总所有合成用户的偏好向量后，可得到用户偏好矩阵 $\mathbf{P} = \begin{bmatrix} \mathbf{p}_1^\top \\ \mathbf{p}_2^\top \\ \vdots \\ \mathbf{p}_U^\top \end{bmatrix} \in \{0,1\}^{U\times|\mathcal{L}|}$ $\mathbf{p}_1 \ldots \mathbf{p}_2$。

P的每一行都编码了一个不同的合成用户在语义标签空间中的推断兴趣，每一列则对应一个特定的标签类别。所有档案中的训练项目出现总次数为 UK；假设采样大致均匀，则每个训练项目平均约出现在 $\frac{UK}{N_{\text{tr}}}$ 个合成用户档案中，从而提高训练分布的覆盖率。

—

5.3 方法

以下细节值得关注：

• **GeMi 变体的参数：** 表1 总结了传导式协议下三个 GeMi 变体（即 GeMi-GCN、GeMi-GAE 和 GeMi-VGAE）的架构与优化配置。在所有情况下，图的构建均基于余弦相似度 k 近邻连接，监督学习采用带 logits 的焦点二元交叉熵，使用 Adam 进行优化，正则化则包括丢弃率、边丢弃和梯度裁剪。评估采用固定的合成用户档案，包含 $k = 5$ 次交互，偏好阈值为 0.2。

为确保不同变体之间具有可比性，我们有意统一了若干核心超参数。这些参数包括丢弃率（0.20）、优化器（Adam）、最大范数为 2.0 的梯度裁剪、基于 kNN 的图构建（各变体对应的 k值分别为 30、30 和 25）、边丢弃（0.10）、图层和头部层中的 ReLU 激活，以及焦点损失参数 $\gamma = 2.0$ 和 $\alpha = 0.25$。此外，所有模型均采用相同的推荐评估设置（$k = 5$ 交互，偏好阈值为 0.2）。

GeMi 变体	参数
GeMi-GCN	图编码器：2 层图卷积网络；GCN 隐藏维度：128；GCN 丢弃率：0.20；训练轮数：450；优化器：Adam；激活函数：ReLU；学习率：3×10^{-4}；权重衰减：2×10^{-3}；梯度裁剪：最大范数 2.0；损失函数（平衡）：焦点二元带 logits 的交叉熵；焦点参数：$\gamma = 2.0$, $\alpha = 0.25$；k近邻邻居（k）：30；树边增强邻居（k_{tree}）：25；树增强最大节点数：2500；边丢弃：0.10；$k = 5$ 交互；偏好阈值：0.2
GeMi-GAE	图编码器：采用两层图卷积网络（GCN）的图自编码器（GAE）编码器；隐藏维度：128；潜在维度（d_z）：64；丢弃率：0.20；训练轮数：400；优化器：Adam；激活函数：ReLU；学习率——率：2×10^{-3}；权重衰减：2×10^{-4}；梯度裁剪：最大范数 2.0；k近邻（k）：30；边丢弃：0.10；自监督损失函数：图自编码器重构损失 $\mathcal{L}_{rec}$；监督损失：焦点二元交叉熵——带 logits 的熵；焦点参数：$\gamma = 2.0$、$\alpha = 0.25$；联合目标函数：$\mathcal{L} = \mathcal{L}_{rec} + \lambda_{sup}\mathcal{L}_{sup}$ 带有 $\lambda_{sup} = 0.60$；增强邻居数：$k_{animal} = 18$、$k_{myth} = 18$、$k_{tree} = 35$；每个标签的增强最大节点数：2500；$k = 5$ 交互；偏好阈值：0.2
GeMi-VGAE	图编码器：基于 GCN 的变分图自编码器（VGAE）参数化；隐藏维度：256；潜在维度（d_z）：128；丢弃率：0.20；对数标准差截断：$\log \sigma \in [-10, 10]$；训练轮数：500；优化器：Adam；激活函数：ReLU；学习率：3×10^{-3}；权重衰减：5×10^{-4}；梯度裁剪：最大范数 2.0；k近邻（k）：25；树增强邻居数：$k_{tree} = 25$；树增强最大节点数：2500；边丢弃：0.10；重构损失函数：变分图自编码器重构损失 $\mathcal{L}_{rec}$；KL 正则化：变分图自编码器KL损失 $\mathcal{L}_{KL}$；KL退火：$\beta_{KL}(t)$ 线性增加至 $\beta_{\max} = 1.0$；自监督变分图自编码器目标：$\mathcal{L}_{vgae} = \mathcal{L}_{rec} + \beta_{KL}\mathcal{L}_{KL}$；监督损失：焦点二元带 logits 的交叉熵；焦点参数：$\gamma = 2.0$, $\alpha = 0.25$；联合自监督学习目标：$\mathcal{L} = \mathcal{L}_{vgae} + \lambda_{ssl}\mathcal{L}_{su\ \ p}$ 使用 $\lambda_{ssl} = 0.60$；$k = 5$ 交互；偏好阈值：0.2

表 1：GeMi 变体的参数

• **使用对称 K 近邻进行平衡图结构学习：**我们不使用全局相似度阈值构建边（这种方式往往会使密集多数区域过度连接，而稀有区域连接不足），而是在所有节点上构建对称 k 近邻（*kNN*）图。设$\mathbf{X} \in \mathbb{R}^{N\times d}$ 为节点特征。首先进行 ℓ_2归一化：$\tilde{\mathbf{x}}_i = \mathbf{x}_i / (\|\mathbf{x}_i\|_2 + \varepsilon)$。我们计算余弦相似度 $s_{ij} = \cos(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)$，并为每个节点 i选取其前 k 个近邻：$\mathcal{N}_k(i) = \text{argtopk}_{j\neq i}\ s_{ij}$。然后添加双向边（对称闭包）：$A_{ij} = \mathbb{I}[j \in \mathcal{N}_k(i)] \vee \mathbb{I}[i \in \mathcal{N}_k(j)]$。也可以应用相似度下限 $s_{ij} \leftarrow \max(s_{ij}, 0)$，以避免出现负边。对称 k 近邻可确保局部邻域大小大致均匀，通过确保每个节点参与的图传播量相近，改善稀疏概念的表征学习。

- **类别不平衡处理策略：** 我们采用了多种不同的策略来处理类别不平衡问题，包括：

1. 面向少数类的边增强，用于树 标签：为明确应对树 类别的极端稀疏性，我们在树正类训练节点之间添加额外的边。令 $\mathcal{V}_{\text{tree}} = \{i \mid y_{i,\text{tree}} = 1\}$ 表示被标记为树正类的节点索引（实际操作中限定为训练节点）。我们计算 $\mathcal{V}_{\text{tree}}$ 内部的余弦相似度，并将每个节点与其排名前 k_{tree} 的邻居相连：$\mathcal{N}_{k_{\text{tree}}}^{\text{tree}}(i) = \text{arg topk}_{j \in \mathcal{V}_{\text{tree}}, j \neq i} \cos(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)$。随后，我们将这些边与基础 k近邻图合并：$\mathcal{E} \leftarrow \mathcal{E} \cup \mathcal{E}_{\text{tree}}$。为确保计算可行并防止连接过度，我们限制用于增强的树节点数量（必要时进行随机子采样）。添加少数类内部边可以增加类内信息流，稳定少数类正样本的嵌入，并降低这些节点被多数类邻域主导的信息消息淹没的可能性。

2. 边丢弃以减少多数连通性过拟合：在每个训练轮次中，我们应用边丢弃，以概率p_e随机移除边。对于边集 $\mathcal{E}$，我们采样伯努利掩码 $\delta_e \sim \text{Bernoulli}(1-p_e)$ 并保留：$\mathcal{E}' = \{e \in \mathcal{E} \mid \delta_e = 1\}$。边丢弃作为结构正则化，增强对缺失边的鲁棒性，并防止依赖一小部分主导连接，从而有利于少数类泛化。

3. 通过正类权重进行类别平衡重加权：我们根据训练 标签计算各标签的正类权重，以抵消出现率的偏斜。对于标签 ℓ，令$P_\ell = \sum_{i \in \text{训练}} y_{i\ell}$ 为正样本，$N_\ell = n_{\text{train}} - P_\ell$ 为负样本。我们设定：$w_\ell = \frac{N_\ell}{P_\ell}$（并进行截断以避免除以零），并使用加权二元交叉熵（带 logits）：

$$\mathcal{L}_{\text{WBCE}} = \frac{1}{n_{\text{train}}} \sum_{i \in \text{train}} \sum_{\ell} \Big(- w_\ell \, y_{i\ell} \log \sigma(z_{i\ell}) - (1 - y_{i\ell}) \log(1 - \sigma(z_{i\ell})) \Big),$$

其中 $z_{i\ell}$ 为 logits，$\sigma(\cdot)$ 为 Sigmoid。若不进行重加权，经验风险会被大量负样本和频繁标签主导，使优化器倾向于保守的预测结果（尤其是对稀有标签）。w_ℓ 增大少数类正样本的梯度贡献，从而有效均衡误分类稀有正样本与频繁正样本的代价。

4. 采用焦点损失强调难例和少数类样本：我们在带 logits 的二元交叉熵之上进一步加入焦点调制。令 $p_{i\ell} = \sigma(z_{i\ell})$。定义

$$p_t = \begin{cases} p_{i\ell}, & y_{i\ell} = 1, \\ 1 - p_{i\ell}, & y_{i\ell} = 0. \end{cases}$$

我们采用聚焦因子 $(1 - p_t)^\gamma$ 和一个 α平衡项：$\alpha_t = \alpha \, y_{i\ell} + (1 - \alpha)(1 - y_{i\ell})$，得到焦点二元交叉熵目标：

$$\mathcal{L}_{\text{focal}} = \frac{1}{n_{\text{train}}} \sum_{i \in \text{train}} \sum_{\ell} \alpha_t \, (1 - p_t)^\gamma \, \text{BCEWithLogits}(z_{i\ell}, y_{i\ell}; w_\ell).$$

焦点损失会降低分类正确的样本的权重，并将学习重点放在困难样本上——通常是边界样本和少数类正样本——从而提高少数类召回率，同时避免假阳性过度增加。

5. 重加权目标下的训练稳定化：重加权（正类权重）和焦点调制可能会增大梯度方差。因此，我们采用以下稳定化措施：

（a）梯度裁剪：反向传播后，我们裁剪全局范数

$$\|\nabla_\theta \mathcal{L}\|_2 \leftarrow \min(\|\nabla_\theta \mathcal{L}\|_2,\ \tau), \tag{3}$$

以防止稀有标签梯度尖峰导致更新不稳定。

（b）正则化与容量调优：我们在图卷积网络中使用更大的隐藏维度和丢弃率，并采用权重衰减，以提升类别不平衡情况下的泛化能力。（c）边丢弃（如上所述）作为结构正则化项，通过防止拓扑过拟合来补充损失平衡。

平衡策略有意放大少数类梯度；裁剪和正则化可确保这些放大的梯度不会导致更新爆炸、振荡或过拟合，尤其是在多标签场景中，因为某些样本会对多个损失函数产生贡献。

- **传导式与归纳式图学习：**我们在传导式和归纳式设置下研究 GeMi 系统——针对每种设置，均采用最先进的同质图基线算法进行比较。此外，在传导式设置下，我们还研究基于异构图构建的协同过滤风格算法。基线算法已在第5.2.3节中讨论，表A1 和表A2 阐明了运行这些算法时使用的参数。

- **相关性估计与推荐：** 我们将相关性估计形式化为一个标签条件语义一致性问题，即推断的用户偏好与推荐项目之间的一致性问题。令 $\mathcal{D}_{\text{tr}}$ 和 $\mathcal{D}_{\text{te}}$ 分别表示训练集和测试集，其基数为 $|\mathcal{D}_{\text{tr}}| = N_{\text{tr}}$ 和 $|\mathcal{D}_{\text{te}}| = N_{\text{te}}$。每个项目（面板）$i \in \mathcal{D}_{\text{tr}} \cup \mathcal{D}_{\text{te}}$ 都关联有一个二元多标签标注

$$\mathbf{y}_i = (y_{i1}, \ldots, y_{i|\mathcal{L}|}) \in \{0,1\}^{|\mathcal{L}|}, \tag{4}$$

其中 $\mathcal{L}$ 表示固定的语义标签集（在我们的设定中，$|\mathcal{L}| = 3$）。标签指示值被视为项目内容的真实语义描述符。我们考虑一组由 $m \in \mathcal{M}$索引的嵌入模型。对于每个模型m，训练集和测试集中的物品均嵌入到一个 d_m维潜在空间中，$\mathbf{e}_i^{(m)} \in \mathbb{R}^{d_m}, \quad i \in \mathcal{D}_{\text{tr}} \cup \mathcal{D}_{\text{te}}$

1. 合成用户偏好模型：令 $u \in \{1, \ldots, U\}$ 表示一个用户。每个用户都关联一个推断得到的二元偏好向量 $\mathbf{p}_u = (p_{u1}, \ldots, p_{u|\mathcal{L}|}) \in \{0,1\}^{|\mathcal{L}|}$，其中 $p_{u\ell} = 1$ 表示用户 u 被推断为偏好语义标签$\ell \in \mathcal{L}$。向量 $\mathbf{p}_u$ 的构建遵循第5.2.4节所述的蒙特卡洛采样和阈值化过程，并在余下的评估中视为固定。

2. 潜在空间中的用户表示：对于合成用户 u，我们通过聚合用于生成用户档案的抽样物品的嵌入，将其关联到一个潜在用户表示 $\mathbf{e}_u^{(m)} \in \mathbb{R}^{d_m}$ 。该表示被解释为模型 m下用户潜在语义兴趣的点估计。

3. 评分与排序函数：对于给定的模型 m 和合成用户 u，为每个测试项目 $t \in \mathcal{D}_{\text{te}}$ 分配相关性分数 $s^{(m)}(u,\ t) = \text{sim}\left(\mathbf{e}_u^{(m)},\ \mathbf{e}_t^{(m)}\right)$，其中 $\text{sim}(\cdot,\ \cdot)$ 表示余弦相似度，$\text{sim}(\mathbf{x},\ \mathbf{z}) = \frac{\mathbf{x}^\top \mathbf{z}}{\|\mathbf{x}\|_2\,\|\mathbf{z}\|_2}$ 大小为 K_{rec} 的推荐列表定义为 $\mathcal{R}_u^{(m)} = \arg\max_{\substack{\mathcal{S}\subset\mathcal{D}_{\text{te}} \\ |\mathcal{S}|=K_{\text{rec}}}} \sum_{t\in\mathcal{S}} s^{(m)}(u,\ t)$，即取得最高-$K_{\text{rec}}$ 个相似度分数的测试项目集合。

4. 标签条件相关性定义 相关性针对每个语义标签独立评估。对于用户 u、项目 t 和标签 ℓ，我们定义按标签划分的相关性指标

$$r_{u,t,\ell} = \mathbb{I}[p_{u\ell} = 1 \ \wedge \ y_{t\ell} = 1]$$

因此，当且仅当满足以下条件时，推荐项目才被视为与标签 ℓ 相关：(i) 推断用户偏好标签 ℓ；(ii) 项目在其真实标注中体现了标签 ℓ。等价地，对于每个用户 u和标签 ℓ，我们定义按标签条件划分的候选集

$$\mathcal{C}_{u,\ell} = \{t \in \mathcal{D}_{\text{te}} \mid p_{u\ell} = 1 \ \wedge \ y_{t\ell} = 1\},$$

并将相关推荐集合表示为

$$\mathcal{RR}_{u,\ell}^{(m)} = \mathcal{R}_u^{(m)} \cap \mathcal{C}_{u,\ell}$$

5. 评估指标（Precision@K）：对于每个合成用户 u、标签 ℓ 和模型 m，我们计算特定标签的 Precision@K_{rec}

$$\text{P@}K_{\text{rec}}^{(m)}(u,\ell) = \frac{|\mathcal{RR}_{u,\ell}^{(m)}|}{K_{\text{rec}}} = \frac{1}{K_{\text{rec}}} \sum_{t\in\mathcal{R}_u^{(m)}} r_{u,t,\ell} \tag{5}$$

由此可为每位用户得到一个按标签划分的精确率分数向量

$$\boldsymbol{\pi}_u^{(m)} = \left(\text{P@}K_{\text{rec}}^{(m)}(u,1), \ldots, \text{P@}K_{\text{rec}}^{(m)}(u,|\mathcal{L}|)\right) \in \mathbb{R}^{|\mathcal{L}|} \tag{6}$$

在合成用户群体上进行汇总后，我们报告经验均值和标准差

$$\mu_\ell^{(m)} = \frac{1}{U}\sum_{u=1}^{U} \text{P@}K_{\text{rec}}^{(m)}(u,\ell), \qquad \sigma_\ell^{(m)} = \sqrt{\frac{1}{U}\sum_{u=1}^{U}\left(\text{P@}K_{\text{rec}}^{(m)}(u,\ell) - \mu_\ell^{(m)}\right)^2} \tag{7}$$

我们将 Precision@K 作为主要指标，因为 GeMi 中的推荐任务被明确界定为面向精确率的检索问题：给定用户已知的

概念偏好，系统会展示一份包含 K 个面板的简短列表，这些面板很可能符合这些偏好。在这种情境下，返回少量高度相关的物品，比穷尽式检索更有价值——浏览艺术收藏的用户，从精心筛选的短列表中获益更多，而不是在数百个面板中追求高召回率。不过，我们也承认，仅凭 Precision@K 无法体现排序质量；未来若纳入归一化折损累计增益或平均准确率均值进行评估，将能提供更全面的评估，尤其是随着数据集不断增大。

5.4 结果

本节将对传导式和归纳式设置进行实证分析，并回答第 5.1 节提出的研究问题。

5.4.1 性能比较（研究问题1）

传导实验

表 2–5 展示了 GEMI 在传导式设置下，与最先进的基于图的基线方法在同质物品–物品图和异质用户-物品图上的比较结果。总体而言， GEMI 在不同概念上表现出最强且最稳定的性能，尤其是在神话和树上，同时在动物上也具有竞争力。结果表明，将大语言模型增强的多模态特征与基于图的表征学习相结合，相较于以往的同质和异质图推荐方法，能够带来显著提升。

在采用 Llama-SigCLIP 特征的同质物品–物品图上，LATTICE、PMGT、HUIGN、DualGNN、FREEDOM、MICRO 和 DGVAE 等基线方法表现中等，其中 MICRO 在动物上的得分为 0.61，HUIGN 在神话上的得分为 0.55。然而， GEMI-GCN 显著提升了树上的表现，达到 0.62，神话上的表现也达到 0.61，这表明在标签信号稀疏时，模型具有很强的关系传播能力。 GEMI-GAE则进一步将动物上的表现提升至 0.59，同时在各标签上保持均衡的结果。尤为突出的是， GEMI-VGAE 在神话上的得分达到 0.69，是此配置下的最高值，表明变分图编码能够更好地捕捉物品关系中的不确定性和潜在结构。因此，即使面对强大的同质图基线，三个 GEMI 变体也都至少能在一个标签维度上达到或超过表现最佳的竞争方法。

在采用 Llama-VAE 特征的同质图上，性能提升更为显著。虽然基线在大多数维度上的得分仍低于 0.60，但 GEMI-GCN 在动物维度上达到 0.60，在神话维度上达到 0.65。 GEMI-GAE 在神话维度上的整体表现最佳，达到 0.78，显著优于此前的方法。 GEMI-VGAE 在动物维度上达到 0.62，再次优于基线。这些结果表明，使用更丰富的生成式文本表示时， GEMI 中的图编码器变体始终优于同质图基线。

异质用户-物品图也呈现出类似趋势。在 Llama-SigCLIP 配置下，MambaRec、PGL、HPM-Rec、COHESION、SMORE 和 CMDL 等具有竞争力的多模态协同过滤基线方法在神话数据集上的数值约为 0.60–0.67，但在树数据集上表现较弱。 GEMI-GCN 将树数据集上的数值提升至 0.36，同时在神话数据集上保持 0.65。 GEMI-GAE 在神话数据集上达到 0.78，是该配置下最优的结果。

模型	动物	神话	Tree
LATTICE	0.57 ± 0.21	0.47 ± 0.25	0.20 ± 0.18
PMGT	0.52 ± 0.21	0.46 ± 0.23	0.33 ± 0.22
HUIGN	0.48 ± 0.16	0.55 ± 0.25	0.10 ± 0.11
DualGNN	0.55 ± 0.21	0.47 ± 0.26	0.25 ± 0.21
FREEDOM	0.58 ± 0.22	0.41 ± 0.24	0.22 ± 0.19
MICRO	0.61 ± 0.20	0.47 ± 0.26	0.26 ± 0.21
DGVAE	0.55 ± 0.21	0.45 ± 0.26	0.24 ± 0.22
GeMi GCN	**0.47 ± 0.27**	**0.61 ± 0.16**	**0.62 ± 0.30**
GeMi GAE	**0.59 ± 0.30**	**0.55 ± 0.13**	**0.49 ± 0.29**
GeMi VGAE	**0.51 ± 0.22**	**0.69 ± 0.16**	**0.48 ± 0.21**

表 2：**传导式**Llama-SigCLIP 特征，同质图

模型	动物	神话	Tree
LATTICE	0.53 ± 0.22	0.54 ± 0.13	0.21 ± 0.25
PMGT	0.47 ± 0.19	0.52 ± 0.17	0.25 ± 0.20
HUIGN	0.49 ± 0.20	0.51 ± 0.12	0.23 ± 0.22
DualGNN	0.41 ± 0.18	0.56 ± 0.32	0.19 ± 0.27
FREEDOM	0.51 ± 0.17	0.55 ± 0.24	0.24 ± 0.16
MICRO	0.45 ± 0.25	0.58 ± 0.23	0.26 ± 0.19
DGVAE	0.44 ± 0.23	0.60 ± 0.21	0.27 ± 0.21
GeMi GCN	**0.60 ± 0.24**	**0.65 ± 0.20**	**0.36 ± 0.26**
GeMi GAE	**0.48 ± 0.21**	**0.78 ± 0.23**	**0.27 ± 0.15**
GeMi VGAE	**0.62 ± 0.25**	**0.56 ± 0.19**	**0.27 ± 0.28**

表 3：**传导式**Llama-VAE 特征，同质图

在该设置下， GeMi-VGAE在动物数据集上达到 0.62，再次优于异质竞争方法。这表明， GeMi 能够有效整合用户–物品协同结构与多模态特征信号。

在 Llama-VAE 异质设置中，这一模式依然存在。尽管 CMDL 和 HPMRec 在动物和神话上表现不俗，但没有任何基线模型达到 GeMi-GAE 在神话上取得的 0.78 分。同样，相较于以往方法， GeMi-GCN 和GeMi-VGAE在动物和树上均保持了出色表现。在所有异质实验中，Tree 标签尤其受益于 GeMi 中明确的图传播和标签感知增强；而以往的协同过滤模型则因数据稀疏而难以取得良好效果。

综合来看，这些发现回答了研究问题1，表明 GeMi通过整合大语言模型/视觉语言模型衍生的特征与基于图的关系学习，在传导式训练下优于同质物品间图神经网络推荐系统和异质协同过滤基线。

模型	动物	神话	Tree
MambaRec	0.50 ± 0.24	0.60 ± 0.17	0.12 ± 0.13
PGL	0.43 ± 0.26	0.60 ± 0.20	0.25 ± 0.18
HPMRec	0.47 ± 0.26	0.63 ± 0.21	0.34 ± 0.21
COHESION	0.41 ± 0.23	0.59 ± 0.20	0.27 ± 0.17
SMORE	0.40 ± 0.20	0.66 ± 0.24	0.24 ± 0.16
CMDL	0.41 ± 0.25	0.67 ± 0.22	0.23 ± 0.18
GeMi GCN	**0.60 ± 0.24**	**0.65 ± 0.20**	**0.36 ± 0.26**
GeMi GAE	**0.48 ± 0.21**	**0.78 ± 0.23**	**0.27 ± 0.15**
GeMi VGAE	**0.62 ± 0.25**	**0.56 ± 0.19**	**0.27 ± 0.28**

表4：**传导式**Llama-SigCLIP 特征（用户-物品）

模型	动物	神话	Tree
MambaRec	0.52 ± 0.19	0.57 ± 0.26	0.15 ± 0.28
PGL	0.44 ± 0.23	0.60 ± 0.21	0.27 ± 0.21
HPMRec	0.46 ± 0.18	0.64 ± 0.22	0.23 ± 0.14
COHESION	0.50 ± 0.24	0.55 ± 0.11	0.25 ± 0.26
SMORE	0.41 ± 0.54	0.59 ± 0.13	0.17 ± 0.31
CMDL	0.57 ± 0.28	0.61 ± 0.15	0.16 ± 0.25
GeMi GCN	**0.60 ± 0.24**	**0.65 ± 0.20**	**0.36 ± 0.26**
GeMi GAE	**0.48 ± 0.21**	**0.78 ± 0.23**	**0.27 ± 0.15**
GeMi VGAE	**0.62 ± 0.25**	**0.56 ± 0.19**	**0.27 ± 0.28**

表5：**传导式**Llama-VAE特征异构图

模型	动物	神话	Tree
GraphSAGE	0.48 ± 0.14	0.56 ± 0.23	0.34 ± 0.28
PinSAGE	0.52 ± 0.21	0.57 ± 0.22	0.49 ± 0.14
GATNE-I	0.52 ± 0.20	0.60 ± 0.21	0.37 ± 0.19
GeMi GCN	**0.56 ± 0.24**	**0.66 ± 0.20**	**0.44 ± 0.20**
GeMi GAE	**0.50 ± 0.22**	**0.63 ± 0.15**	**0.34 ± 0.23**
GeMi VGAE	**0.52 ± 0.22**	**0.70 ± 0.21**	**0.44 ± 0.20**

表 6：**归纳式**Llama-SigCLIP 特征（物品-物品）

归纳实验

表6–7总结了同质物品-物品图下的归纳性能。总体而言，表现最佳的 GeMi 变体取决于标签。

- 动物：最强的归纳式项目—项目结果由采用 LLM-SigCLIP 特征的 GeMi-GCN 取得（0.56），与采用 LLM-VAE 的最佳归纳竞争方法表现相近。

模型	动物	神话	Tree
GraphSAGE	0.56 ± 0.25	0.58 ± 0.12	0.33 ± 0.16
PinSAGE	0.56 ± 0.25	0.57 ± 0.12	0.32 ± 0.17
GATNE-I	0.56 ± 0.25	0.60 ± 0.19	0.34 ± 0.22
GeMi GCN	**0.46 ± 0.22**	**0.32 ± 0.19**	**0.56 ± 0.34**
GeMi GAE	**0.51 ± 0.29**	**0.61 ± 0.18**	**0.29 ± 0.23**
GeMi VGAE	**0.50 ± 0.27**	**0.61 ± 0.23**	**0.29 ± 0.23**

表7：**归纳式**Llama-VAE 特征（物品-物品）

模型	动物	神话	Tree
GraphSAGE	0.63 ± 0.18	0.59 ± 0.19	0.26 ± 0.18
PinSAGE	0.50 ± 0.21	0.65 ± 0.17	0.21 ± 0.15
GATNE-I	0.51 ± 0.20	0.60 ± 0.21	0.36 ± 0.19
GeMi GCN	**0.71 ± 0.25**	**0.39 ± 0.05**	**0.35 ± 0.13**
GeMi GAE	**0.47 ± 0.20**	**0.71 ± 0.20**	**0.20 ± 0.22**
GeMi VGAE	**0.52 ± 0.23**	**0.77 ± 0.11**	**0.53 ± 0.19**

表 8：**归纳式** Llama-SigCLIP 特征异构图

模型	动物	神话	Tree
GraphSAGE	0.58 ± 0.24	0.52 ± 0.15	0.26 ± 0.12
PinSAGE	0.46 ± 0.20	0.61 ± 0.11	0.39 ± 0.22
GATNE-I	0.56 ± 0.24	0.61 ± 0.20	0.33 ± 0.22
GeMi GCN	**0.36 ± 0.12**	**0.59 ± 0.07**	**0.35 ± 0.13**
GeMi GAE	**0.69 ± 0.25**	**0.77 ± 0.11**	**0.34 ± 0.15**
GeMi VGAE	**0.53 ± 0.19**	**0.78 ± 0.10**	**0.53 ± 0.19**

Tab第 9 列：**归纳式**Llama-VAE 特征异构

一致的

• 神话：最佳归纳式结果由采用 LLM-SigCLIP 特征的 GeMi-VGAE 取得（0.70）。表现最强的竞争基线是 GATNE-I（0.60），而 GraphSAGE 和 PinSAGE 的表现更低。

• 树：总体而言，树是最难预测的标签，但 GeMi 在归纳式物品间性能上仍表现最佳。在 LLM+SigCLIP 下，GeMi-GCN 达到 0.56 ± 0.34，明显优于 GraphSAGE/PinSAGE/GATNE-I。一种可能的解释是，“树”通常是许多面板中的次要元素或背景元素：它所占区域很小，视觉显著性低于主体，而且在配套文本中可能鲜少提及（甚至完全没有提及）。因此，与神话人物或动物等语义上更突出的类别相比，多模态特征和诱导邻域结构为该标签提供的判别信号都有限。

在用户–物品归纳设置（表 8–9）中，GeMi 各变体仍具竞争力，并且有时会在特定标签上进一步提升（例如，动物：采用 LLM-SigCLIP 的 GeMi-GCN（0.71）；神话：采用 LLM-VAE 的 GeMi-VGAE（0.78）；树：采用 LLM-VAE 的 GeMi-VGAE（0.53））。这是意料之中的，因为用户–物品二部图提供了纯基于特征的物品邻域之外的额外交互结构。

5.4.2 特征生成（研究问题2）

传导实验我们首先在相同的 GeMi 骨干网络（图卷积网络、图自编码器、变分图自编码器）中，将基础特征表示（SigCLIP+ 图和变分自编码器 + 图，见表 10 和 11）与其经大语言模型增强的对应版本进行比较。随后，我们分析大语言模型 + SigCLIP 与大语言模型 +变分自编码器在不同标签维度下的相对表现。

GeMi 中大语言模型增强特征与基础特征的比较比较表格 10和11与表格 2–5 的结果可见，使用大语言模型生成的特征能够带来一致且显著的提升。例如，在同质物品-物品图下，基础SigCLIP + 图卷积网络在动物和神话上的得分分别为 0.48 和 0.70，而大语言模型-SigCLIP + 图卷积网络将动物得分提升至 0.60，同时在神话上仍保持优异表现。同样，基础VAE + 图卷积网络在动物和树上的得分分别为 0.49 和 0.38，而大语言模型-VAE + 图卷积网络将动物得分提升至 0.60，并在树上保持了有竞争力的表现，得分为 0.36。

对于自编码器变体，这种提升更加显著。基础VAE +图自编码器在动物和神话上的表现分别仅为0.33和0.47，而LLM-VAE +图自编码器则分别提升至0.48和0.78。在神话数据集上的这一巨大差距表明，经过大语言模型增强的表示能够更好地编码有利于基于重构的图目标的语义结构。同样，LLM-SigCLIP +变分图自编码器将动物上的表现从0.51（基础）提升至0.62，并在其他标签上保持了稳定的性能。

在异质用户-物品图中，也呈现出相同的规律。尽管MambaRec、HPMRec、SMORE和CMDL等具有竞争力的基线在神话上的表现仍处于0.57–0.67范围内，但经过大语言模型增强的GeMi变体最高可达到0.78。这些结果证实，特征质量，而不仅仅是图架构，是影响性能的关键因素。

大语言模型 + SigCLIP 与大语言模型 + 变分自编码器 对两种大语言模型增强的特征流水线进行比较，可以发现它们在不同标签上的表现具有互补性。在动物类别上，LLM-SigCLIP +变分图自编码器达到 0.62，略高于 LLM-VAE + 图卷积网络的 0.60，这表明判别式对比对齐有利于视觉基础类别。在神话类别上，LLM-VAE + 图自编码器以 0.78 的成绩取得最佳结果，超过了 LLM-SigCLIP 变体，说明生成式多模态对齐能够更有效地捕捉抽象的叙事语义。对于树类别，LLM-SigCLIP + 图卷积网络达到 0.36，而在异质设置下，LLM-VAE + 图卷积网络也取得相近的 0.36，表明两种流水线与图传播结合时，都能够捕捉结构类别线索。

总体而言，结果表明，在所有 GeMi 骨干网络中，大语言模型增强的视觉语言表征始终优于相应的基础表征。对于神话等语义丰富的标签，这种提升尤为显著，因为大语言模型的条件化提供了更清晰的上下文信号。

模型	动物	神话	Tree
SigCLIP + 图卷积网络	0.48 ± 0.25	0.70 ± 0.21	0.13 ± 0.15
SigCLIP + 图自编码器	0.54 ± 0.28	0.52 ± 0.17	0.20 ± 0.17
SigCLIP + 变分图自编码器	0.51 ± 0.25	0.62 ± 0.19	0.18 ± 0.18

表 10：**传导式**基础SigCLIP

模型	动物	神话	Tree
变分自编码器 + 图卷积网络	0.49 ± 0.23	0.58 ± 0.20	0.38 ± 0.24
变分自编码器 + 图自编码器	0.33 ± 0.20	0.47 ± 0.21	0.23 ± 0.20
变分自编码器 + 变分图自编码器	0.34 ± 0.20	0.53 ± 0.13	0.20 ± 0.18

表 11：**传导式**基础VAE特征

归纳实验 在传导式分析之后，我们考察大语言模型增强特征是否能提升以下模型的归纳性能： GEMI。表 12–13报告了采用基础 SigCLIP 和基础变分自编码器特征的归纳结果，而表 6–9则报告了在相同归纳协议下，采用大语言模型增强管线（ LLM-SigCLIP / LLM-VAE）的归纳结果。总体而言，大语言模型增强特征优于基础特征，提升幅度因标签而异。

• 动物：在 SigCLIP 流程下，LLM 增强将最佳 GEMI得分从 0.52（基础SigCLIP + GCN，表 12）提升至 0.56（LLM-SigCLIP +GCN，表 6），这表明归一化文本输入能够增强用于归纳式邻域传播的视觉–语言嵌入。

• 神话：最佳基础SigCLIP结果为 0.57（表 12），而 LLM-SigCLIP 达到 0.70（表 6）。这表明，LLM 增强的文本消息对神话类别有所帮助，因为其中的语义消歧往往对文本较为敏感。

• 树：改进幅度更为复杂。基础SigCLIP在 0.44 处达到峰值（表 12），而 LLM-SigCLIP 保持在相近范围内（0.44，表 6）；相比之下，VAE 流程显示出更大的跃升，从 0.51（基础VAE +VGAE，表 13）提升至 0.56（LLM-VAE + GCN，表7）。这与树属于视觉上较为细微的类别相一致。

我们进一步比较归纳式设置下的两种大语言模型增强的特征流水线: LLM-SigCLIP（表 6 和表 8）与 LLM-VAE（表 7 和表 9）。总体而言，表现更好的流水线取决于标签，而且同质物品-物品图与异质用户-物品图呈现的模式略有不同。在同质图中，我们观察到以下情况:

• 动物：对于动物，LLM-SigCLIP 的 GEMI 配置表现最强（0.56）；LLM-VAE 的基线表现相近，但未能超过该标签下的最佳 GEMI得分。

• 神话：LLM-SigCLIP 明显更强： GEMI-VGAE 达到 0.70，而 LLM-VAE 的最佳 GEMI 结果为 0.61。

• 树：趋势出现反转：在项目—项目归纳式图中，LLM-VAE 在树上表现更佳， GEMI-GCN 达到 0.56，而表现最佳的 LLM-SigCLIP 约为 0.44

模型	动物	神话	Tree
SigCLIP + 图卷积网络	0.52 ± 0.26	0.57 ± 0.14	0.39 ± 0.27
SigCLIP + 图自编码器	0.50 ± 0.25	0.54 ± 0.23	0.37 ± 0.27
SigCLIP + 变分图自编码器	0.51 ± 0.35	0.50 ± 0.27	0.44 ± 0.21

表12：**归纳式**基础SigCLIP

模型	动物	神话	Tree
变分自编码器 + 图卷积网络	0.47 ± 0.27	0.51 ± 0.28	0.45 ± 0.23
变分自编码器 + 图自编码器	0.39 ± 0.23	0.61 ± 0.22	0.30 ± 0.25
变分自编码器 + 变分图自编码器	0.51 ± 0.25	0.33 ± 0.18	0.51 ± 0.30

表13：**归纳式**基础VAE特征

GEMI 这表明，VAEfusion 或许能在归纳式场景下更好地保留树的背景语义。

另一方面，在异质设置下，两种大语言模型增强管线的表现更为接近。对于动物概念，LLM-SigCLIP 的得分为 0.71，高于 LLM-VAE 的 0.69。对于神话概念，LLM-VAE 略高（0.78），而 LLM-SigCLIP 为 0.77；对于树，两种用户-物品表中的最佳结果基本持平（均为 0.53）。

总体而言，归纳式结果表明，与基础特征相比，大语言模型增强始终能带来益处；而 LLM-SigCLIP 与 LLM-VAE 之间的相对优势则取决于标签和图：在同质图中，SigCLIP 往往更适用于神话，而 VAE 对树等视觉上不明显的标签更有帮助

5.4.3 图结构学习（研究问题3）

我们研究传导式或归纳式图训练哪种方式能带来更好的推荐性能。我们通过比较两种训练范式下 GCN、图自编码器和变分图自编码器在 (i) 同质物品-物品图和 (ii) 异质用户-物品图上的表现来进行分析。

同质物品-物品图（Llama-SigCLIP）对比表2（传导式）以及对应的归纳式表格显示出标签依赖行为。对于动物概念，传导图自编码器（0.59）与归纳图卷积网络（0.56）的表现相近，但传导图自编码器仍略胜一筹。对于神话概念，归纳变分图自编码器（0.70）略微优于传导变分图自编码器（0.69），这表明，对于语义丰富的类别，当图结构稳定时，归纳学习能够很好地泛化。然而，对于树概念，传导图卷积网络（0.62）明显优于归纳图卷积网络（0.44），这说明训练期间在全图上进行稠密传播有利于结构稀疏的标签。

同质物品-物品图（Llama-VAE）在变分自编码器特征下（表3），传导式训练始终表现更强。对于动物概念，直推变分图自编码器（0.62）优于归纳式变体（约 0.50）。对于神话，直推式

图自编码器的得分为 0.78，显著高于归纳式图自编码器/变分图自编码器（约 0.61）。对于树，直推图卷积网络（0.36）和归纳图卷积网络（0.56）各有优势，但归纳式的优势仅体现在图卷积网络上，而图自编码器和变分图自编码器在传导式设置下表现更佳。

异质用户-物品图 对于 Llama-SigCLIP 特征（表4），传导式GeMi 变体始终优于归纳式基线。传导式图卷积网络在 Animal、Mythology 和 Tree 标签上的得分分别为 0.60、0.65 和 0.36，而归纳式 Graph-SAGE、PinSAGE 和 GATNE-I 至少在两个标签上的得分都低于这些数值。Llama-VAE 特征（表 5）也呈现出相同趋势：传导式图自编码器在 Mythology 上取得 0.78，明显高于归纳式对应模型（最高约为 0.61）。

模型层面的趋势（图卷积网络 vs 图自编码器 vs 变分图自编码器）在各种设置下，图卷积网络在 Tree 等结构导向型标签上受益于传导式传播最多。在传导式训练下，图自编码器和变分图自编码器在 Mythology 上表现出明显优势，尤其是使用 LLM增强特征时，基于重构的目标能够利用全图连通性。归纳式训练在个别标签上有时能达到或略微超过传导式训练的表现（例如，在同质 SigCLIP 中，变分图自编码器在 Mythology 上的表现），但这些优势并未在所有标签或图类型中持续出现。

总体而言，传导式训练能带来更可靠且持续优异的性能，尤其是在异质用户-物品图中，以及对于神话这类语义复杂的标签。归纳式训练在某些同质设置中也能取得有竞争力的泛化效果，但并不能持续优于直推式学习。因此，对于本文研究的推荐场景，传导式图结构学习在不同图结构形式和 GeMi 变体中都能取得更强、更稳定的结果。

6 讨论

- 在第 4.1.1 节中，我们介绍了一种基于大语言模型的语义文本规范化流程，并已在 GeMi 系统中使用 TinyLlama（Zhang 等人(2024)）实现。可替代 TinyLlama 的其他大语言模型变体包括 BLOOM（LeScao 和 Fan (2023)）、Gemma2（Team 等人 (2024)）、通义千问（Hui 和 Yang (2024)）以及其他开源变体。使用这些预训练模型开展实验以研究其对性能的影响，是一个活跃的研究领域。
- 近年来的一种趋势是，不再将预训练语言模型适配到下游任务，而是采用预训练、提示与推理 范式，以硬提示/软提示重新表述下游推荐任务（Liu 等人(2023c)）。在这一范式中，可以避免微调，预训练模型本身可直接用于预测小组评分，甚至输出与推荐目标相关的子任务，例如解释。例如，Geng 等人Geng 等人(2023)利用一个预训练CLIP组件将图像转换为词元。这些词元被添加到项目的文本词元中，从而创建个性化多模态软提示。不同的提示策略，例如固定预训练

模型、免调优提示或提示与预训练调优，均可在GeMi系统中加以探索。

- 在当前版本的 GeMi 系统中，我们尚未探索利用扩散模型进行内容生成的可能性（Wei 和 Fang (2025)；Lin 等人 (2025a)）。不过，鉴于多模态数据并非总能良好对齐（例如，可能存在歌曲文本，却没有相应的插图），或许可以运用生成式人工智能技术来生成内容，从而有望增强推荐系统的个性化程度。

- 我们当前的在线系统 UPFAR（见附录E）可用于生成有关系统用户的数据，并了解他们的偏好。我们已将这些真实世界数据纳入 GeMi。

- 研究叙事长卷动态定价机制的算法，可能有助于改善边缘化艺术家的处境，因为他们往往面临严重的经济困境。这将在 GeMi 的未来版本中予以探讨。

7 结论

歌唱画师是一群流动说书人，他们通过长篇叙事长卷画讲述道德故事和神话故事。这种表演艺术如今濒临失传，目前仍从事这一艺术形式的画师仅有数百人。为将他们的艺术作品和歌曲歌词保存给后代，人们正通过广泛的实地调查和数据收集开展保护工作，随后对数据进行仔细标注和处理，进而开发出推荐系统 GeMi。这个多模态基于图的推荐系统的性能可与最先进的传导式与归纳式图学习算法相媲美，并像协同过滤一样纳入用户偏好。GeMi 系统有望进一步融入生成式人工智能能力或基于提示的推荐，也可结合定价模型，使用户能够通过在线平台购买叙事长卷。

致谢。 第一作者感谢以下机构提供的支持：亚洲研究所颁发的Gnamm奖、纽约州立大学布法罗分校国际教育办公室提供的教职人员资助，以及美国国家人文基金会和美国印度研究所提供的奖学金。作者感谢Daricha基金会在新冠疫情期间协助数据收集，感谢Jishnu Basak编辑图像，也感谢研究生课程“机器学习中的优化方法”和“分布式计算与大数据技术”的学生协助处理实地研究收集的数据和标注。最后，本研究的灵感源于与Elise Auerbach、James Nye，以及芝加哥大学数字印度学习倡议的同事们进行的多次讨论。

附录A 基于图神经网络的推荐系统

图神经网络已广泛用于分析图结构数据，以执行节点分类（Kipf 和 Welling (2017年）；Zhu等人(2021））、链接预测（Zhang和Chen(2018））以及信息检索任务（Zhang 等人(2021c））。近年来，图神经网络在推荐系统中的应用（Wu等人(2022b）；Gao等人(2023）；Anand和Maurya(2024）；Sharma等人(2024））已十分普遍。大规模推荐系统通常包含三个阶段：(a) 匹配——从庞大的物品集合（通常有数百万个）中筛选候选物品，从而降低问题的规模和复杂度；(b) 排序——使用排序模型为候选物品评分；(c) 重排序——更新排序列表，以满足其他标准或业务需求。图神经网络[4] 已知可用于上述所有三个阶段。

A.0.1 使用图神经网络进行匹配

在匹配阶段，基于图神经网络的模型常用于构建复杂的架构嵌入表示（Zhang 等人（2021c）；Gao 等（2023）；Wang等人（2019）；He等人（2020）；Wu等人（2021）；Sun 等人（2020a）；Ying等人（2018）；Liu 等人（2021a, 2023b）；Yu 和 Qin（2020）；Mu等人（2022）；Liu 等人（2023a））。这些模型能够捕捉用户与物品之间的高阶相似性，也能捕捉结构连通性。这类模型的一个指导原则是，交互相似的用户往往具有相似的偏好，而这些偏好可以通过多轮信息传播加以揭示。van den Berg等人（2018）提出了一种图自编码器框架，该框架基于在表示电影评分的二部用户-物品交互图上进行可微消息传递。神经图协同过滤（NGCF）（Wang 等（2019））的作者指出，通过预定义的用户-物品特征得到的嵌入无法充分捕捉协同过滤效应；因此，他们利用用户-物品图结构，在图上进行嵌入传播。Sun 等人（2020a）提出了邻居交互感知（NIA）-GCN，该模型能够显式地建模邻居节点之间的关系信息，并利用用户-物品二部图的异质性。Wang 等（2020）认为，用户-物品交互通常以统一的方式建模，无法刻画多样化关系，也无法在表示中解耦用户意图。他们提出的模型解耦图协同过滤（DGCF）通过为每个用户-物品交互建模意图分布，并迭代优化意图感知交互图及其表示，解决了这一问题。边级调制（EGO, Chen 等人（2022））融合通过自适应提取边级多模态信息，并在其他模态的监督下学习调制单模态节点，从而增强用户-物品交互。自适应抗瓶颈多模态图学习网络（Cai 等人（2022a））设计了协同表示学习模块和语义模块，分别用于学习用户和微视频的表示。Du等人（2022）学习了伪相关。

[4]全面综述图神经网络及其应用超出了本文的范围，但我们向有兴趣的读者推荐以下综述：（Wu等人(2022b）；Gao等人(2023）；Anand和Maurya (2024）；Sharma等人(2024））

Liu 等人(2021a)注意到，基于图卷积网络的推荐模型存在过平滑问题；在图卷积操作中，与用户没有共同兴趣的高阶邻居用户也可能参与该用户的嵌入学习。因此，为解决这一问题，研究者提出了一种新颖的兴趣感知消息传递图卷积网络（IMP-GCN）推荐模型，该模型在子图内部执行高阶图卷积。HUIGN（Wei 等人(2022)）以分层图结构呈现用户意图。

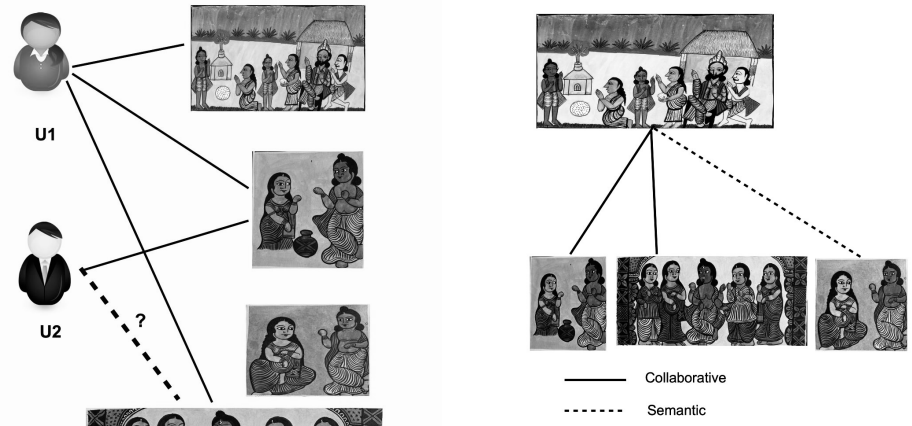
行为驱动特征适配器（BeFA Fan 等人 (2025)）可减少提取特征中的信息漂移和遗漏。多模态推荐的多层次自监督学习（MENTOR，Xu 等人 (2025c)）通过增强特征和图卷积网络，解决标签稀疏和模态对齐问题。EliMRec（Liu 等人 (2022c)）旨在消除单模态偏差。Dual-GNN（Wang等人 (2023a)）解决了缺失模态导致的多模态融合问题。预训练模态解耦表示模型（PAMDHan 等人 (2022)）探讨了用于匹配的模态对齐。在自监督图学习（SGL）（Wu 等人 (2021)）中，会生成节点的多个视图，以最大化同一节点不同视图之间的一致性，并使其高于不同节点视图之间的一致性。这种方法被认为可以纠正高度数节点对推荐产生更大影响以及噪声交互所引发的问题。MMGCL（Yi 等人 (2022)）设计了模态边丢弃和模态掩码，以生成用户和物品的多个视图。

DGMRec（Kim 等人 (2025)）从信息视角解耦模态特征，并通过整合其他模态中对齐的特征、利用用户对模态的偏好来生成缺失的模态特征。模态对齐问题在 Tran 和 Lauw (2022)的研究中得到进一步探讨。COHESION（Xu 等人(2025a)）采用双阶段融合策略，以减少无关信息的影响并完善所有模态。GUIDER（Li 等人(2025a)）对用户反馈进行去噪。MoDiCF（Li 等人 (2025b)）解决了因捕捉模态分布的能力有限而难以准确生成缺失数据的问题。

汉明空间图卷积网络（HS-GCN）（Liu 等人 (2023b)）通过用户-物品二部图构建哈希框架，以在汉明空间中建模一阶和高阶相似度。Yu 和 Qin (2020)指出，在处理超大型图时，为提高计算效率，卷积操作会被过度简化，但这种过度简化会损害其正常功能。因此，他们设计了低通协同过滤器（LCF），以去除观测数据中由量化造成的噪声，同时以不损害图卷积功能的方式降低其复杂度。为应对模态编码器规模庞大以及融合高维特征的复杂性，Zhuang 等人 (2025)提出采用轻量级 Wasserstein 知识蒸馏框架。

虽然图神经网络模型可以高效地将高阶连接注入用户-物品二部图的嵌入中，但它们已知存在若干局限性：

- **建模项目之间的语义关系：**图A1展示了 GeMi 中的一个示例，其中，项目之间的语义关系对于理解潜在图结构的重要性不亚于协同信号。在此示例中，面板 $\{1, 2, 4\}$ 是第一位用户感兴趣的，$\{2\}$ 则只有



(a) 协同物品关系　　(b) 物品之间的语义关系

图 A1：GeMi 中展示两种物品关系的一个简单示例——协同关系和语义关系。两者均可用于学习潜在图结构。

第二位用户点赞过；协同信号有助于判断面板 {4} 是否受到关注。此外，由于四个面板讨论的是同一个角色，因此面板 {1} 与 {3} 之间存在语义关系，而这种关系无法通过该协同信号直接捕获，必须由多模态表示学习得出。据我们所知，只有少数研究能够捕获这类语义信息（例如Meng 等人(2025a））。

• **应对最先进的视觉语言模型与多模态融合：**许多推荐系统会利用用户和物品的多模态信息来建模偏好（Xu 等人 (2025a及其中引用的文献），据我们所知，至少有两个用于实现多模态系统模型的开源代码库/工具箱：MMRec（Zhou 等人 (2023a,c）和 Cornac（Salah 等人 (2020）；Truong 等人 (2021b,a））。例如，LGMRec（Guo 等人 (2024）首先构建局部图嵌入模块，利用局部拓扑关系，分别学习用户和物品的协同嵌入及模态相关嵌入。随后，该方法设计了全局超图嵌入模块，通过建模有意义的全局依赖关系来捕捉用户和物品的全局嵌入。Liu 等人(2021b）提出了一种预训练策略，通过同时考虑物品侧信息及其关系来学习物品表示。多模态图注意力网络（MGAT）（Tao 等人(2020）尝试在模态层面解缠个人兴趣。因此，首先为

每种模态构建多模态交互图；随后使用门控注意力机制，确定不同模态对用户偏好的重要性得分。多模态知识图注意力网络（MKGAT）（Sun 等人(2020b)）采用多模态图注意力技术，在知识图中传播信息。在 PGMT（Liu 等人(2021b)）中，使用预训练的 Inception-v4 网络提取图像特征，并使用预训练的 BERT 模型提取句子特征。LATTICE（Zhang 等人 (2021a)）为每种模态学习物品-物品结构[5] 并聚合多种模态以获得潜在物品图。为探索用户之间高度复杂的关系，研究者使用模态感知异构信息图（M-HIG）生成高质量的用户和微视频嵌入（Cai 等人 (2022b)）。该模型包含两个关键组成部分：（a）一种异构图，利用基于随机游走的采样策略为用户和微视频采样邻居；（b）一个分层的特征内与特征间聚合网络，以更好地捕捉异构图中的复杂结构和丰富语义信息。

除了图神经网络之外，还有一些算法使用图自编码器或变分图自编码器（VGAE）进行特征提取。在 GraphCAR（Xu 等人(2018)）中，使用两个图卷积网络对用户和物品的潜在向量空间进行自编码，并通过计算两个潜在因子向量的内积来获得偏好分数。在 MVGAE（Yi 和 Chen (2022)）中，模态特定变分编码器为每个节点学习一个高斯变量——均值向量表示语义信息，方差向量则表示相应模态的噪声水平。此外，模态特定高斯节点嵌入依据专家乘积原理进行融合，其中每种模态的语义信息根据估计的不确定性水平赋予不同权重。我们探索了在 GeMi 系统中使用 MVGAE 进行排序的方式。在 GeMi 系统中，我们使用一个判别式视觉语言模型（对比语言—图像预训练（CLIP））和一个生成式视觉语言模型（VAE）（详见第 4.1节）。最后，我们也了解到一些基于非图神经网络的技术（如 PAMD），但它们与此处的讨论并无直接关联。

- **利用预训练大语言模型探索嵌入的增强语言属性（如句子级或篇章级属性）及其可控性（如提示技术）：** 大语言模型[6] 已被广泛用作推荐的辅助工具（Wei 等人 (2024); Chen 等人 (2025b); Ji 等人(2025); Meng 等人 (2025b)），例如，利用用户-物品数据、物品元数据，以及其他多模态推荐系统生成的候选物品来设计提示。将大语言模型用作现有推荐系统中的辅助组件，相关研究还较少（Fioretti 等人 (2025)）。GeMi 框架探索了大语言模型的这种辅助作用，以探究预训练模型在句子和篇章层面的语言属性（见第 4.1.1节）。

[5]视觉特征取自已发表文献；文本特征嵌入则通过 sentence-transformers 拼接每个项目的标题、描述、类别和品牌，获得 1024维表示。[6]关于大语言模型如何以及在何处应用于推荐系统，我们推荐读者参阅以下几篇综述（Lin 等人(2024)；Huang 等人(2025)；Wu等人(2024)；Shehmir 和 Kashef(2025)）。这超出了本文的讨论范围。

A.0.2 使用图神经网络进行排序

在推荐的排序阶段，系统会确定一小组候选物品并生成排序分数。基于图神经网络的排序模型（Li 等人（2019）；Su 等人（2020）；Guo 等人（2021）；Li 等人（2021）；Wu等人（2022a）；Liu 等人（2022a））通常由两个分量构成：一个编码器，用于捕捉所需的特征交互；另一个是解码器（预测器），通过整合不同的特征嵌入来估计排序分数。Li 等人（2019）指出，简单、无结构地组合特征字段，往往会限制模型刻画复杂特征交互的能力，因此提出以图结构表示多字段特征，其中每个节点对应一个特征字段，不同字段可通过边进行交互。Zheng 等人（2023）通过提出价格感知的用户偏好建模（PUP）研究价格对推荐模型的影响。该方法首先构建一个由四种不同节点组成的异构图——用户、物品、价格和类别，并在其上进行嵌入传播。最后，借鉴因子分解机（Rendle（2010））的思想，构建一个基于成对交互的解码器。Liu 等人（2022b）设计了属性感知注意力图卷积网络（A^2-GCN），该网络能够处理缺失属性，并允许用户对物品持有不同偏好。Su 等人（2022）提出了一种基于超图神经网络的模型，可生成任意阶的特征交互，从而帮助学习推荐结果。Guo 等人（2021）设计了双图增强嵌入，利用分而治之以及受课程学习启发的有组织学习，来应对特征稀疏性和用户行为导致的稀疏性。Li 等人（2021）开发了序列感知异构图神经协同过滤（SHCF）模型，该模型融合高阶异构协同信号与序列信息，以解决数据稀疏性问题并纳入动态用户偏好。GeMi 框架探索了三种不同的图架构——图卷积网络（GCN）、图自动编码器（GAE）和变分图自动编码器（VGAE），详见第4.5节。

A.0.3 使用图神经网络进行重排序

人们普遍认为，重排序能够提升亚马逊和谷歌等大规模商业平台上的推荐性能（Pei 等人 (2019年)；Liu 等人 (2020)）。生成并排序候选物品后，通常会在排名靠前的候选物品上运行重排序算法，以进一步优化推荐结果。在**物**品**关**系图**神**经网络用于**个**性化**推**荐（IRGPR Liu 等人 (2020)）中，研究者将两个图——项目关系图和用户-项目评分图——融合，以从异质节点表示中获取多样化的语义信息。随后，使用全连接图神经网络对候选物品进行重排序。尽管神经模型可以用于重排序阶段，但基于用户偏好或物品关系开发简单的基于规则的函数也并不少见。在 GeMi 中，我们使用标签一致性规则作为重排序函数。

模型	参数
LATTICE	图卷积层：2；图混合系数（λ）：0.6；损失函数：贝叶斯个性化排序（BPR）；优化器：Adam；学习率：5×10^{-4}；权重衰减：1×10^{-4}；激活函数：线性图传播化；嵌入归一化：ℓ_2 归一化；
PMGT	每个序列的邻居数：32；隐藏层大小：128；Transformer层：4；注意力头：4；前馈维度：256；激活函数：GELU；丢弃率：0.1；损失函数：带 logits 的二元交叉熵 $+$ 均方误差；掩码比例：0.16；随机替换比例：0.02；优化器：AdamW；学习率：8×10^{-4}；权重衰减：1×10^{-4}
HUIGN	层次化意图层（$K^{(1)}$，$K^{(2)}$，$K^{(3)}$）：$[32, 8, 4]$；协同交互阈值：≥ 2；损失项：贝叶斯个性化排序 $+$ 熵正则化 $+$ 独立性正则化 $+ \ell_2$ 权重衰减；熵权重（λ_1）：0.1；独立性权重（λ_2）：0.1；权重衰减（λ_3）：1×10^{-5}；优化器：Adam；学习率：1×10^{-4}
DualGNN	图传播深度：2；图混合系数（α）：0.6；对比学习温度（τ）：0.2；节点特征丢弃：0.2；损失函数项：InfoNCE 对齐 $+$ 正交性正则化 $+$ 图平滑正则化；Loss 权重：λ_{align}=1.0, λ_{ortho}=0.05, λ_{smooth}=0.05；优化器：Adam；学习率：8×10^{-4}；权重衰减：1×10^{-4}
FREEDOM	多模态图混合：α_v=0.1（图像）和 α_t=0.9（文本）；项目-项目传播层（L_{ii}）：1；用户-物品传播层（L_{ui}）：2；度敏感边剪枝比率（ρ）：0.8；训练损失：贝叶斯个性化排序 $+$ 多模态辅助BPR正则项；正则化权重（λ）：1×10^{-3}；优化器：Adam；学习率：1×10^{-3}；权重衰减：1×10^{-5}
MICRO	物品视图层：1（每种模态）；图混合系数（α）：0.7；对比目标：InfoNCE（温度 τ=0.5）；正交正则项：跨视图残差去相关；损失权重：β_c=0.1（con-对比式），β_s=0.1；优化器：Adam；学习率：5×10^{-4}；权重衰减：1×10^{-4}
DGVAE	模态融合权重（α_v）：0.1（视觉），$1 - \alpha_v$（文本）；图卷积层：2；潜在因子数量（K）：4；软分配温度（τ）：0.1；先验尺度（σ_0）：0.1；目标：带 logits 的二元交叉熵 $+$ KL 正则化 $+$ 互信息代理正则化；正则化权重：β=0.1（KL），λ_{MI}=0.2（MI）；优化器：Adam；学习率：5×10^{-4}；权重衰减：1×10^{-4}

表 A1：传导式同质物品-物品基线参数

附录 B 基线传导算法的参数

表格A1 和A2 展示了第5.2.3节中所述同质与异质传导基线算法的参数。

模型	参数
MambaRec	用户-物品图层（L_{ui}）：2；物品-物品层（L_{mm}）：2；降维因子：$r = 2$；归一化 +丢弃率：层归一化（d） + 丢弃率（0.2）；损失函数：BPR成对排序损失 + ℓ_2 正则化 + 对比式InfoNCE +可选MMD对齐；对比权重：1.0；InfoNCE温度：0.2；MMD权重：1×10^{-2}；MMD核：带宽为 $\sigma = 1.0$的径向基函数；正则化权重：1×10^{-3}；优化器：Adam；学习率：1×10^{-3}；权重衰减：1×10^{-4}
PGL	用户-物品图层（L_{ui}）：2；物品-物品层（L_{mm}）：2；丢弃率：0.2；损失函数：BPR成对排序 +对比式InfoNCE损失；对比温度ature (τ)：0.2；对比项上的正则化权重（λ）：1×10^{-3} 优化器：Adam；学习率：1×10^{-3}；权重衰减：1×10^{-4}
HPMRec	超复代数维度（A）：4；用户-物品图层（L_{ui}）：2；项目–项目多模态图层（L_{mm}）：2；超复交互：Cayley–Dickson 乘法，缩放系数为 $\alpha = 0.1$；丢弃率：0.2；损失函数：贝叶斯个性化排序成对排序 + ℓ_2 正则化 + 自监督超复损失函数；正则化权重：1×10^{-3}；自监督学习权重：0.05；优化器：Adam；学习率：1×10^{-3}；权重衰减：1×10^{-4}
COHESION	用户-物品图层（L_{ui}）：2；多模态物品-物品层（L_{mm}）：2；用户-物品边丢弃率：0.2；融合后稳定器：层归一化 + 线性层 +带泄漏的ReLU；用户-用户邻域聚合：$k_{uu} = 40$；损失函数：BPR成对排序排序 + ℓ_2 正则化；正则化权重：1×10^{-3}；优化器：Adam；学习率：1×10^{-3}；权重衰减：1×10^{-4}
SMORE	用户-物品图层（L_{ui}）：2；物品-物品层（L_{ii}）：2；频谱（快速傅里叶变换）卷积：对特征维度执行实数快速傅里叶变换（rFFT），使用可学习的复数滤波器-ters（3个分支：w_a，w_b，w_f）；非线性融合分支：频域（快速傅里叶变换(z) $\odot$快速傅里叶变换(z)）$\odot w_f$然后逆快速傅里叶变换；丢弃率：0.2；辅助对比损失：InfoNCE；对比权重：0.15；温度（τ）：0.2；正则-化：轻度 ℓ_2；正则化权重：1×10^{-4}；优化器：Adam；学习率：2×10^{-3}；权重衰减：1×10^{-4}；梯度裁剪：最大范数5.0
CMDL	用户-物品图层（L_{ui}）：2；内容视图增强：特征随机失活 $p = 0.25$ + 高斯噪声 $\sigma = 0.05$；物品内容门控；损失函数：BPR 对-wise 排序 + ℓ_2 正则化 +对比对齐；对比权重：0.1；对比温度：$\tau_{pos} = 0.2$（物品），$\tau_{neg} = 0.2$；正则化权重：1×10^{-3}；优化器：AdamW；学习率：1×10^{-3}；权重衰减：1×10^{-4}

表 A2：传导式异质用户–物品基线参数

附录C 基线归纳式算法的参数

表格 B3 和 B4 展示了第 5.2.3 节中所述的同质和异质归纳式基线算法的参数。

模型	参数
GraphSAGE	基准百分位数：90，图神经网络层：2，SAGEConv 聚合器：max，SAGE-Conv 设置：归一化：False，隐藏通道：12，输出通道：2，激活函数：ReLU（层间），损失函数：物品标签上的交叉熵，优化器：Adam，学习率：0.0001。
PinSAGE	百分位数阈值：90，随机游走邻域：50，游走长度：5，新邻居数/每个节点(topm)：20，图神经网络编码器：与 GraphSAGE 相同
GATNE-I	同上（2 层，隐藏层 12，输出层 2，Adam 学习率 0.0001；交叉熵）。在训练物品上使用 k近邻（k ＝ 30），相似度下限：0.0，特征预处理：计算余弦相似度前对特征进行 L2 归一化，嵌入维度：128，负样本：10，优化器：Adam，学习率：0.001，权重衰减：0.000001。

表 B3：归纳式同质物品-物品基线参数

模型	参数
GraphSAGE	图神经网络层：2，SAGEConv 聚合器：max，SAGEConv 设置：project=True，normalize=False，隐藏通道：12，输出通道：2，激活函数：ReLU， 损失函数：基于物品标签的交叉熵， 优化器：Adam，学习率：1e-4，权重衰减：默认。
PinSAGE	百分位数阈值：90，随机游走邻域：50，游走长度：5，每个节点的新邻居（topm）：20，图神经网络编码器与 GraphSAGE 相同
GATNE-I	同上（2 层，隐藏层 12，输出 2，Adam，交叉熵）基于训练物品的 k近邻（k = 30），相似度下限：0.0，特征预处理：在计算余弦相似度前对特征进行 L2 归一化，嵌入维度：128，负样本：10，优化器：Adam，学习率：$1e{-}3$，权重衰减：$1e{-}6$。不使用测试交互进行学习。

表B4：归纳式异构用户—物品基线参数

附录D 消融研究

本节针对 GeMi 系统在传导式和归纳式设置下的各种参数进行了消融研究。

传导实验

我们针对表格中呈现的共享超参数进行敏感性分析1。分析旨在量化 GeMi 在训练正则化、图稀疏性、监督平衡强度和评估配置受到扰动时的鲁棒性。通过调整前k 推荐截断值、边丢弃率和焦点损失参数，同时保持架构特定组件不变，我们得以分离这些全局设计选择的贡献。

我们分析了 GeMi 在传导式设置下对 Top-k 推荐截断值的敏感性（图 D2）。在所有变体中， k取中等值时，性能趋势最稳定、最可靠；而取值过小或过大时， k 会导致评估表现不理想。

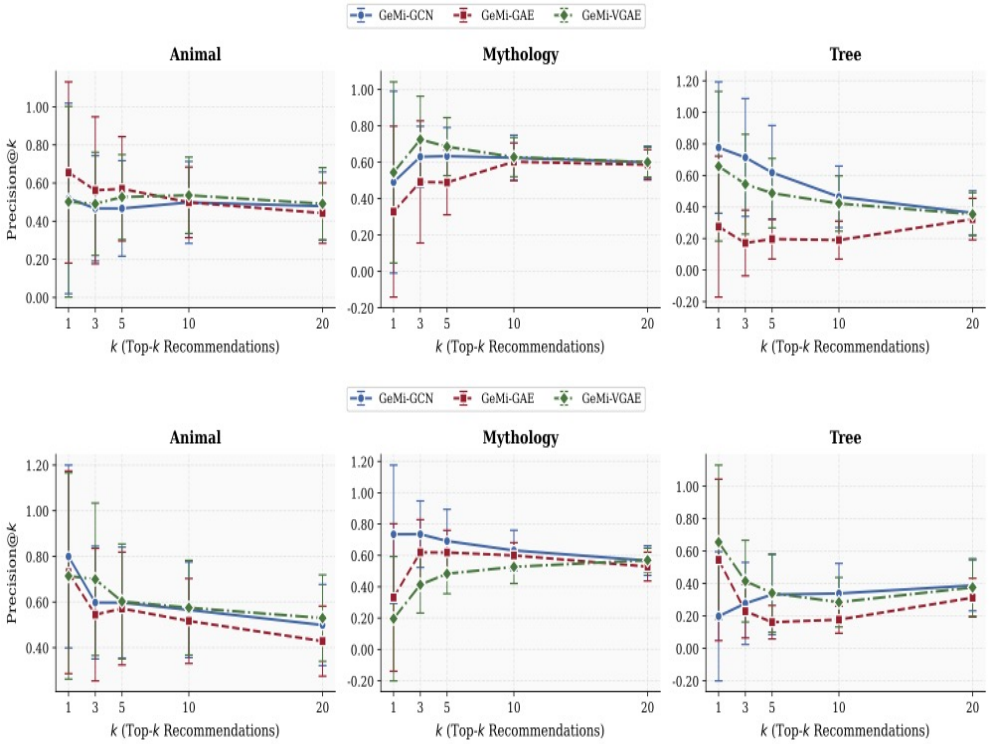


图 D2：不同特征骨干网络（LlamaSigCLIP 和 LlamaVAE）下，GeMi 变体在动物、神话和树类别中的 Top-k 推荐截断值对 Precision@5 的影响。结果以多次运行的均值 $\pm$ 标准差报告。

对于动物，Precision@5 在较低的 k （约为 $k = 1$–5）时最高，并随着 k 增大而逐渐下降，这表明最相关的物品排名靠前，而较长的推荐列表会纳入相关性较低的候选项，从而降低精确率。GeMi-GCN 的峰值表现最出色，其次是 GeMi-VGAE；随着 k 增大，GeMi-GAE 的表现下降幅度相对更大。

对于神话类别，随着 k 增大，性能总体保持相对稳定，在中等取值（$k = 5$–10）时略有提升，随后略微下降。这表明，语义较抽象的类别适合稍长的推荐列表；但随着 k 增大，由于纳入了相关性较低的物品，收益会逐渐减小。GeMi-VGAE 在整个 k 范围内表现出更强的鲁棒性。

对于相对稀疏且更具挑战性的树，Precision@5 在 k 非常小时较低，并随中等取值（$k = 5$–10）而提升，此后趋于稳定。这表明略大的候选池有助于检索相关物品，但进一步增大则会引入噪声。GeMi-VGAE 再次表现出更稳定的行为，而 GeMi-GAE 最为敏感。

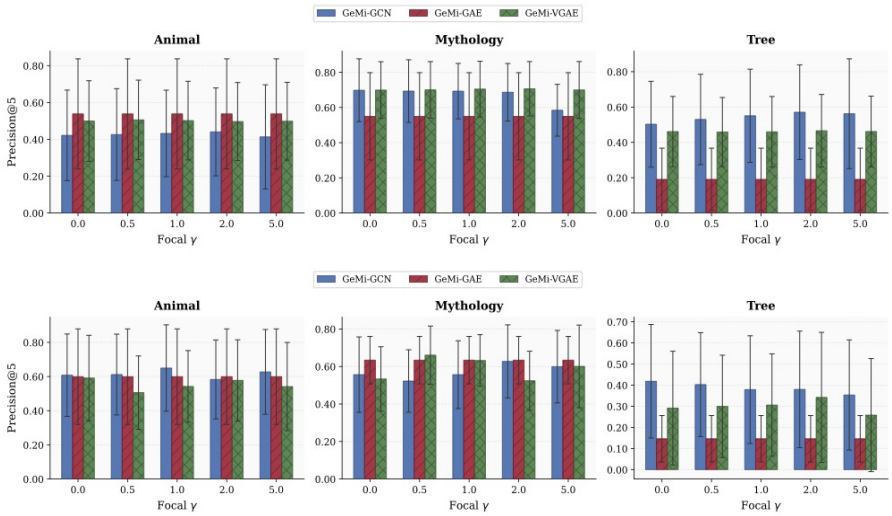


图 D3：在不同特征骨干网络（LlamaSigCLIP和LlamaVAE）下，GeMi 变体在动物、神话和树类别上的 Precision@5 随焦点损失聚焦参数 γ 的变化。结果以多次运行的均值 $\pm$ 标准差报告。

总体而言，结果凸显了由推荐截断值主导的权衡：较小的 k 强调高置信度预测结果，而较大的 k 则以牺牲精确率为代价提高覆盖率。GeMi 对该参数表现出可控的敏感性，各变体的性能趋势保持一致。GeMi-GCN 取得了最强的峰值性能，而 GeMi-VGAE 在不同 k 下提供了更高的鲁棒性，从而支持采用适中的 Top-k 取值，以实现均衡且有效的推荐性能。

我们进一步分析 GeMi 对焦点损失聚焦参数 γ 的敏感性，并使用 LlamaSigCLIP 和 LlamaVAE 特征骨干网络，如图 D3 所示。总体而言，模型在较大范围的 γ 值内表现相对稳定，这表明 GeMi 对焦点调制强度的适度变化并不十分敏感。

在 LlamaSigCLIP 设置下，动物类别在 $\gamma \in [0, 5]$ 上的变化幅度很小，GeMi-GAE 和 GeMi-VGAE 的表现保持稳定，而 GeMi-GCN 的表现略低但同样稳定。这表明，对于以视觉为基础且类别区分相对明显的类别，增加 γ 的收益有限，因为图结构和特征质量已足以应对类别不平衡。相比之下，神话类别的表现随着 γ 从 0 增加到约 1–2 而略有提升，之后趋于饱和。这说明，适度聚焦有助于突出语义抽象类别中的难分类样本，但过度聚焦会带来递减收益。对于树类别，表现总体较为稳定，在适度的 γ下略有提升，这表明焦点重加权对稀疏类别有轻微益处，但不会显著改变表现趋势。

在 LlamaVAE 设置下，可以观察到类似但略为明显的趋势。在动物类别中，适度的 γ （约为 1）使 GeMi-GCN 受益并取得最佳表现，而更高的取值会导致轻微波动，这可能是因为对困难样本的强调过度。GeMi-GAE 和 GeMi-VGAE 的表现相对稳定，不过 GeMi-VGAE 在极端的 γ 取值下略有退化。在神话类别中，适度的 γ （约为 0。5–1）能提升 GeMi-GAE 和 GeMi-VGAE 的表现，反映出它们对困难和模糊样本的处理有所改善；而非常高的 γ 会因过度聚焦而导致表现略有下降。对于类别不平衡最严重的树类别，适度的 γ 能提升 GeMi-VGAE 的表现，但随着 γ 增加，GeMi-GCN 的表现略有下降，这表明过度聚焦可能会抑制有用的多数类信号。

在两种特征设置下，结果均表明， γ 控制着这样一种权衡：既要强调困难样本，又要保持来自较容易样本的梯度贡献稳定。较低的取值（$\gamma \approx 0$）会使焦点损失退化为标准的二元交叉熵，从而带来更均匀的学习；较高的取值（$\gamma > 2$）则会过度放大困难样本，可能导致优化不稳定。实验结果显示，适中的取值（$\gamma \in [0.5, 2]$）在不同类别和模型变体上始终能取得最佳或接近最优的性能。

总之，GeMi 对焦点损失参数表现出良好的鲁棒性，性能随 γ平滑变化。适度聚焦能够带来稳定收益，尤其适用于语义复杂且类别不平衡的类别；而极端取值并无必要，且可能略微降低性能。这些发现支持将 $\gamma = 2.0$ 选作一种平衡且有效的默认值，适用于不同的特征骨干网络和 GeMi 变体。

最后，我们分析了 GeMi 对边丢弃的敏感性。在 LlamaSigCLIP 和 LlamaVAE 特征骨干网络下，边丢弃会在训练期间对交互图进行随机稀疏化（图 D4）。总体而言，GeMi 表现出很强的鲁棒性；当丢弃率从 0.0 变为 0.4 时，Precision@5 仅有轻微变化，表明其在结构扰动下仍能稳定地进行表征学习。

在 LlamaSigCLIP 设置下，所有丢弃率下的性能总体保持稳定。对于动物，GeMi-GCN 在中等丢弃率（~ 0.1）时略有提升，体现出轻微的正则化收益；而较高的丢弃率会因边移除过多而导致性能略有下降。GeMi-GAE 和 GeMi-VGAE 的表现几乎不变。在神话数据集上，GeMi-VGAE 始终取得最佳性能（~ 0.69），且变化可以忽略不计；GeMi-GCN 在中等丢弃率下也略有提升。在树数据集上，GeMi-GCN 和 GeMi-VGAE 的表现保持稳定，表明即使在稀疏情形下，适度稀疏化也不会妨碍信息传播。

在 LlamaVAE 设定下，观察到略高的敏感性。对于动物，GeMi-GCN 受益于中等丢弃率（峰值约为 0.1），表明通过去除冗余边改善了泛化，而更高的丢弃率则会降低性能。在神话中，GeMi-GCN 在中等丢弃率（$\sim$ 0.2）处达到峰值，表明噪声过滤有效，而过度稀疏化则会降低性能。GeMi-GAE 保持稳定，GeMi-VGAE 则表现稳定但较低。在

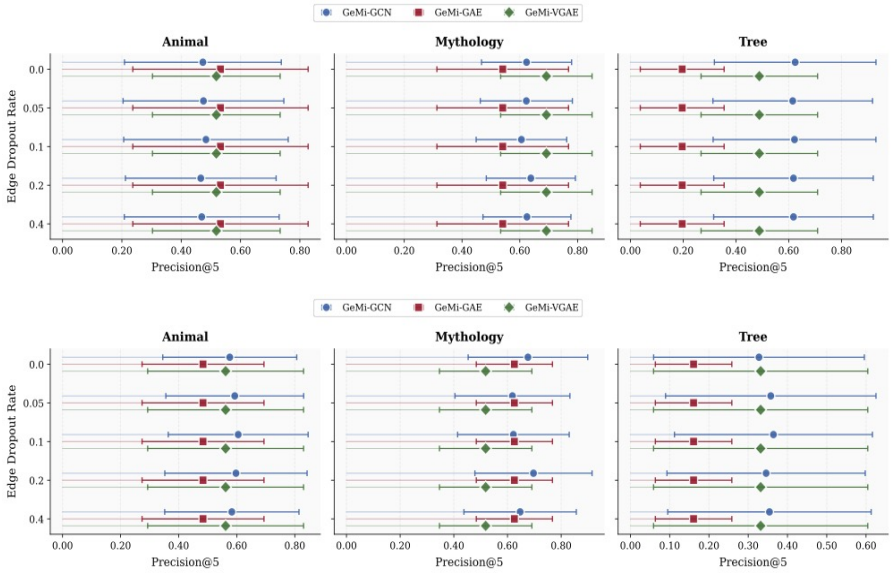


图 D4：不同特征骨干网络（LlamaSigCLIP 和 LlamaVAE）下，GeMi 变体在动物、神话和树类别上的 Precision@5 随边丢弃参数变化的效果。结果以多次运行的均值 ± 标准差报告。

树，中等丢弃率再次使 GeMi-GCN 受益，而更高的丢弃率会降低连通性并略微损害性能。

总体而言，结果凸显了正则化与信息损失之间的权衡。较低的丢弃率保留了完整连通性，但也保留了噪声；适中的丢弃率（0.05–0.2）提升了鲁棒性和泛化能力；较高的丢弃率（≥ 0.4）则削弱了消息传递。GeMi-VGAE 表现出最高的鲁棒性，而 GeMi-GCN 从适度正则化中获益最大。这些发现支持采用适度的边丢弃，将其作为一种有效且稳定的设计选择。

归纳实验

我们研究归纳性能对 k近邻图参数 k 的敏感性，该参数控制 (i) 训练项目间子图的密度，以及 (ii) 每个测试项目与训练集的连接情况。在不同标签和骨干网络下，k 的影响呈非单调变化：增大 k 并非总能带来好处，在某些情况下，较小的 k 反而会产生更好的性能。这一结果表明存在偏差-方差权衡：较大的 k 会引入更多相似度较低的邻居，可能导致过度平滑；而过小的 k 则可能造成图过于稀疏，使信息传递减弱。

非单调性：即使固定模型和特征流水线，k 的影响也可能表现出显著的非单调性。例如，在 SigCLIP + 图自编码器设置下，将 k 从 30 降至 10 时，性能显著提升，但随后

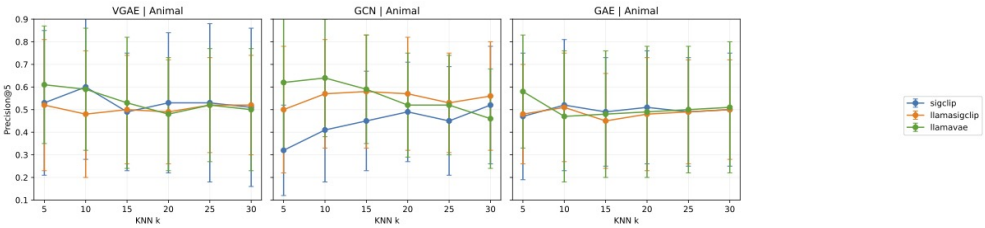


图 D5：RQ4（k近邻消融）：Animal 的 Precision@5 与 k近邻图构建参数 k的关系。误差条表示多次运行的均值 $\pm$ 1 标准差。

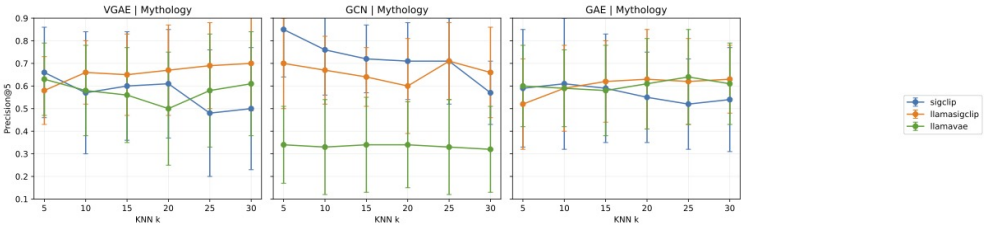


图 D6：RQ4（k近邻消融）：Mythology 的 Precision@5 与 k近邻图构建参数 k的关系。误差条表示多次运行的均值 $\pm$ 1 标准差。

在 $k = 5$ 时又有所下降。这表明，最适合图密度的 k近邻值处于中间范围，而非极端值。

标签依赖敏感性：最优的 k 取决于标签，但不同骨干网络和特征流水线中的敏感性模式并不一致。在许多设置中，Tree 标签往往对 k 更敏感，而 Mythology 的敏感程度则从相对稳定到对 k高度敏感不等。一种合理的解释是，Tree 的信号较弱（通常在视觉上较为隐晦，在文本中也不总是占主导），因此邻域大小的选择会更直接地影响能否聚合到足够的信息性信号；此外，这也促使我们采用树特定边增强。

特征依赖性：首选的 k 也取决于特征。不同的嵌入会带来不同的近邻质量。较小的 k 在一个嵌入空间中可以保留质量更高的近邻，但在另一个空间中可能会使图过度稀疏。因此， k 在消息传递中起到平滑超参数的作用，其适用范围取决于基础特征几何结构。

骨干网络依赖性：与图卷积网络和图自编码器相比，变分图自编码器往往对 k表现出更高的敏感性，其性能有时会随着 k 的变化而出现更明显的上升和下降。这与变分图自编码器的生成式目标相符，因为图密度的变化可能会重塑潜在结构，并导致局部最优解。

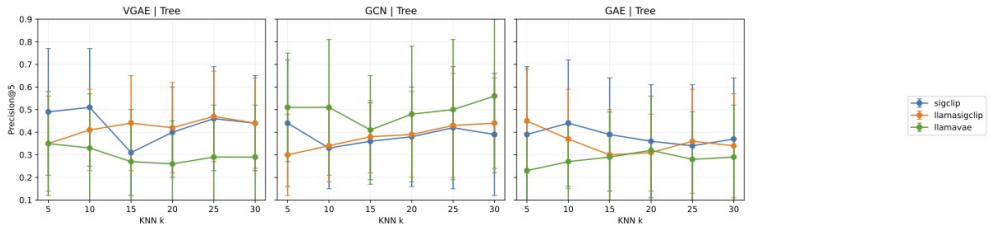


图 D7：RQ4（k近邻消融）：树的 Precision@5 与 k近邻图构建参数 k的关系。误差条表示多次运行的均值 $\pm$ 1 标准差。



Fig. E8：使用 UP 的用户特征 FAR

附录 E 民间艺术推荐在线工具

在正在开展的工作中（Hotwani 和 Dutta （2025）），作者开发了一个系统——民间艺术推荐用户偏好（UPFAR）——用于收集相关信息（例如文化接触、博物馆参观、艺术与音乐熟悉度、观看叙事性手卷画演出的经历）以及对面板的评分（例如（= 1）/不（= 0））。图E8展示了从用户处收集的数据类别，图E9展示了架构图以及部署于线上的工具（图E10）。该系统收集的数据已输入 GeMi 流程。以下细节值得关注：

•数据预处理：通过删除无效条目和重复项，并仅保留属于训练画面的交互，对用户与画面交互数据进行清理，以避免测试信息泄漏到用户画像中。

• 计算偏好：为将原始评分转换为用户对标签的偏好，使用 $[0, 1]$中的最小-最大归一化对评分进行归一化。然后，

– 对每位用户，计算其基准评分（其归一化评分的平均值）。– 对每个标签（动物、神话、树），计算包含该标签的物品的平均评分。

据此，相对提升定义如下：

$$lift = \text{average rating for that label} - \text{user baseline}$$

这意味着，如果用户对某个概念的评分高于其平均水平，该概念就会获得正向提升度，否则则为负向提升度。

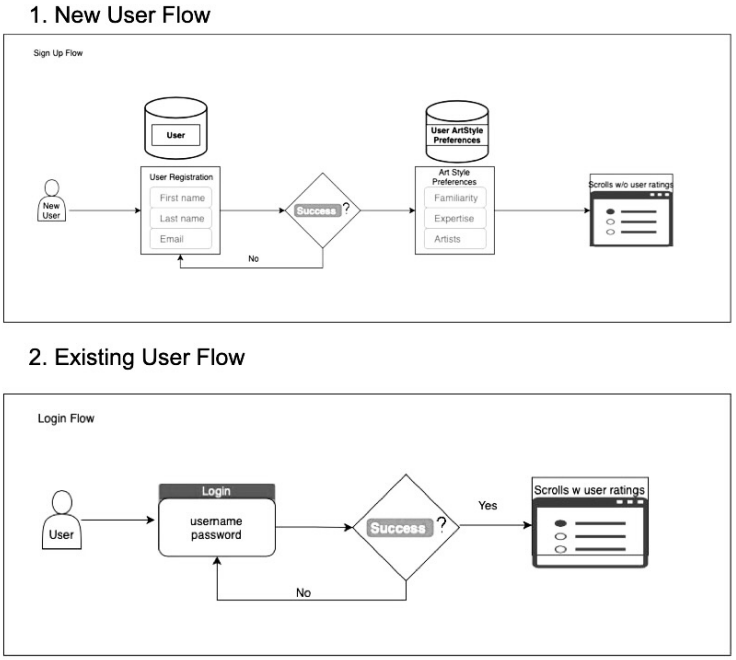


图 E9：UPFAR 的架构图

- 处理稀疏性：许多用户对某个标签的交互次数很少。因此，采用了贝叶斯式平滑，将用户的提升度与全局先验提升度（基于所有用户计算）相结合。该结果根据用户拥有的样本数量进行加权，从而避免稀疏用户的估计不稳定。
- 构建二元偏好：每个用户的提升值通过 Sigmoid 函数映射为类似概率的偏好，具体如下：

$$\text{preference} = \sigma(\text{gain} \times \text{lift}).$$

  增益控制偏好偏离 0.5 的程度。这样我们就得到了一个简洁的偏好矩阵，其中每个用户对于动物、神话和树概念的取值都在 0 到 1 之间。同时，对于每个用户，我们还根据最高评分记录其互动次数最高的前 k 个面板（$k = 5$）。这将在评估期间用于构建用户嵌入。
- 用户画像增强：从 UPFAR 收集到的真实世界数据非常少。为了解决这一问题，我们通过自助法增强生成了一个规模较大且真实的数据集。对于每个合成用户（目标为 $= 10,000$ 名用户），我们随机选取一名真实用户作为基础用户，从该用户排名前 k 的项目中（有放回地）抽取 5 次交互，并以一定概率（$p = 0.3$）随机替换少数项目，换成其他已观测项目，以增加多样性。对于偏好，我们使用基础用户的偏好向量进行增益缩放（使部分

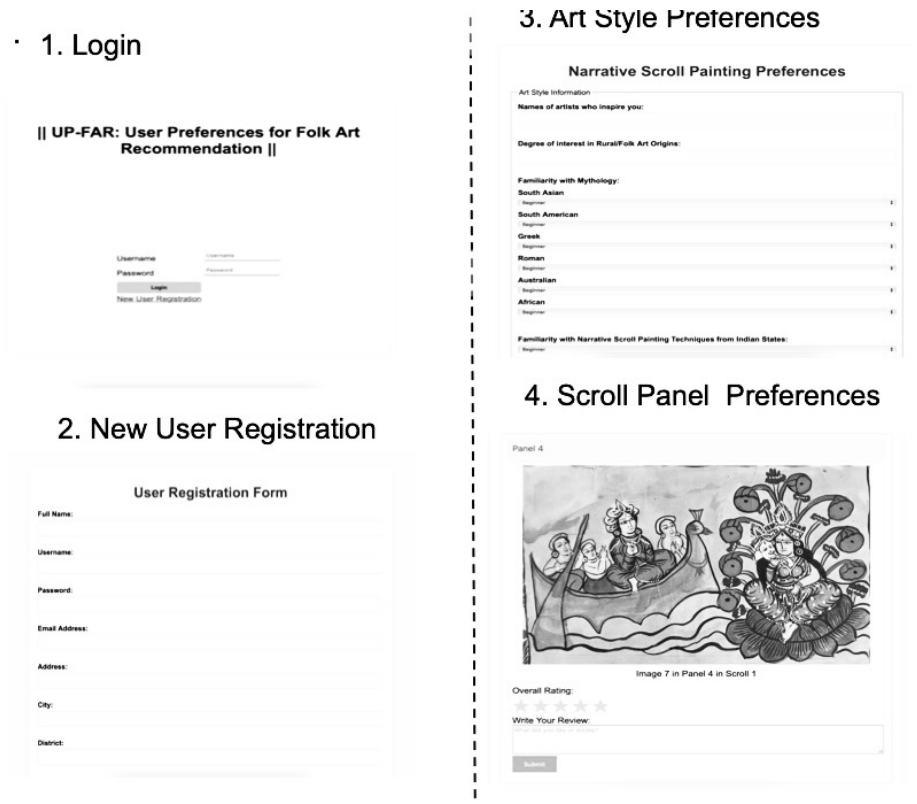


图 E10：UPFAR 部署的屏幕

用户的偏好更极端或更中性），施加小幅偏差偏移（以模拟用户倾向），并为每个标签添加高斯噪声（以增加多样性），然后将值裁剪回所需范围 $[0, 1]$。

表E5 展示了传导式设置下的结果。结果表明，使用变分图自编码器进行图结构学习对这些数据有显著益处，但也可以使用 SIGCLIP 或变分自编码器来学习概念。有趣的是，通过大语言模型规范化进行增强（例如使用 SIGCLIP 特征学习神话概念）仅带来微小收益。表 E6 展示了归纳式设置下的结果。使用 GCN 或图自编码器表示的 SIGCLIP 能够提供良好的推荐；然而，对于真实世界数据，树概念仍然难以推荐。

参考文献

Anand V, Maurya AK（2024）基于图神经网络的推荐系统综述。ACMTrans Inf Syst 43(1)

模型	动物	神话	Tree
SIGCLIP + 图卷积网络	0.52 ± 0.22	0.68 ± 0.17	0.40 ± 0.20
SIGCLIP + 图自编码器	0.55 ± 0.22	0.49 ± 0.14	0.33 ± 0.22
SIGCLIP + 变分图自编码器	0.44 ± 0.16	0.74 ± 0.18	0.43 ± 0.22
变分自编码器 + 图卷积网络	0.27 ± 0.09	0.59 ± 0.04	0.01 ± 0.04
变分自编码器 + 图自编码器	0.50 ± 0.22	0.69 ± 0.11	0.22 ± 0.16
变分自编码器 + 变分图自编码器	0.60 ± 0.11	0.65 ± 0.09	0.60 ± 0.21
Llama SigCLIP + 图卷积网络	0.56 ± 0.19	0.71 ± 0.16	0.44 ± 0.25
Llama SigCLIP + 图自编码器	0.43 ± 0.20	0.58 ± 0.15	0.40 ± 0.17
Llama SigCLIP + 变分图自编码器	0.59 ± 0.16	0.69 ± 0.15	0.44 ± 0.13
Llama 变分自编码器 +图卷积网络	0.45 ± 0.19	0.53 ± 0.19	0.39 ± 0.24
Llama 变分自编码器 +图自编码器	0.43 ± 0.20	0.43 ± 0.23	0.33 ± 0.16
Llama 变分自编码器 +变分图自编码器	0.48 ± 0.17	0.39 ± 0.20	0.31 ± 0.22

表 E5：使用真实世界数据进行**传导式** 模型训练，以指导用户偏好矩阵生成。

模型	动物	神话	Tree
SIGCLIP + 图卷积网络	0.45 ± 0.14	0.79 ± 0.25	0.23 ± 0.23
SIGCLIP + 图自编码器	0.39 ± 0.20	0.58 ± 0.20	0.24 ± 0.26
SIGCLIP + 变分图自编码器	0.65 ± 0.27	0.71 ± 0.13	0.30 ± 0.21
变分自编码器 + 图卷积网络	0.64 ± 0.18	0.61 ± 0.09	0.24 ± 0.30
变分自编码器 + 图自编码器	0.53 ± 0.09	0.54 ± 0.13	0.19 ± 0.16
变分自编码器 + 变分图自编码器	0.60 ± 0.20	0.64 ± 0.19	0.25 ± 0.19
Llama SigCLIP + 图卷积网络	0.56 ± 0.27	0.52 ± 0.19	0.25 ± 0.20
Llama SigCLIP + 图自编码器	0.49 ± 0.22	0.54 ± 0.13	0.36 ± 0.23
Llama SigCLIP + 变分图自编码器	0.47 ± 0.21	0.59 ± 0.17	0.26 ± 0.14
Llama 变分自编码器 +图卷积网络	0.31 ± 0.23	0.31 ± 0.15	0.33 ± 0.17
Llama 变分自编码器 +图自编码器	0.51 ± 0.18	0.56 ± 0.12	0.38 ± 0.12
Llama 变分自编码器 +变分图自编码器	0.53 ± 0.19	0.48 ± 0.13	0.37 ± 0.17

表 E6：使用真实世界数据进行**归纳式** 模型训练，以指导用户偏好矩阵生成的结果。

Aroyo L, Brussee R, Rutledge L, 等（2007）个性化博物馆体验：荷兰国立博物馆案例。在：博物馆与网络 2007年

Bajpai LP（2015）孟加拉帕塔奇特拉绘画中的神话与民间故事：传统与现代性。《Chitralekha国际艺术与设计期刊》5(1)：124–131

Bennett J，Lanning S（2007）Netflix竞赛。在：KDD Cup Workshop 2007论文集。ACM，第3–6页，网址 http://www.cs.uic.edu/~liub/KDD-cup-2007/NetflixPrize-description.pdf

Bentley JL（1975年）寻找固定半径近邻的复杂性。《ACM通讯》18(8)：421–425

van den Berg R、Kipf T、Welling M（2018）图卷积矩阵补全。KDD 深度学习日，1–7页；Bohnert F、Zukerman I（2009）使用基于关键词的方法自适应地预测对博物馆展品的兴趣。载于：澳大拉西亚人工智能进展联合会议论文集，第656–665页；Cai D、Qian S、Fang Q等（2022a）用于个性化微视频推荐的自适应反瓶颈多模态图学习网络。载于：ACM国际多媒体会议论文集，第581–590页；Cai D、Qian S、Fang Q等（2022b）用于个性化微视频推荐的异质分层特征聚合网络。IEEE 多媒体汇刊 24：805–818

Cao M, Yang FF, Jin Y, 等（2025）基于亚马逊共购图的商品推荐图神经网络。网址 https://arxiv.org/abs/2508.14059

Cen Y, Zou X, Zhang J, 等（2019）属性多重异质网络的表征学习。载于：ACM SIGKDD知识发现与数据挖掘国际会议论文集。ACM，第 1358–1368 页

Chen F, Wang J, Wei Y, 等（2022）打破孤立：通过边级调制进行多模态图融合，以实现多媒体推荐。载于：ACM国际多媒体会议论文集，第 385–394 页

Chen Y, Wu L, Zaki M (2020) 面向图神经网络的迭代式深度图学习：更优且稳健的节点嵌入。载于：神经信息处理系统进展，第 33 卷。Curran Associates, Inc.，第 19314–19326 页 Chen Z, Xu J, Wang H, 等 (2025a) 超复数提示感知多模态推荐。网址 https://arxiv.org/abs/2508.10753，arXiv:2508.10753Chen Z, Xu J, Wang H, 等 (2025b) 超复数提示感知多模态推 - 荐。载于：ACM 信息与知识管理国际会议论文集，第 403–414 页 Deldjoo Y, Schedl M, Cremonesi P, 等 (2020) 利用多媒体内容的推荐系统。ACM Comput Surv 53(5)。https://doi.org/10.1145/3407190Dimock EC (编) (1963) 《爱的窃贼：孟加拉宫廷与乡村故事》。芝加哥大学出版社 Du X, Wu Z, Feng F, 等 (2022) 面向多媒体推荐的不变表示学习。载于：ACM 多媒体国际会议论文集，第 619–628 页

Dutta G (1939) 帕图阿音乐。加尔各答大学

Dutta H, Tiwari A(2024) 基于多模态图神经网络的内容式艺术推荐。载于：IEEE国际知识图谱会议 (ICKG)，第 67–74 页

Fan Q, Yu P, Tan Z, 等 (2025) Befa：一种通用的多媒体推荐行为驱动特征适配器。载于：AAAI人工智能会议论文集

Fioretti ML, Laterza N, Preziosa A, 等 (2025) 大语言模型支持多模态推荐的能力有多强？对 LLMRec 的一项可复现性研究。载于：ACM推荐系统会议论文集，第 774–782 页

Frost S, Thomas MM, Forbes AG (2019) “我不喜欢的艺术”：面向视觉艺术的反推荐系统。博物馆与网络

Gao C, Zheng Y, Li N, 等 (2023) 推荐系统图神经网络综述：挑战、方法与发展方向。ACM 推荐系统汇刊 1(1)

Geng S, Tan J, Liu S 等（2023）VIP5：面向推荐的多模态基础模型。在：ACM推荐系统会议论文集

Grieser K, Baldwin T, Bohnert F 等（2011）运用本体相似性和文档相似性估算博物馆展品之间的相关性。《计算与文化遗产》3(3)

Guo W, Su R, Tan R 等（2021）用于点击率预测的双图增强嵌入神经网络。在：ACM SIGKDD知识发现与数据挖掘会议论文集，第496–504页

Guo Z, LiJ,Li G 等（2024）LGMRec：面向多模态推荐的局部与全局图学习。在：AAAI人工智能会议论文集

Hamilton WL, Ying R, Leskovec J（2017）大型图上的归纳式表示学习。在：第31届神经信息处理系统国际会议论文集，第1025–1035页

Han T, WangP,Niu S, 等 (2022) 模态匹配模态：面向推荐的模态解耦物品表示预训练。见：ACM网络会议论文集，第2058–2066页

Harper FM, Konstan JA (2015年) MovieLens数据集：历史与背景。ACM Trans Interact Intell Syst5(4)

He R,McAuley J (2016) Vbpr：基于隐式反馈的视觉贝叶斯个性化排序。载于：第三十届 AAAI 人工智能会议论文集。AAAI 出版社，第 144–150 页

He X, Deng K, Wang X, et al (2020) Lightgcn：简化并增强用于推荐的图卷积网络。载于：第43届国际 ACM SIGIR 信息检索研究与发展会议论文集，第 639–648 页 Hotwani S, Dutta H (2025) 一款用于收集用户对民间艺术推荐系统偏好的在线工具。MSS体验式学习博览会，纽约州立大学布法罗分校 Huang C, Huang H, Yu T, et al (2025) 基础模型驱动的推荐系统综述：从基于特征、生成式到智能体范式。网址 https://arxiv.org/abs/2504.16420

Hui B, Yang Jea（2024）Qwen2.5-Coder 技术报告。网址 https://arxiv.org/abs/2409.12186，arXiv:2409.12186

Jain J（编）（1998）《图像说书人：印度艺术叙事传统见解》。Marg Publications，印度

Ji Y、Luo D、Liu C 等（2025）为多模态推荐生成负样本。载于：ACM国际多媒体会议论文集，第 6007–6016 页 Kim J、Kang H、Kim S 等（2025）针对模态缺失场景，为推荐解耦并生成模态。载于：国际 ACM SIGIR 信息检索研究与发展会议论文集，第 1820–1829 页

Kipf TN、Welling M（2016）变分图自编码器。网址 https://arxiv.org/abs/1611.07308

Kipf TN, Welling M（2017年）基于图卷积网络的半监督分类。载于：国际学习表征会议论文集，ICLR'17

Korom FJ（2006）《画家之村：西孟加拉邦的叙事卷轴》。新墨西哥博物馆出版社，圣菲。Krippendorff K（编）（2004）《内容分析：方法论导论》。Sage Publications。Kuflik T、Stock O、Zancanaro M 等（2011）互动式博物馆中的访客指南：展示、交流与反思。《计算与文化遗产》3。Lachaud G、Conde-Cespedes P、Trocan M（2023）使用图神经网络在真实引文网络中比较归纳式学习与传导式学习。载于：2022年IEEE/ACM社会领域进展国际会议论文集

《网络分析与挖掘》，第 534–540 页

Le Scao T、Fan Aea（2023）BLOOM：一个拥有1760亿参数的开放获取多语言语言模型。网址 https://arxiv.org/abs/2211.05100，arXiv:2211.05100Li C、Hu L、Shi C 等（2021）序列感知异构图神经协同过滤。载于：SIAM国际数据挖掘会议（SDM）论文集，第64–72页。Li H、Du H、Li Y 等（2025a）教我如何去噪：通过引导校准实现多模态推荐系统去噪的通用框架。载于：ACM国际网络搜索与数据挖掘会议论文集，第782–791页。Li J、Wang S、Zhang Q 等（2025b）以公平性为基础生成：面向不完整多模态推荐的模态扩散反事实框架。载于：2025年ACM网络会议论文集，第2787–2798页。Li R、Wang S、Zhu F 等（2018）自适应图卷积神经网络。载于：第三十二届AAAI人工智能会议论文集。AAAI 出版社。Li X、She J（2017）用于推荐系统的协同变分自编码器。载于：ACM SIGKDD 知识发现与数据挖掘国际会议论文集，第305–314页。Li Z、Cui Z、Wu S 等（2019）Fi-GNN：通过图神经网络建模特征交互以进行点击率预测。载于：ACM 信息与知识管理国际会议论文集，第539–548页。Liang D、Krishnan RG、Hoffman MD 等（2018）用于协同过滤的变分自编码器。载于：万维网会议论文集，第689–698页。Lin J、Dai X、Xi Y 等（2024）推荐系统如何从大语言模型中受益：综述。ACM信息系统汇刊。Lin J、Cao Y、Yu Y 等（2025a）推荐系统的扩散模型：从内容分布到内容生成。载于：ACM SIGKDD知识发现与数据挖掘会议论文集，第6074–6085页。Lin X、Liu R、Cao Y 等（2025b）用于多模态推荐的对比模态解耦学习。《ACM信息系统汇刊》43(3)：1–31。Liu F、Cheng Z、Zhu L 等（2021a）用于推荐的兴趣感知消息传递图卷积网络。载于：万维网会议论文集，第1296–1305页

Liu F, Cheng Z, Zhu L, 等（2022a）一种用于推荐中属性缺失问题的属性感知注意力图卷积网络模型。《IEEE 知识与数据工程汇刊》34(9):4077–4088

Liu F, Cheng Z, Zhu L, 等（2022b）一种用于推荐中属性缺失问题的属性感知注意力图卷积网络模型。《IEEE 知识与数据工程汇刊》34(9):4077–4088

Liu F, Chen H, Cheng Z, 等（2023a）面向推荐的解耦多模态表示学习。《多媒体汇刊》25:7149–7159

Liu H, Wei Y, Yin J, 等（2023b）Hs-gcn：用于推荐的汉明空间图卷积网络。《IEEE 知识与数据工程汇刊》35(6):5977–5990

Liu P, Zhang L, Gulla JA（2023c）预训练、提示与推荐：推荐系统中语言建模范式适配的综合综述。《计算语言学协会汇刊》11:1553–1571

Liu Q、Wu S、Wang L（2017年）Deepstyle：学习用户对视觉推荐的偏好。载于：第40届ACM国际信息检索研究与发展会议论文集，SIGIR '17，第841–844页

Liu W、Liu Q、Tang R 等（2020）基于物品关系的电子商务个性化重排序。载于：ACM信息与知识管理国际会议论文集，第925–934页

Liu X、Tao Z、Shao J 等（2022c）EliMRec：消除多媒体推荐中的单模态偏差。载于：ACM国际多媒体会议论文集，第687–695页

Liu Y、Yang S、Lei C 等（2021b）利用多模态侧信息对图变换器进行预训练以实现推荐。载于：ACM国际多媒体会议论文集，第2853–2861页

Ma J, Zhou C, Cui P, 等（2019）学习用于推荐的解耦表示。载于：神经信息处理系统国际会议论文集。Mair VH（编）（2019）《绘画与表演：中国图像说唱及其印度起源》。夏威夷大学出版社。Markwald M, Demidova E（2024）Refuel：面向不平衡神经节点分类的规则提取。《Machine Learning》113(9):6227–6246。McCutchion DJ, Bhowmick SK（1999）《孟加拉的帕图亚人与帕图亚艺术》。Firma KLM。McPherson M, Smith-Lovin L, Cook JM（2003）物以类聚：社交网络中的同质性。《社会学年度评论》27:415—-444

Meng F, Meng Z, Jin R, 等（2025a）Tamer：用于多模态推荐的兴趣树增强模态图推荐器。载于：ACM国际多媒体会议论文集，第5998–6006页。Meng F, Meng Z, Jin R, 等（2025b）Doge：大语言模型增强型超知识图谱推荐器，用于多模态推荐。AAAI人工智能会议论文集39:12399–12407。Messina P, Dominguez V, Parra D, 等（2019）基于内容的艺术作品推荐：将绘画元数据与神经网络及人工设计的视觉特征相结合。《用户建模与用户适应交互》29(2):251–290。Mu Z, Zhuang Y, Tan J, 等（2022）学习用于多媒体推荐的混合行为模式。载于：ACM国际多媒体会议论文集，第376–384页

Ni J, Li J, McAuley J（2019年）使用远程标注评论和细粒度方面为推荐提供依据。载于：自然语言处理实证方法会议和自然语言处理国际联合会议（EMNLP-IJCNLP）论文集，第188–197页

Ong RK, Khong AWH（2025年）基于频谱的模态表示融合图卷积网络，用于多模态推荐。载于：第十八届ACM国际网络搜索与数据挖掘会议论文集。ACM，WSDM '25，第773–781页，https://doi.org/10.1145/3701551.3703561，网址 http://dx.doi.org/10.1145/3701551.3703561

Pariser E（2011年）《过滤气泡：互联网对你隐瞒了什么》。企鹅出版集团

Pei C, Zhang Y, Zhang Y 等（2019年）用于推荐的个性化重排序。载于：ACM推荐系统会议论文集，第3–11页

RadfordA、Kim JW、HallacyC 等（2021）从自然语言监督中学习可迁移视觉模型。载于：第 38 届国际机器学习会议论文集，《机器学习研究论文集》第 139 卷。PMLR，第 8748–8763 页

Ren K、Ju CY、Lee DH（2025）用于多模态推荐的多尺度双向注意力模态对齐。载于：第 34 届 ACM 信息与知识管理国际会议论文集。ACM，CIKM '25，第 2493–2502 页，https://doi.org/10.1145/3746252.3761222，网址 http://dx.doi.org/10.1145/3746252.3761222

Rendle S（2010）因子分解机。载于：2010年 IEEE 数据挖掘国际会议论文集，第 995–1000 页

Salah A、Truong Q、Lauw HW（2020）Cornac：多模态推荐系统比较框架。《机器学习研究杂志》21(95)：1–5。网址 http://jmlr.org/papers/v21/19-805.htmlSemeraro G、Lops P、De Gemmis M 等（2012）用于个性化访问数字艺术品的基于民众分类法的推荐系统。《计算与文化遗产》5(3) Sharma K、Lee Y、Nambi S 等（2024）社交推荐系统图神经网络综述。ACM Comput Surv 56(10) Shazeer N（2020）GLU变体改进 Transformer。CoRR abs/2002.05202 Shehmir S、Kashef R（2025）Llm4rec：大语言模型在推荐系统中的集成综述——方法、应用与挑战。《未来互联网》17(6) Stock O、Zancanaro M、Busetta P 等（2007）面向博物馆参观者的自适应智能信息呈现系统 PEACH。《用户建模与用户适应交互》17(3)：257–304 Su J 、 Ahmed M 、 Lu Y 等（2024）Roformer：采用旋转位置嵌入的增强型 Transformer。《神经计算》568 Su Y、Zhang R、Erfani SM 等（2020a）检测推荐系统中的有益特征交互。载于：AAAI人工智能会议，网址 https://api.semanticscholar.org/CorpusID:234771712Su Y、Zhao Y、Erfani S 等（2022）检测推荐系统中任意阶的有益特征交互。载于：ACM SIGKDD知识发现与数据挖掘会议论文集，第 1676–1686 页 Sun J、Zhang Y、Guo W 等（2020a）用于推荐的邻居交互感知图卷积网络。载于：国际 ACM SIGIR 信息检索研究与发展会议论文集，第 1289–1298 页 Sun R、Cao X、Zhao Y 等（2020b）用于推荐系统的多模态知识图谱。载于：第29届ACM信息与知识管理国际会议论文集，第 1405–1414 页 Tao Z、Wei Y、Wang X 等（2020）MGAT：用于推荐的多模态图注意力网络。《信息处理与管理》57(5)：102277 Team G、Riviere M、Pathak Sea（2024）Gemma 2：以实用规模改进开放语言模型。网址 https://arxiv.org/abs/2408.00118，arXiv:2408.00118Touvron H、Lavril T、Izacard G 等（2023）Llama：开放且高效的基础语言模型。arXiv预印本 arXiv:230213971

Tran N, Lauw HW（2022）对齐来自评分和文本内容的双重解耦用户表示。载于：ACM SIGKDD知识发现与数据挖掘会议论文集，第1798–1806页。Truong Q, Salah A, Lauw H（2021a）多模态推荐系统：实践探索。载于：第十五届ACM推荐系统会议论文集，第834–837页。Truong Q, Salah A, Tran T, 等（2021b）探索推荐系统中的跨模态利用。IEEE互联网计算 25(4):50–57。Wang Q, Wei Y, Yin J, 等（2023a）Dualgnn：用于多媒体推荐的双图神经网络。Trans Multi 25:1074–1084。Wang Q, Wei Y, Yin J, 等（2023b）Dualgnn：用于多媒体推荐的双图神经网络。IEEE 多媒体汇刊 25:1074–1084。Wang X,He X, Wang M, 等（2019）神经图协同过滤。载于：国际 ACM SIGIR 信息检索研究与发展会议论文集，SIGIR'19，第165–174页。Wang X, Jin H, Zhang A, 等（2020）解耦图协同过滤。载于：国际 ACM SIGIR 信息检索研究与发展会议论文集，第1001–1010页。Wang Y, Stash N, Aroyo L, 等（2008）基于语义增强博物馆藏品的推荐。Web Semant 6(4):283–290。Wang Y, Aroyo LM, Stash N, 等（2009）在线及现场打造个性化博物馆导览。跨学科科学评论 34(2):141–156。Wei TR, Fang Y（2025）推荐系统中的扩散模型：综述。arXiv:2501.10548Wei W, Ren X, Tang J, 等（2024）Llmrec：结合图增强进行推荐的大语言模型。载于：ACM国际网络搜索与数据挖掘会议论文集，第806–815页。Wei Y, Wang X, Nie L, 等（2019）Mmgcn：用于微视频个性化推荐的多模态图卷积网络。载于：第27届ACM国际多媒体会议论文集，第1437–1445页。Wei Y, Wang X, Nie L, 等（2020）利用隐式反馈进行多媒体推荐的图精炼卷积网络。载于：第28届ACM国际多媒体会议论文集，第3541–3549页。Wei Y, Wang X, He X, 等（2022）用于多媒体推荐的层次化用户意图图网络。IEEE 多媒体汇刊 24:2701–2712

Wu J, Wang X, Feng F, 等（2021）用于推荐的自监督图学习。载于：国际 ACM SIGIR 信息检索研究与发展会议论文集，第726–735页。Wu J, He X, Wang X, 等（2022a）用于上下文感知推荐系统的图卷积机。Front Comput Sci 16(6)。Wu L, Zheng Z, Qiu Z, 等（2024）面向推荐的大语言模型综述。World Wide Web 27(60)。Wu S, Sun F, Zhang W, 等（2022b）推荐系统中的图神经网络：综述。ACM Comput Surv 55(5)。Wu S, Sun F, Zhang W, 等（2023）推荐系统中的图神经网络：综述。ACM Computing Surveys 55(5)。Xu D, Ruan C, Korpeoglu E, 等（2020）时序图上的归纳式表示学习。网址 https://arxiv.org/abs/2002.07962, arXiv:2002.07962Xu J, Chen Z, Wang W, 等（2025a）Cohesion：采用双阶段融合的多模态推荐复合图卷积网络。载于：国际 ACM SIGIR 信息检索研究与发展会议论文集，第1830–1839页。Xu J, Chen Z, Wang W, 等（2025b）Cohesion：采用双阶段融合进行多模态推荐的复合图卷积网络。网址 https://arxiv.org/abs/2504.04452, arXiv:2504.04452Xu J, Chen Z, Yang S, 等（2025c）Mentor：用于多模态推荐的多层次自监督学习。载于：AAAI人工智能会议论文集。Xu Q, Shen F, Liu L, 等（2018）Graphcar：利用图自编码器进行内容感知多媒体推荐。载于：国际 ACM SIGIR 信息检索研究与发展会议论文集，第981–984页。Yazijian E（2007年）从通俗表演到文学：kavikankana mukundarama cakravarti 的 candimangala。博士论文，芝加哥大学。Yi J, Chen Z（2022）用于推荐系统的多模态变分图自编码器。IEEE 多媒体汇刊 24:1067–1079。Yi Z, Wang X, Ounis I, 等（2022）用于微视频推荐的多模态图对比学习。载于：国际 ACM SIGIR 信息检索研究与发展会议论文集，第1807–1811页。Yilma BA, Leiva LA（2023）视觉艺术推荐的要素：绘画潜在语义表示的学习。载于：CHI 会议论文集

Ying R、He R、Chen K 等（2018）用于大规模 Web 推荐系统的图卷积神经网络。载于：ACM SIGKDD 知识发现与数据挖掘国际会议论文集，第 974–983 页。Yu P、Tan Z、Lu G 等（2025）关注个体信息！用于多媒体推荐的主图学习。载于：AAAI人工智能会议论文集，第 13096–13105 页。Yu W、Qin Z（2020）基于低通协同滤波器的推荐图卷积网络。载于：国际机器学习会议论文集。JMLR.org

Zhang J、Zhu Y、Liu Q 等（2021a）挖掘用于多媒体推荐的潜在结构。载于：第 29 届 ACM国际多媒体会议论文集，第 3872–3880 页

Zhang J、Zhu Y、Liu Q 等（2021b）挖掘用于多媒体推荐的潜在结构。载于：第 29 届 ACM国际多媒体会议。ACM，MM '21，第 3872–3880 页，https://doi.org/10.1145/3474085.3475259，网址http://dx.doi.org/10.1145/3474085.3475259

Zhang J, Zhu Y, LiuQ 等（2023）基于对比模态融合的多媒体推荐潜在结构挖掘。IEEE Transactions on Knowledge and Data Engineering 35(9):9154–9167。https://doi.org/10.1109/TKDE.2022.3221949

Zhang L, Liu Y, Zhou X, 等（2022）基于扩散的图对比学习：用于隐式反馈推荐。载于：高级应用数据库系统国际会议，第 232–247 页。Zhang M, Chen Y（2018）基于图神经网络的链接预测。载于：神经信息处理系统国际会议论文集，第 5171–5181 页。Zhang P, Zeng G, Wang T, 等（2024）Tinyllama：一种开源小语言模型。arXiv:2401.02385Zhang X, Zitnik M（2020）GNNGuard：防御图神经网络免受对抗攻击。载于：神经信息处理系统进展，第 33 卷。Curran Associates, Inc.，第 9263–9275 页。Zhang Y, Zhang J, Cui Z, 等（2021c）用于即席检索的基于图的相关性匹配模型。载于：AAAI人工智能会议。Zhao T, Zhang X, Wang S（2021）GraphSMOTE：基于图神经网络的图不平衡节点分类。载于：ACM 国际

网络搜索与数据挖掘会议，第 833–841 页

Zheng Y, Gao C, He X 等（2023）将价格纳入基于图卷积网络的推荐。IEEE《知识与数据工程汇刊》35(2):1609–1622

Zhi-Han Y（2022）使用自编码变分贝叶斯训练潜变量模型：教程。Zhou H, Zhou X, Zeng Z 等（2023a）多模态推荐系统综合综述：分类体系、评估与未来方向。arXiv abs/2302.04473

Zhou H, Zhou X, Zhang L 等（2023b）利用同质图增强多模态推荐中的二元关系。载于：ECAI，第3123–3130页

Zhou X, Miao C（2024）具有可解释性的解耦图变分自编码器，用于多模态推荐。URL https://arxiv.org/abs/2402.16110，arXiv:2402.16110Zhou X, Shen Z（2023）两个图的故事：为多模态推荐冻结并去噪图结构。载于：第31届ACM国际多媒体会议论文集。ACM，MM '23，第935–943页，https://doi.org/10.1145/3581783.3611943，URL http://dx.doi.org/10.1145/3581783.3611943Zhou X, LinD, Liu Y 等（2023c）用于推荐的层级细化图卷积网络。载于：国际数据工程会议论文集，ICDE 2023，第1247–1259页 Zhu Y, Xu Y, Yu F 等（2021）采用自适应增强的图对比学习。载于：万维网会议论文集，第2069–2080页 Zhuang Z, Du H, Han H 等（2025）弥合差距：教师辅助的瓦瑟斯坦知识蒸馏，实现高效的多模态推荐。载于：ACM 网络会议论文集，第2464–2475页