CLEAR：多模态推荐中的跨模态去冗余：零空间投影

Hao Zhan*合肥工业大学 2023213432 @mail.hfut.edu.cn

Yihui Wang*合肥工业大学 2023213406 @mail.hfut.edu.cn

Yonghui Yang†新加坡国立大学 yh_yang@nus.edu.sg

Danyang Yue 华中科技大学 danyangy@hust.edu.cn

Yu Wang 合肥工业大学 w angyu20001162@gmail.com

Pengyang Shao 新加坡国立大学 shaopymark@gmail.com

Fei Shen 新加坡国立大学 shenfei29@nus.edu.sg

Fei Liu 合肥工业大学 feiliu@mail.hfut.edu.cn

Le Wu 合肥工业大学 lewu.ustc@gmail.com

arXiv:2603.01536v1 [cs.IR] 2 Mar 2026

摘要

多模态推荐已成为一种有效范式，通过融入异构内容模态来增强协同过滤。现有多模态推荐器主要致力于强化跨模态一致性，以促进多模态融合。然而，我们发现，多模态表示往往存在显著的跨模态冗余，即不同模态间占主导地位的共享成分存在重叠。这种冗余会限制互补信息的有效利用，也解释了为何引入更多模态并不总能提升性能。在本研究中，我们提出 CLEAR，这是一种轻量、即插即用的多模态推荐跨模态去冗余方法。CLEAR 不再强制加强跨模态对齐，而是通过建模视觉与文本表示之间的跨模态协方差，显式刻画不同模态间的冗余共享子空间。CLEAR 借助奇异值分解识别占主导地位的共享方向，并将多模态特征投影到互补零空间，从而重塑多模态表示空间，在抑制冗余跨模态成分的同时保留模态特定信息。这种子空间层面的投影能够隐式调节表示学习的动态过程，防止模型在训练期间反复放大冗余的共享语义。值得注意的是，CLEAR 无需修改现有多模态推荐器的架构或训练目标，即可无缝集成其中。在三个公开基准数据集上的大量实验表明，显式减少跨模态冗余能够持续提升各种多模态推荐模型的推荐性能。

CCS 概念

• **信息系统 → 推荐系统**。

关键词

跨模态冗余、多模态推荐、零空间投影

1 引言

多模态推荐已成为一种广泛采用的范式，通过整合来自多个模态的辅助信息来增强协同过滤 [8, 16]。多模态推荐器利用用户–物品交互数据之外的异质物品内容，旨在学习更丰富、更具表现力的物品表示。现有的多模态推荐方法大致可分为两类：利用从单个模态中提取的语义特征来增强物品表示 [3, 8]；以及利用由模态特定相似度诱导的潜在物品–物品结构 [38]。在这些建模范式的基础上，越来越多的近期研究进一步引入对齐驱动目标来改善多模态融合，通常通过在共享潜在空间中强制实现跨模态一致性 [15, 39]。

尽管现有多模态推荐模型效果显著，但它们往往默认不同模态能够提供互补信息。然而在实践中，辅助模态可能存在大量语义重叠，从而导致显著的跨模态冗余 [27, 37]。这种冗余不仅会削弱引入额外模态所带来的边际增益，还会引入噪声，并在不加区分地融合冗余信号时增加过拟合风险。此外，对齐驱动目标虽然有助于促进跨模态一致性，却可能因促使不同模态趋于相似的潜在表示而无意中加剧这一问题

*这些作者对本作品的贡献相同。
†通讯作者。

允许免费制作本作品的全部或部分内容的数字副本或纸质副本，用于个人或课堂用途，但前提是这些副本不得用于营利或商业利益，不得为此制作或分发，并且副本首页须附有本声明和完整引文。对于本作品中由作者以外的其他人拥有版权的内容，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、重新出版、发布到服务器或分发到列表，必须事先获得特定许可和/或支付费用。请通过 permissions@acm.org 申请许可。*Conference'17*，美国华盛顿哥伦比亚特区© 2026 版权所有者/作者(s)保留版权。出版权已授权给 ACM。ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM https://doi.org/10.1145/nnnnnnn.nnnnnnn

嵌入 [2, 34]。因此，学得的多模态表征可能无法充分利用模态特有的特征，从而限制多模态推荐的全部潜力。

为更好地理解跨模态冗余的本质，我们开展了简单的实证分析，以说明跨模态冗余在真实世界推荐数据中的普遍性。如图1所示，给定一个锚点物品，我们分别基于预训练的视觉特征和文本特征检索相似物品。我们发现，两种模态下检索物品的相似度分数分布存在显著重叠，这表明基于视觉和文本的检索会在表示空间中形成高度相似的物品邻域。此外，视觉模态与文本模态检索出的前K个物品之间的重叠率始终比随机期望高出一个数量级。这些观察结果表明，即使在联合建模之前，不同模态也可能产生高度冗余的物品间结构。在实践中，这种冗余往往表现为多模态推荐器的性能饱和：在已有信息丰富的表征基础上，仅仅引入额外的内容模态并不一定能提升推荐效果 [32, 33]。因此，简单地聚合多种模态并不足够；关键挑战在于有效利用互补的模态特定信息，同时抑制冗余的共享语义。

为应对冗余挑战，我们提出 **C**ross-moda**L**de-r**E**dundant subsp**A**ce p**R**ojection（CLEAR），这是一种轻量、即插即用的框架，用于抑制多模态推荐中的跨模态冗余。*CLEAR* 的核心思想是将跨模态冗余视为一个低维共享子空间：当视觉和文本模态编码了重叠的语义时，它们的表征往往会沿着少数几个主导方向相关联，而这些方向在训练过程中会被反复强化，从而削弱模态特定信号。为将这一思想付诸实践，CLEAR 首先显式建模视觉表示与文本表示之间的跨模态协方差，然后通过奇异值分解提取主导共享方向。CLEAR 不会强制加强对齐，也不会引入辅助目标，而是通过将多模态表征投影到与这些共享方向正交的零空间，实施子空间级干预，从而防止冗余分量在训练过程中被强化。这种投影会重塑多模态表示空间，在保留互补信息的同时，减轻共享语义的放大效应。

CLEAR 进一步提供了两个直观的控制因素，以平衡冗余抑制与信息保留：冗余秩 k，用于指定被视为冗余的主导共享方向数量；以及投影强度 λ，用于控制抑制程度。通过在表征层对冗余进行轻量级控制，CLEAR 在训练期间对多模态表示施加零空间投影，无需进行架构修改或添加辅助损失项。得益于其简洁性与通用性，CLEAR 可作为即插即用模块，无缝集成到现有的多模态推荐器中。在三个基准数据集上进行的大量实验表明，明确减少跨模态冗余能够持续提升推荐性能

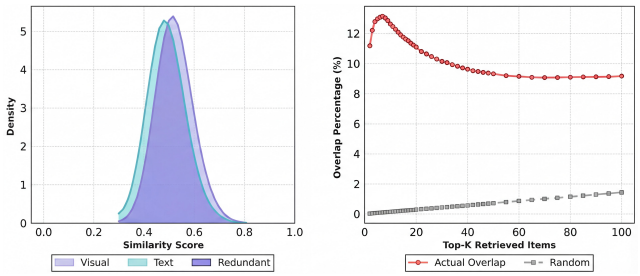


图1：跨模态冗余示意图。左：由视觉和文本模态检索到的物品的相似度分数密度分布。右：不同模态下前K个检索物品之间的重叠比率。

在各种多模态推荐架构中均如此。我们的贡献总结如下：

- 我们重新审视多模态推荐，指出其有效性取决于对不同模态进行有原则的冗余控制，而非增加建模容量或强制加强跨模态对齐。
- 我们提出 CLEAR，一种零空间投影框架，通过对多模态特征进行投影，在表征层面进行干预以抑制跨模态冗余，从而在保留模态特定信息的同时实现可控的冗余抑制。
- 我们在三个基准数据集上进行了大量实验，结果表明，使用 CLEAR 显式减少跨模态冗余，能够在各种多模态推荐架构中持续提升推荐性能。

2 预备知识

2.1 问题定义

在多模态推荐中，我们旨在结合用户-物品交互与物品多模态特征，提升偏好预测。形式化地，设 $\mathrm{U} = \{u_1, u_2, \ldots, u_M\}$ 和 $\mathrm{I} = \{i_1, i_2, \ldots, i_N\}$ 分别表示用户集合和物品集合。观测到的交互构成二值矩阵 $\mathrm{R} \in \{0,1\}^{M \times N}$，其中 $r_{ui} = 1$ 表示用户 u 与物品 i 之间存在交互。每个物品 i 都关联原始视觉特征 $\mathbf{v}_i^{\text{raw}} \in \mathbb{R}^{d_v}$ 和文本特征 $\mathbf{t}_i^{\text{raw}} \in \mathbb{R}^{d_t}$，分别组成矩阵 $\mathbf{V}^{\text{raw}} \in \mathbb{R}^{N \times d_v}$ 和 $\mathbf{T}^{\text{raw}} \in \mathbb{R}^{N \times d_t}$。我们的任务是学习表征 $\mathbf{e}_u$ 和 $\mathbf{e}_i$，使评分函数 $f_\Theta(\mathbf{e}_u, \mathbf{e}_i)$ 能够准确预测 $\hat{y}_{ui}$，其中 Θ 表示模型参数。本文聚焦于通过沿着前-k 个主导相关方向施加强度为 λ 的零空间投影，消除视觉特征与文本特征之间的跨模态冗余。

2.2 多模态推荐公式化

多模态推荐器通常遵循一条通用流程，将辅助内容特征融入协同过滤。原始多模态特征 $\mathbf{V}^{\text{raw}}$ 和 $\mathbf{T}^{\text{raw}}$ 首先通过特征编码器转换为稠密嵌入：

$$\mathbf{V} = \phi_v(\mathbf{V}^{\text{raw}}), \quad \mathbf{T} = \phi_t(\mathbf{T}^{\text{raw}}), \tag{1}$$

其中 ϕ_v 和 ϕ_t 将原始特征投影到共享嵌入空间 $\mathbb{R}^d$。随后，这些嵌入通过表征学习模块进行细化，例如基于用户-物品交互的图卷积网络、注意力机制或对比学习目标，以捕获协同信号和跨模态依赖关系。所得的模态特定表示通过加权聚合或拼接进行融合，形成统一物品嵌入$\mathbf{e}_i$。最后，通过评分函数预测用户偏好，并根据这些预测分数对物品进行排序以生成推荐结果。

尽管这一范式十分有效，但也面临一个关键挑战：视觉和文本模态通常会编码语义重叠，从而导致**跨模态冗余**，削弱模态特有信号，并使优化过程反复放大共享成分。这促使我们采用一种方法，对此类冗余进行显式建模并加以抑制，详见第 3 节。

3 方法

本节将详细分析 CLEAR 的关键机制和流程，包括跨模态方差矩阵的构建、奇异值分解以及零空间投影操作。图 2 展示了 CLEAR 的架构和训练流程。

3.1 CLEAR 概述

CLEAR 通过直接对学习得到的表征进行几何干预来处理跨模态冗余，而非依赖辅助损失。该方法首先通过跨模态协方差矩阵显式建模视觉特征与文本特征之间的统计依赖关系。随后，我们对该协方差矩阵进行奇异值分解，以确定冗余相关能量集中的主导方向。将这些主要奇异向量视为冗余的主轴后，CLEAR 会应用软零空间投影，衰减沿这些方向的分量，同时完整保留正交补中的信息，从而使模态特定信号更加突出，不易被共享语义掩盖。投影后的特征随后被无缝输入下游图神经网络，以捕获协同过滤信号。在下一节中，我们将详细介绍 CLEAR。

3.2 跨模态冗余建模

在前文中，我们将冗余定义为视觉和文本模态之间的语义重叠。具体而言，这种语义重叠在数学上表现为跨模态协方差：当两个模态编码相同概念时，它们相应的特征维度会在不同条目上呈现同步变化模式。遵循信息论中的既有定义 [24]，我们将跨模态冗余定义为与任务无关的共享信息。对于视觉特征$\mathbf{V}$ 和文本特征$\mathbf{T}$，总互信息可分解为：

$$I(\mathbf{V};\mathbf{T}) = \underbrace{I(\mathbf{V};\mathbf{T}|Y)}_{\text{task-relevant}} + \underbrace{[I(\mathbf{V};\mathbf{T}) - I(\mathbf{V};\mathbf{T}|Y)]}_{\text{redundancy } \mathcal{R}} \tag{2}$$

其中，Y表示推荐任务标签。冗余项 $\mathbf{R}(\mathbf{V},\ \mathbf{T}|Y)$量化了模态之间共享但与推荐性能无关的信息。

3.2.1 冗余及其干扰机制的特征分析。冗余与对齐之间的一个关键区别在于：对齐衡量跨模态对应关系，而冗余则反映系统性的重叠。这一区分揭示了冗余从三个方面损害推荐性能的原因：表征坍缩、融合失衡和优化病态。语义重叠主要编码模态通用信息，其判别能力有限，也难以捕获更具区分性的信号。在训练过程中，冗余信息往往会导致信息分配不均，使加权和注意力过度偏向重复的编码模式。具体而言，在随机梯度下降 [25], 模型中，模型通常优先拟合这些更易学习的重复模式，从而削弱学习更具判别力的模态特定信息的动力。如果不抑制冗余，多模态学习便会逐渐退化为对单模态信号的放大，进一步加剧对重复信息的过拟合。我们在附录A.2中为上述三点提供了严格的理论证明，以说明解决冗余问题的必要性。

3.2.2 跨模态协方差构建。为量化视觉特征与文本特征之间的相关性结构，我们构建跨模态协方差矩阵。给定经多层感知机变换后得到的特征矩阵$\mathbf{V} \in \mathbb{R}^{N\times d_v}$ 与 $\mathbf{T} \in \mathbb{R}^{N\times d_t}$，我们首先进行均值中心化，以消除全局偏移并聚焦于关系变化：

$$\bar{\mathbf{V}} = \mathbf{V} - \mathbf{1}_N \boldsymbol{\mu}_v^\top, \quad \bar{\mathbf{T}} = \mathbf{T} - \mathbf{1}_N \boldsymbol{\mu}_t^\top, \tag{3}$$

其中 $\bar{\mathbf{V}}$ 与 $\bar{\mathbf{T}}$分别表示均值中心化后的视觉特征矩阵与文本特征矩阵；$\mathbf{1}_N$为全一列向量；$\boldsymbol{\mu}_v = \frac{1}{N}\sum_{i=1}^{N} \mathbf{v}_i$ 与 $\boldsymbol{\mu}_t = \frac{1}{N}\sum_{i=1}^{N} \mathbf{t}_i$为按列均值。均值中心化确保协方差捕捉的是真正的共同变化模式，而非不同模态间幅度差异或偏移所带来的偏差。跨模态协方差矩阵随后定义为：

$$\mathbf{C} = \frac{1}{N}\bar{\mathbf{V}}^\top \bar{\mathbf{T}} \in \mathbb{R}^{d_v \times d_t}. \tag{4}$$

该矩阵 $\mathbf{C}$由统计学中的经验协方差估计量 [19, 20], 经 $1/N$缩放得到，从而对视觉特征与文本特征之间的总体协方差给出无偏估计。元素 $c_{ij} = \frac{1}{N}\sum_{n=1}^{N} \bar{v}_{ni}\bar{t}_{nj}$ 表示在全部 N个项目上，第i个视觉维度与第 j个文本维度之间的协方差。$\mathbf{C}$的每一行可解释为某一视觉维度（例如颜色通道）与所有文本维度（例如词嵌入分量）之间的相关性强度。$\mathbf{C}$ 中绝对值较大表明存在强烈的共同变化，预示潜在冗余，即一个模态的变化能够系统地预测另一模态的变化。

3.2.3 通过奇异值分解定位冗余。为提取主导相关性的结构，我们对跨模态协方差矩阵进行奇异值分解（SVD）：

$$\mathbf{C} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{W}^\top = \sum_{i=1}^{d} \sigma_i \mathbf{u}_i \mathbf{w}_i^\top \tag{5}$$

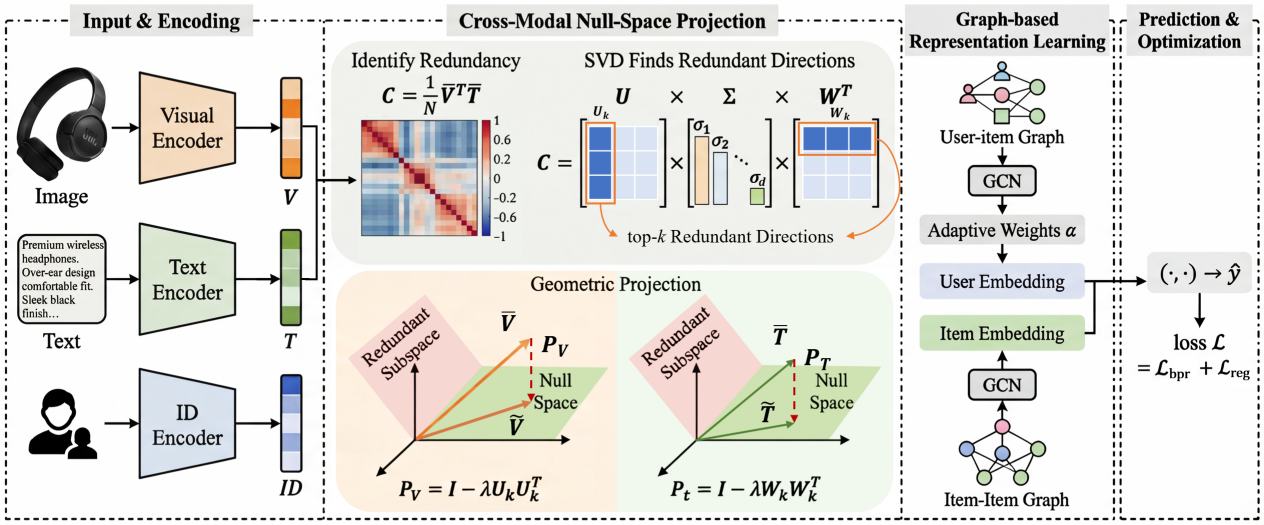


图2：CLEAR的整体架构。视觉特征和文本特征经过编码后，构建跨模态协方差矩阵C，并通过奇异值分解进行分解，以识别前 k 个冗余方向。零空间投影$\mathbf{P}_V$和$\mathbf{P}_T$在抑制冗余的同时保留模态特定信息，用于下游基于图的推荐。

其中，$\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_d] \in \mathbb{R}^{d \times d}$包含对应于视觉特征空间的左奇异向量，而$\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_d] \in \mathbb{R}^{d \times d}$包含对应于文本特征空间的右奇异向量。这一对应关系直接源于跨模态协方差矩阵$\mathbf{C} = \frac{1}{N}\bar{\mathbf{V}}^\top \bar{\mathbf{T}}$的构造，其中行对应视觉维度，列对应文本维度。因此，$\mathbf{U}$张成视觉子空间，而$\mathbf{W}$张成文本子空间。对角矩阵$\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_d)$中的奇异值按降序排列，其中$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d \geq 0$。每个奇异值$\sigma_i$衡量对应奇异向量对所捕获的跨模态相关性的强度，其大小直接反映该方向上的冗余能量。

对跨模态协方差矩阵进行奇异值分解，本质上是在所有可能的线性方向对中，寻找一个能使视觉与文本之间的协同变化最大化的正交基。奇异向量对并非人为定义的语义轴，而是在全局统计意义上从协方差结构中自动涌现。相应的奇异值σ_i量化了该方向上的跨模态相关性强度。需要注意的是，奇异值分解仅在线性意义上提取最优的协同方向。在真实的推荐数据中，特征空间中的不同语义因子往往并非严格正交，而是高度纠缠、彼此重叠。因此，每个奇异方向通常并不对应单一、纯粹的语义概念，而更像是多个有效信息成分与无效噪声的线性混合。我们承认，分解操作可能会影响任务相关信息，但仍可通过理论 A.4 证明，冗余抑制依然能够带来性能提升。

3.3 用于去冗余的零空间投影

识别出冗余后，我们接下来将说明 CLEAR 如何在训练过程中抑制冗余。关键在于，一旦识别出冗余方向，它们往往会在训练过程中被反复强化，因为它们主导着各模态的表征。CLEAR 在表征层面进行干预，将嵌入投影到与冗余方向正交的子空间上，从而避免共享语义成分被放大。形式上，给定一个由矩阵 $\mathbf{M} \in \mathbb{R}^{d \times k}$ 的列张成、且具有正交归一列的子空间 S，当且仅当向量$\mathbf{v} \in \mathbb{R}^d$满足以下条件时，它才位于 S 的零空间中

$$\mathbf{M}^\top \mathbf{v} = \mathbf{0}. \tag{6}$$

相应的到该零空间上的正交投影算子定义为

$$\mathbf{P}^\perp = \mathbf{I} - \mathbf{M}\mathbf{M}^\top, \tag{7}$$

其中$\mathbf{I}$表示单位矩阵。从几何上看，将$\mathbf{P}^\perp$应用于向量会去除其沿$\mathbf{M}$所张成方向上的分量，仅保留与该子空间正交的分量。定理5证明了零空间投影在消除冗余方面的有效性。

设$\mathbf{U}_k = [\mathbf{u}_1、\ldots\ldots、\mathbf{u}_k] \in \mathbb{R}^{d \times k}$和$\mathbf{W}_k = [\mathbf{w}_1、\ldots\ldots、\mathbf{w}_k] \in \mathbb{R}^{d \times k}$分别表示前$k$个左奇异向量和右奇异向量。投影到由$\mathbf{U}_k$张成的子空间上的正交投影算子为$\mathbf{U}_k\mathbf{U}_k^\top$。要投影到其零空间，我们使用：

$$\mathbf{P}_v^\perp = \mathbf{I} - \mathbf{U}_k\mathbf{U}_k^\top, \quad \mathbf{P}_t^\perp = \mathbf{I} - \mathbf{W}_k\mathbf{W}_k^\top \tag{8}$$

其中$\mathbf{I} \in \mathbb{R}^{d \times d}$是单位矩阵。

为灵活控制抑制强度并维持训练稳定性，我们引入带参数 $\lambda \in [0, 1]$ 的软投影：

$$\mathbf{P}_v = \mathbf{I} - \lambda\mathbf{U}_k\mathbf{U}_k^\top, \quad \mathbf{P}_t = \mathbf{I} - \lambda\mathbf{W}_k\mathbf{W}_k^\top \tag{9}$$

当 $\lambda = 0$ 时，特征保持不变；当 $\lambda = 1$ 时，则应用完整的零空间投影。介于两者之间的取值可实现部分抑制，从而在冗余降低与信息保留之间取得平衡。投影后的特征为：

$$\tilde{\mathbf{V}} = \bar{\mathbf{V}}\mathbf{P}_v, \quad \tilde{\mathbf{T}} = \bar{\mathbf{T}}\mathbf{P}_t \tag{10}$$

从几何角度看，该操作将沿前k个冗余方向的分量衰减至原来的（$1-\lambda$），同时保持正交分量不变。虽然奇异值分解无法将关键判别信号与冗余语义重叠完美分离，但选择 $\lambda < 1$ 可以让模型保留冗余方向上的部分信息，从而减轻与冗余纠缠的任务相关信号遭受的附带损害。

3.4 理论基础：低秩冗余假设

我们的方法基于一个核心理论假设：跨模态冗余集中在一个内在维度固定且依赖于数据的低维子空间中。这一视角从根本上将我们的方法与自适应策略区分开来，并为在整个训练过程中采用固定的冗余秩 k 提供了有力依据。

根据流形假设 [1, 5], , 高维多模态数据位于低维流形上或其附近。在推荐场景中，视觉模态和文本模态共享一个内在维度为 $d_s \ll$ 的共同语义流形 $\mathbf{M}_s$，其维度为 $\min(D_v,\ D_t)$，形式化定义为能够忠实近似该数据的最小 m，即数据可由一个 m 维流形忠实近似：

$$d_s = \min\left\{ m : \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{x}_i - \mathrm{proj}_{\mathcal{M}_m}(\mathbf{x}_i)\|^2 < \epsilon \right\}. \tag{11}$$

这一共享流形主要源自一组有限的、与模态无关的语义因子——其中最显著的是类别级概念，其数量 c通常远少于特征维度 d 和样本量 N。因此，跨模态协方差矩阵$\mathbf{C}$ 呈现出显著的低秩结构 [40]：

$$\mathrm{rank}(\mathbf{C}) \le c \ll \min(N, d). \tag{12}$$

值得注意的是，近期关于神经网络优化 [13]的研究表明，这种内在维度由底层数据分布决定，而非瞬态的训练动态；目标函数景观的有效自由度在整个学习过程中保持稳定。将这一结论应用于跨模态冗余，意味着冗余方向k^*的最优数量本质上是一个数据集常数，而不是一个应在优化过程中不断变化的量：

$$k^* = \arg\min_k \left\{ \sum_{i>k} \sigma_i^2(\mathbf{C}) + \lambda k \right\}. \tag{13}$$

因此，我们采用**固定秩投影**策略，在验证集上一次性选择 k，并在训练期间保持不变。此设计确保投影子空间稳定，使模型能够持续抑制相同的冗余方向，并可靠地保留正交补中的模态特定信息。相比之下，自适应秩选择（例如能量比阈值法）会引入不必要的不稳定性：奇异值会随小批次和优化步骤而波动，导致投影目标不断变化。此外，高能量分量可能源自模态特定噪声，而非真正的语义冗余 [22], ，因此，纯粹基于谱特征的自适应方法容易出错。

表 1：实验数据集的统计信息。

数据集	#用户	#物品	#Inters	稀疏性
Baby	19,445	7,050	160,792	99.88%
Sports	35,598	18,357	296,337	99.95%
服装	39,387	23,033	278,677	99.97%

3.5 优化与训练

我们的训练过程将零空间投影无缝整合到标准的推荐优化流程中，无需引入辅助损失函数。投影矩阵按训练轮次计算并缓存，以兼顾计算效率与特征分布适应性，而模型参数则通过小批次梯度下降进行更新。具体而言，在每个训练轮次开始时（或每经过 τ 个训练轮次），我们根据当前经 MLP 变换的特征更新投影矩阵$\mathbf{P}_v$ 和$\mathbf{P}_t$，并在该轮次的所有小批次中一致地应用这些矩阵。在前向传播过程中，投影后的特征$\mathbf{V} = \mathbf{V}\mathbf{P}_v$和$\mathbf{T} = \mathbf{T}\mathbf{P}_t$经过图卷积层，生成特定模态的嵌入；随后，通过可学习的模态特定注意力权重将这些嵌入融合，并通过物品间图聚合进一步细化，从而构建用于偏好预测的最终用户和物品表示。遵循隐式反馈推荐中的标准做法，我们采用贝叶斯个性化排序（BPR）损失 [23] 来优化成对偏好排序：

$$\mathcal{L}_{\mathrm{BPR}} = -\sum_{(u,i,j)\in O} \log \sigma(\hat{y}_{ui} - \hat{y}_{uj}) \tag{14}$$

其中，$O = \{(u, i, j) : r_{ui} = 1, r_{uj} = 0\}$ 表示由观测到的正向交互和采样得到的负样本物品构成的训练三元组集合，$\hat{y}_{ui} = \mathbf{e}_u, \mathbf{e}_i$ 表示预测的偏好得分，$\sigma(\cdot)$ 是 Sigmoid 函数。为防止过拟合，我们对用户偏好嵌入和融合权重施加 L_2 正则化：

$$\mathcal{L}_{\mathrm{reg}} = \|\mathbf{E}_u^v\|_2^2 + \|\mathbf{E}_u^t\|_2^2 + \|\mathbf{w}\|_2^2 \tag{15}$$

其中，$\mathbf{E}_u^v$ 和 $\mathbf{E}_u^t$ 分别表示融合前的视觉和文本用户偏好嵌入，$\mathbf{w}$ 表示可学习的模态融合权重。最终的训练目标结合了 BPR 损失和正则化：

$$\mathcal{L} = \mathcal{L}_{\mathrm{BPR}} + \gamma \mathcal{L}_{\mathrm{reg}} \tag{16}$$

其中，γ为正则化系数。关键在于，奇异值分解的计算基于已分离梯度的特征进行，以防止梯度流经投影矩阵的构建过程，从而确保$\mathbf{P}_v$ 和$\mathbf{P}_t$ 在每个训练轮次中都作为固定的几何约束，而非可学习参数。这一设计使模型能够学习到既有利于推荐任务的判别性表征，又兼容冗余抑制的表征。

4 实验

4.1 实验设置

本节中，我们进行了大量实验，以验证我们的假设，并展示方法的有效性、泛化能力和效率。实验结果有力地回答了以下关键问题。

- **研究问题1：**我们的 CLEAR 架构是否始终优于最先进方法?
- **研究问题2：**我们的设计原则在实践中是否成立?
- **RQ3：**超参数对性能有何影响?
- **RQ4：**CLEAR 的计算开销、普遍适用性和表征改进效果如何?

4.1.1 数据集。我们在亚马逊评论数据 [18]中的三个广泛使用的数据集上开展实证研究：**Baby**、**Sports**和**服装**，这些数据集包含充足的视觉和信息。表 1 总结了这些数据集的统计信息。遵循以往研究 [8, 30], ，我们采用 5-core过滤以确保数据质量。此外，我们使用预先提取的 4096 维视觉特征，并通过预训练的句子Transformer提取 384 维文本特征。附录 C.1 介绍了数据收集的详细信息。

4.1.2 基线方法。我们将 CLEAR 模型与十四种经典推荐模型进行了比较，其中包括三种传统的协同过滤模型：MF-BPR [23],、LightGCN [9], 和 LayerGCN [42]；以及十一种多模态推荐模型：VBPR [8], 、MMGCN [30], 、DualGNN [28], 、LATTICE [38], 、SLM-Rec [26], 、BM3 [44], 、MMSSL [29], 、FREEDOM [43], 、MGCN [35],、LGM-Rec [7], 和 MENTOR [31]。基线方法的完整列表及其实现细节见附录 C.2。

4.1.3 评估方案：为确保对我们方法进行公平评估，我们采用Recall@K和NDCG@K，其中K取 $K \in \{10, 20\}$。我们将数据随机划分为训练集、验证集和测试集，比例为8:1:1。

4.1.4 实现细节。我们借助MMRec [41]开源框架进行模型开发并复现各种基线方法。实验使用PyTorch，并在单张配备24GB内存的NVIDIA RTX 4090 GPU上完成。对于我们的模型，用户嵌入和物品嵌入的嵌入维度均设为64。所有训练参数均采用Xavier初始化 [6], ，训练过程中使用Adam优化器 [12] 。对于零空间投影模块，我们搜索投影强度

λ in $\{0.3, 0.5, 0.7, 0.9\}$，冗余秩 k in $\{2, 4, 8, 16, 20\}$。根据验证集上的表现选择最优超参数。

4.2 总体性能（RQ1）

表 2 展示了我们的方法与所有基线方法在三个数据集上的性能比较。我们有以下观察：

同时考虑模态融合与解耦的方法，优于仅关注对齐的方法 SLMRec、BM3 和 MMSSL 等对比学习方法通过拉近跨模态特征来学习多模态表示，其性能优于基础多模态架构。然而，与仅关注对齐的方法相比，明确平衡模态对齐与解耦的方法始终表现更优。例如，MGCN 采用双通道架构，明确分离共享表示与独有表示，因此在各种数据集上的表现均优于基于对齐的方法。

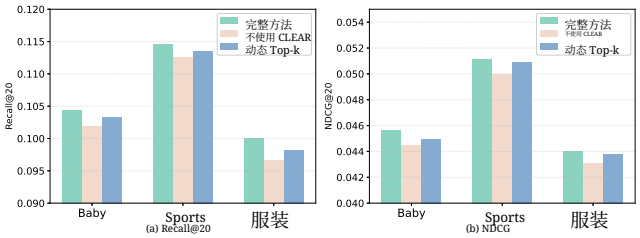


图 3：在三个数据集上对不同变体进行的消融研究。柱状图展示了以下方法的 R@20 和 N@20 指标：完整模型（采用零空间投影和固定的 top-k）、去除零空间，以及动态比例（基于自适应比例进行选择）。

我们的方法优于基于对齐和基于解耦的方法。 我们的方法同时融入了两项核心原则，同时超越了传统的权衡框架：它既不强制对齐，也不预设共享与特有部分的边界。相反，它通过数据驱动的协方差分析，直接识别并抑制冗余方向。在各个数据集上，以 R@20 指标为例，CLEAR 相较 MMSSL 提升了 7.6% 至 25.6%，相较 MGCN 提升了 3.6% 至 5.9%。本研究结果表明，精确定位并抑制冗余方向，比强化对齐或依赖预定义边界更为有效。

我们的方法在所有数据集上均实现了最先进性能，且提升稳定。与各数据集上表现最强的基线相比，我们的方法在 Recall@10 上分别提升了 6.0%（Baby）、2.7%（Sports）和 5.3%（服装），归一化折损累计增益指标也呈现出类似趋势。这三个数据集在规模和稀疏性方面各有特点。在这些不同场景中均取得稳定提升，表明跨模态冗余普遍存在，而现有方法尚未充分解决这一问题；无论数据集属性如何，我们的冗余抑制策略都具有良好的泛化能力。

4.3 消融研究（RQ2）

为验证关键设计选择，我们进行了消融实验，考察了三种配置：（1）使用固定 top-k 进行零空间投影的完整模型；（2）完全移除零空间投影；（3）将固定的 k 替换为动态比例选择，根据奇异值能量分布确定 k。图 3 总结了实验结果。

零空间投影的作用。如图 3 所示，移除零空间投影（w/o null-space）会导致所有数据集的性能持续下降。在 Recall@20 指标上，Baby、Sports 和服装数据集的降幅分别达到 2.4%、1.7% 和 3.4%；NDCG@20 也呈现出类似趋势，降幅为 2.0%–2.4%。这些

表 2：基线方法与我们框架在 Recall@K（R@K）和 NDCG@K（N@K）方面的性能比较。

数据集	Baby				Sports				服装			
方法	R@10	R@20	N@10	N@20	R@10	R@20	N@10	N@20	R@10	R@20	N@10	N@20
MF-BPR	0.0357	0.0575	0.0192	0.0249	0.0432	0.0653	0.0241	0.0298	0.0187	0.0279	0.0103	0.0126
LightGCN	0.0479	0.0754	0.0257	0.0328	0.0569	0.0864	0.0311	0.0387	0.0340	0.0526	0.0188	0.0236
LayerGCN	0.0529	0.0820	0.0281	0.0355	0.0594	0.0916	0.0323	0.0406	0.0371	0.0566	0.0200	0.0247
VBPR	0.0423	0.0663	0.0223	0.0284	0.0558	0.0856	0.0307	0.0384	0.0281	0.0415	0.0158	0.0192
MMGCN	0.0378	0.0615	0.0200	0.0261	0.0370	0.0605	0.0193	0.0254	0.0218	0.0345	0.0110	0.0142
DualGNN	0.0448	0.0716	0.0240	0.0309	0.0568	0.0859	0.0310	0.0385	0.0454	0.0683	0.0241	0.0299
LATTICE	0.0547	0.0850	0.0292	0.0370	0.0620	0.0953	0.0335	0.0421	0.0492	0.0733	0.0268	0.0330
SLMRec	0.0529	0.0775	0.0290	0.0353	0.0663	0.0990	0.0365	0.0450	0.0452	0.0675	0.0247	0.0303
BM3	0.0564	0.0883	0.0301	0.0383	0.0656	0.0980	0.0355	0.0438	0.0422	0.0621	0.0231	0.0281
MMSSL	0.0613	0.0971	0.0326	0.0420	0.0673	0.1013	0.0380	0.0474	0.0531	0.0797	0.0291	0.0359
FREEDOM	0.0627	0.0992	0.0330	0.0424	0.0717	0.1089	0.0385	0.0481	0.0629	0.0941	0.0341	0.0420
MGCN	0.0620	0.0964	0.0339	0.0427	0.0729	0.1106	0.0397	0.0496	0.0641	0.0945	0.0347	0.0428
LGMRec	0.0644	0.1002	0.0349	0.0440	0.0720	0.1068	0.0390	0.0480	0.0555	0.0828	0.0302	0.0371
MENTOR	0.0648	0.1035	0.0349	0.0448	0.0736	0.1130	0.0399	0.0501	0.0659	0.0976	0.0357	0.0437
CLEAR												

表3：Baby数据集上的泛化分析。

模型	R@10	R@20	N@10	N@20
VBPR	0.0423	0.0663	0.0223	0.0284
VBPR +CLEAR	0.0469	0.0776	0.0258	0.0336
提升	*10.87%*	*17.04%*	*15.70%*	*18.31%*
FREEDOM	0.0624	0.0984	0.0328	0.0420
FREEDOM + CLEAR	0.0634	0.0992	0.0337	0.0427
提升	*1.60%*	*0.81%*	*2.74%*	*1.67%*

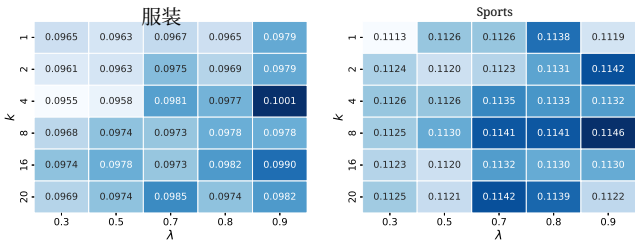


图 4：服装和体育数据集上的超参数分析。热力图展示了不同λ 和 k 组合下的 R@20 性能。颜色越深，结果越好。

结果验证了我们的核心假设：跨模态冗余会使模态通用因素掩盖模态特定判别信息，从而阻碍推荐性能。零空间投影能够有效抑制这些冗余方向，使模型更好地利用独特的贡献

各模态的贡献。

固定顶部-k 与动态比例选择的比较。 采用固定 k 的完整方法始终优于动态比例选择。其根本原因在于，动态选择将冗余秩视为依赖训练的量；随着奇异值谱演变，它会造成不断变化的优化目标。这会导致不稳定，使模型持续适应不断变化的投影子空间。相比之下，固定 k 在整个训练过程中提供恒定的正则化目标，使模型能够学习稳定表示。结果表明，跨模态冗余集中在一个低维子空间中，其维度是数据领域的内在属性，而非训练动态的暂时特征。

4.4 超参数分析（RQ3）

4.4.1 投影强度的影响 λ。 随着

λ λ = 增加，

0.9

两个数据集的表现均持续提升，并在处达到最佳性能。这验证了更强的冗余抑制通常有助于推荐，因为它能让模态特定语义占据主导地位。在较高 λ 值下表现稳定，说明我们的方法能够有效抑制模态通用因素，同时不去除与任务相关的信息。

4.4.2 冗余秩的影响 k。 热力图揭示了一种倒 U 型关系： k 过小时，无法充分抑制冗余；而 k 过大则可能移除有用信息。不同数据集的最优取值各不相同——服装数据集偏好 $k = 4$ ，而 Sports 更适合$k = 8$——这反映了领域特定冗余结构的差异。值得注意的是，在适中的 k 范围内，性能保持稳定，表明模型对超参数选择具有鲁棒性。

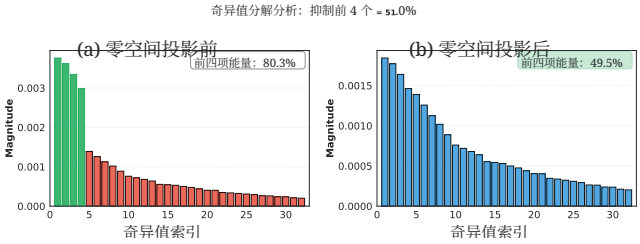


图 5：零空间投影前后跨模态协方差矩阵的奇异值分布。

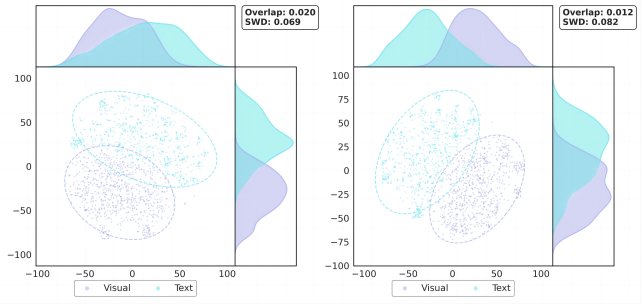


图6：Baby数据集上的t-SNE可视化。我们的零空间投影显著降低了重叠度，并大幅增大了切片Wasserstein距离（SWD）。

4.5 框架效率、泛化与可视化分析（RQ4）

4.5.1 计算效率。我们分析了零空间投影模块带来的计算开销。时间复杂度主要由奇异值分解 $O(d^3)$ 和特征投影 $O(Nd^2)$ 决定；由于 $d \ll N$，且投影矩阵会被缓存，因此这部分开销可以忽略不计。空间复杂度仅增加 $O(d^2)$，用于存储两个投影矩阵。算法1概述了我们方法的训练流程。表4展示了计算复杂度方面的详细实验结果。结果表明，我们的模型在训练效率和内存使用量方面均具有竞争力。

4.5.2 泛化分析。为了验证我们的零空间投影是一种适用于各种骨干网络的通用框架，我们将其集成到两个架构各异的代表性多模态推荐模型中。表 3 展示了在 Baby数据集上的结果。我们观察到，VBPR 和 FREEDOM 在经过零空间投影后均有不同程度的性能提升；由于 FREEDOM 本身具有强大的多图卷积结构，其性能提升幅度略低。

4.5.3 可视化分析。 为了直观理解零空间投影如何影响学习得到的表示，我们提供了可视化分析。**奇异值分布**。 图5显示，Baby数据集上的奇异值分布呈长尾形态；经过零空间投影后，主导奇异值被大幅抑制，能量分布明显变得更加平坦。**t-SNE [17] 可视化**。 图6展示了零空间投影前后的分布。上方和右侧显示了边缘分布。通过重叠度和切片Wasserstein距离指标，我们可以直观地观察到，CLEAR方法在表征层显著降低了冗余。定理A.1证明了这一论断。

5 相关工作

5.1 MMRec中的模型架构

模型层面的方法专注于设计复杂的架构，以捕捉用户、物品与多模态特征之间的高阶交互。早期研究主要利用图卷积网络（GCN）传播多模态信号。例如，MMGCN [30] 构建模态专属二部图，而 LATTICE [38] 学习潜在的物品-物品结构，以增强原始交互图。然而，近期研究开始质疑复杂图结构的必要性。FREEDOM [43]指出，原始多模态图往往包含噪声，并提出一种度引导去噪策略来提升鲁棒性。除了图卷积网络之外，TMLP [11] 还舍弃了开销较大的图卷积，转而采用轻量级的基于多层感知机的架构，证明只要恰当利用多模态信号，高效的特征变换就能胜过复杂的消息传递。尽管这些模型层面的方法在架构上有所创新，但它们往往将多模态特征视为静态输入，未能解决训练过程中固有的语义冗余问题。

5.2 MMRe中的优化策略 c

近年来，研究已显著转向优化层面的创新，重点关注目标函数与训练范式如何塑造表示空间。对比学习已成为主导范式。SLMRec [26]和BM3 [44] 等方法采用自监督目标，将模态特定表征与协同信号进行对齐。此外，该领域正朝着无ID或ID无关的推荐发展，以缓解冷启动问题与ID偏差 [36]。一些工作 [14] 主张完全去除ID嵌入，或使用基于softmax的对齐损失来替代传统的BPR损失，从而实现更好的跨模态收敛。为应对模态坍塌现象 [2],，已引入解耦技术。MGCN [35] 与双重对齐方法试图将模态共享偏好与模态特定偏好分离开来。然而，这些优化策略通常依赖于人工设计的辅助损失或复杂的正交约束，这往往导致对齐与可区分性之间的权衡。我们的工作有所不同，它引入几何视角，通过零空间投影消除冗余，同时不牺牲主任务的完整性。

5.3 机器学习中的零空间投影

零空间投影将向量投影到给定子空间的正交补上，已在包括公平性学习和表示学习在内的多种机器学习应用中展现出有效性。例如，在推荐领域，AlphaFuse [10] 通过奇异值分解（SVD）将语言嵌入空间分解为语义丰富的行空间和语义稀疏的零空间，并将协同信号注入零空间，在融入

行为模式的同时保留语义信息。在分类公平性方面，INLP [21] 通过迭代训练线性分类器来预测敏感属性，并将表征投影到分类器权重的零空间中，从而有效消除歧视性相关性，同时保留任务相关信息。在大语言模型的背景下，AlphaEdit [4] 在知识编辑过程中将参数扰动投影到保留的知识表征的零空间上，确保更新不会改变现有事实的输出。本文将零空间投影引入多模态推荐，以实现跨模态冗余消除。

6 结论

本文识别并形式化了多模态推荐中的跨模态冗余问题。为解决这一问题，我们提出了 CLEAR，通过零空间投影抑制冗余，促使视觉和文本模态学习互补而非重复的表征。我们的方法无需预先定义冗余信息或特定信息的边界，无需辅助损失即可通过纯几何投影实现冗余抑制，并且可作为即插即用模块应用于任何多模态框架。在三个数据集上的实验验证了其稳定的性能提升。我们的方法还可以应用于涉及两种以上模态的任务。除多模态推荐外，我们的方法也可用于跨模态检索，因为跨模态冗余同样会阻碍互补学习。未来的工作可以进一步研究更精确的跨模态信息构建与分解。

参考文献

[1] Bradley CA Brown、Anthony L Caterini、Brendan Leigh Ross、Jesse C Cresswell 和 Gabriel Loaiza-Ganem。2022年。验证图像数据的流形并集假设。*arXiv* 预印本 *arXiv:2207.02862*（2022年）。[2] Abhra Chaudhuri、Anjan Dutta、Tu Bui 和 Serban Georgescu。2025年。深入探究多模态表征坍塌。*arXiv* 预印本 *arXiv:2505.22483*（2025年）。[3] Jingyuan Chen、Hanwang Zhang、Xiangnan He、Liqiang Nie、Wei Liu 和 Tat- Seng Chua。2017年。注意力协同过滤：通过项目级和组件级注意力实现多媒体推荐。载于*SIGIR*。335–344。[4] Junfeng Fang、Houcheng Jiang、Kun Wang、Yunshan Ma、Shi Jie、Xiang Wang、Xiangnan He 和 Tat-Seng Chua。2024年。Alphaedit：面向语言模型的零空间约束知识编辑。*arXiv* 预印本 *arXiv:2410.02355*（2024年）。[5] Charles Fefferman、Sanjoy Mitter 和 Hariharan Narayanan。2016年。检验流形假设。《美国数学会杂志》29卷4期（2016年），983–1049。[6] Xavier Glorot 和 Yoshua Bengio。2010年。理解深度前馈神经网络训练的难点。载于第十三届人工智能与统计国际会议论文集。JMLR研讨会与会议论文集，249–256。[7] Zhiqiang Guo、Jianjun Li、Guohui Li、Chaoyang Wang、Si Shi 和 Bin Ruan。2024年。Lgmrec：面向多模态推荐的局部与全局图学习。载于*AAAI*人工智能会议论文集，第38卷。8454–8462。[8] Ruining He 和 Julian McAuley。2016年。VBPR：基于隐式反馈的视觉贝叶斯个性化排序。载于*AAAI*人工智能会议论文集，第30卷。[9] Xiangnan He、Kuan Deng、Xiang Wang、Yan Li、Yongdong Zhang 和 Meng Wang。2020年。Lightgcn：简化并增强用于推荐的图卷积网络。载于第*43*届国际*ACM SIGIR*信息检索研究与发展会议论文集。639–648。[10] Guoqing Hu、An Zhang、Shuo Liu、Zhibo Cai、Xun Yang 和 Xiang Wang。2025年。Alphafuse：在语言嵌入的零空间中学习用于序列推荐的ID嵌入。载于第*48*届国际*ACM SIGIR*信息检索研究与发展会议论文集。1614–1623。[11] Junjie Huang、Jiarui Qin、Yong Yu 和 Weinan Zhang。2025年。超越图卷积：采用拓扑感知多层感知机的多模态推荐。载于*AAAI*人工智能会议论文集，第39卷。11808–11816。[12] Diederik Kinga、Jimmy Ba Adam 等。2015年。一种随机优化方法。载于《国际学习表征会议》（ICLR），第5卷。加利福尼亚州；[13] Chunyuan Li、Heerad Farkhoor、Rosanne Liu 和 Jason Yosinski。2018年。衡量目标函数景观的内在维度。arXiv 预印本 arXiv:1804.08838（2018年）。[14] Guohao Li、Li Jing、Jia Wu、Xuefei Li、Kai Zhu 和 Yue He。2025年。从基于ID到无ID：重新思考ID在多模态协同过滤推荐中的有效性。arXiv 预印本 arXiv:2507.05715（2025年）。[15] Kang Liu、Feng Xue、Dan Guo、Peijie Sun、Shengsheng Qian 和 Richang Hong。2023年。用于多媒体推荐的多模态图对比学习。IEEE多媒体汇刊 25（2023年），9343–9355。[16] Qidong Liu、Jiaxi Hu、Yutian Xiao、Xiangyu Zhao、Jingtong Gao、Wanyu Wang、Qing Li 和 Jiliang Tang。2024年。多模态推荐系统：综述。《计算机调查》57卷2期（2024年），1–17。[17] Laurens van der Maaten 和 Geoffrey Hinton。2008年。使用 t-SNE 可视化数据。《机器学习研究杂志》9卷，11月（2008年），2579–2605。[18] Julian McAuley、Christopher Targett、Qinfeng Shi 和 Anton Van Den Hengel。2015年。关于风格与替代品的基于图像的推荐。载于 SIGIR。43–52。[19] Ryumei Nakada、Halil Ibrahim Gulluk、Zhun Deng、Wenlong Ji、James Zou 和 Linjun Zhang。2023年。理解多模态对比学习并纳入未配对数据。载于《国际人工智能与统计会议》。PMLR，4348–4380。[20] Changdae Oh、Hyesu Lim、Mijoo Kim、Dongyoon Han、Sangdoo Yun、Jaegul Choo、Alexander Hauptmann、Zhi-Qi Cheng 和 Kyungwoo Song。2024年。迈向视觉-语言模型的校准式鲁棒微调。《神经信息处理系统进展》37（2024年），12677–12707。[21] Shauli Ravfogel、Yanai Elazar、Hila Gonen、Michael Twiton 和 Yoav Goldberg。2020年。将其归零：通过迭代零空间投影保护受保护属性。arXiv 预印本 arXiv:2004.07667（2020年）。[22] Mohammad Zia Ur Rehman、Devraj Raghuvanshi、Umang Jain、Shubhi Bansal 和 Nagendra Kumar。2025年。一种结合跨模态关系与层次交互注意力的多模态多任务框架，用于语义理解。《信息融合》（2025年），103628。[23] Steffen Rendle、Christoph Freudenthaler、Zeno Gantner 和 Lars Schmidt-Thieme。2012年。BPR：基于隐式反馈的贝叶斯个性化排序。arXiv 预印本 arXiv:1205.2618（2012年）。[24] Fazlollah M Reza。1994年。《信息论导论》。Courier Corporation。[25] Sebastian Ruder。2016年。梯度下降优化算法概述。arXiv 预印本 arXiv:1609.04747（2016年）。[26] Zhulin Tao、Xiaohao Liu、Yewei Xia、Xiang Wang、Lifang Yang、Xianglin Huang 和 Tat-Seng Chua。2022年。用于多媒体推荐的自监督学习。IEEE多媒体汇刊 25（2022年），5107–5116。[27] Hui Wang、Jinghui Qin、Wushao Wen、Qingling Li、Shanshan Zhong 和 Zhongzhan Huang。2025年。用于多模态推荐的多模态表征解耦信息瓶颈。arXiv 预印本 arXiv:2509.20225（2025年）。[28] Qifan Wang、Yinwei Wei、Jianhua Yin、Jianlong Wu、Xuemeng Song 和 Liqiang Nie。2021年。DualGNN：用于多媒体推荐的双图神经网络。IEEE多媒体汇刊 25（2021年），1074–1084。[29] Wei Wei、Chao Huang、Lianghao Xia 和 Chuxu Zhang。2023年。用于推荐的多模态自监督学习。载于《2023年ACM万维网大会论文集》。790–800。[30] Yinwei Wei、Xiang Wang、Liqiang Nie、Xiangnan He、Richang Hong 和 Tat-Seng Chua。2019年。MMGCN：用于微视频个性化推荐的多模态图卷积网络。载于《第27届ACM国际多媒体会议论文集》。1437–1445。[31] Jinfeng Xu、Zheyu Chen、Shuo Yang、Jinze Li、Hewei Wang 和 Edith CH Ngai。2025年。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》第39卷。12908–12917。[32] Yonghui Yang、Le Wu、Zhuangzhuang He、Zhengwei Wu、Richang Hong 和 Meng Wang。2025年。少即是多：通过信息瓶颈去噪的多媒体推荐。arXiv 预印本 arXiv:2501.12175（2025年）。[33] Yu Ye、Junchen Fu、Yu Song、Kaiwen Zheng 和 Joemon J Jose。2025年。多模态嵌入是否真的有利于推荐？深入探究整体模态与单一模态。arXiv 预印本 arXiv:2508.07399（2025年）。[34] Lingjie Yi、Raphael Douady 和 Chao Chen。2025年。解读多模态对比学习中的模态差距：从收敛表征到成对对齐。arXiv 预印本 arXiv:2510.03268（2025年）。[35] Penghang Yu、Zhiyi Tan、Guanming Lu 和 Bing-Kun Bao。2023年。用于多媒体推荐的多视图图卷积网络。载于《第31届ACM国际多媒体会议论文集》。6576–6585。[36] Zheng Yuan、Fajie Yuan、Yu Song、Youhua Li、Junchen Fu、Fei Yang、Yunzhu Pan 和 Yongxin Ni。2023年。推荐系统的下一步该走向何方？重新审视基于ID与基于模态的推荐模型。载于《第46届ACM国际信息检索研究与发展大会论文集》。2639–2649。

[37] Jinghao Zhang、Qiang Liu、Shu Wu 和 Liang Wang。2023年。挖掘稳定偏好：面向多媒体推荐的自适应模态去相关。载于第*46*届*ACM*国际信息检索研究与发展大会。443–452。[38] Jinghao Zhang、Yanqiao Zhu、Qiang Liu、Shu Wu、Shuhui Wang 和 Liang Wang。2021年。挖掘多媒体推荐的潜在结构。载于 *MM*。3872–3880。[39] Jinghao Zhang、Yanqiao Zhu、Qiang Liu、Mengqi Zhang、Shu Wu 和 Liang Wang。2022年。通过对比模态融合挖掘多媒体推荐的潜在结构。*IEEE TKDE*（2022年）。[40] Xu Zheng、Yuanhuiyi Lyu 和 Lin Wang。2024年。学习适用于任意模态语义分割的模态无关表示。载于 欧洲

计算机视觉会议。Springer，146–165。[41] Xin Zhou。2023年。MMRec：简化多模态推荐。载于第*5*届*ACM*亚洲多媒体国际会议研讨会论文集。1–2。[42] Xin Zhou、Donghui Lin、Yong Liu 和 Chunyan Miao。2023年。用于推荐的层精炼图卷积网络。载于*2023*年*IEEE*第*39*届国际数据工程会议（*ICDE*）。IEEE，1247–1259。[43] Xin Zhou 和 Zhiqi Shen。2023年。两个图的故事：冻结和去噪多模态推荐中的图结构。载于第*31*届*ACM*国际多媒体会议论文集。935–943。[44] Xin Zhou、Hongyu Zhou、Yong Liu、Zhiwei Zeng、Chunyan Miao、Pengwei Wang、Yuan You 和 Feijun Jiang。2023年。用于多模态推荐的自举潜在表示。载于*2023*年*ACM*万维网大会论文集。845–854。

A 定理证明

A.1 切片Wasserstein距离（SWD）

SWD衡量分布之间的总体差异。给定两组 d维特征$\mathbf{V} = \{\mathbf{v}_i\}_{i=1}^N$ 和$\mathbf{T} = \{\mathbf{t}_i\}_{i=1}^N$，SWD通过随机的一维投影近似Wasserstein距离：

$$\mathrm{SWD}(\mathbf{V},\mathbf{T}) = \frac{1}{L}\sum_{l=1}^{L} W_1\left(\boldsymbol{\theta}_l^\top \mathbf{V}, \boldsymbol{\theta}_l^\top \mathbf{T}\right), \quad \boldsymbol{\theta}_l \in \mathbb{S}^{d-1} \tag{17}$$

其中，$\{\boldsymbol{\theta}_l\}_{l=1}^L$ 是从（$d-1$）维单位球面上均匀采样的随机单位向量，$W_1(\cdot,\cdot)$表示一维投影分布之间的 1-Wasserstein距离，通过对样本分位数排序计算。SWD 值越高，表明分布散度越大，这意味着每种模态保留了更多独特信息。

A.2 冗余损害推荐：三重分析

（i）表征坍缩。

定理 1（可判别性退化）。将特征分解为 $\mathbf{v} = \mathbf{v}_\parallel + \mathbf{v}_\perp$，其中 $\mathbf{v}_\parallel = \mathbf{U}_k\mathbf{U}_k^\top \mathbf{v}$ 位于冗余子空间中。对于来自不同类别的项目 $c_i \neq c_j$，由 *Fisher* 线性判别分析测得的类间分离满足：

$$FDR_\parallel := \frac{\|\mathbb{E}_{i\in c_1}[\mathbf{v}_{i,\parallel}] - \mathbb{E}_{j\in c_2}[\mathbf{v}_{j,\parallel}]\|^2}{Var_{i\in c_1}[\mathbf{v}_{i,\parallel}] + Var_{j\in c_2}[\mathbf{v}_{j,\parallel}]} \ll FDR_\perp \tag{18}$$

其中，$FDR_\perp$ 的计算方式与$\mathbf{v}_\perp$类似。

证明。冗余子空间捕获了跨模态共享的类别级语义。在 $\mathrm{span}(\mathbf{U}_k)$ 中，类内方差大于类间方差：

$$\mathrm{Var}_{\text{within}}[\mathbf{v}_\parallel] = \mathbb{E}_c[\mathrm{Var}_{i\in c}[\mathbf{U}_k^\top \mathbf{v}_i]] \approx \sigma_1^2 \tag{19}$$

$$\mathrm{Var}_{\text{between}}[\mathbf{v}_\parallel] = \mathrm{Var}_c[\mathbb{E}_{i\in c}[\mathbf{U}_k^\top \mathbf{v}_i]] = O(\sigma_k^2) \ll \sigma_1^2 \tag{20}$$

Thus $\mathrm{FDR}_\parallel = O(\sigma_k^2/\sigma_1^2) \to 0$ as spectral gap increases. □

（ii）融合失衡。

命题 2（融合中的冗余放大）。对于加权融合$\mathbf{h} = \alpha\mathbf{v} + \beta\mathbf{t}$，其中 α、$\beta > 0$，将$\mathbf{v} = \mathbf{v}_\parallel + \mathbf{v}_\perp$ 和$\mathbf{t} = \mathbf{t}_\parallel + \mathbf{t}_\perp$分解。期望能量分布为：

$$\mathbb{E}[\|\mathbf{h}\|^2] = \underbrace{(\alpha^2+\beta^2+2\alpha\beta\rho)\mathbb{E}[\|\mathbf{v}_\parallel\|^2]}_{\textit{redundant: amplified}} + \underbrace{\alpha^2\mathbb{E}[\|\mathbf{v}_\perp\|^2] + \beta^2\mathbb{E}[\|\mathbf{t}_\perp\|^2]}_{\textit{specific: preserved}} \tag{21}$$

其中 $\rho = \mathbb{E}[\langle \mathbf{v}_\parallel, \mathbf{t}_\parallel\rangle]/(\|\mathbf{v}_\parallel\|\|\mathbf{t}_\parallel\|) \approx 1$ 表示冗余分量。

推论 3（失衡比率）。冗余分量相对于特定分量的能量比为：

$$\frac{(\alpha+\beta)^2\mathbb{E}[\|\mathbf{v}_\parallel\|^2]}{\alpha^2\mathbb{E}[\|\mathbf{v}_\perp\|^2] + \beta^2\mathbb{E}[\|\mathbf{t}_\perp\|^2]} \geq 1 + \frac{2\alpha\beta}{\alpha^2+\beta^2} \geq 1 \tag{22}$$

仅当 $\alpha = 0$ 或 $\beta = 0$（单模态）时等号成立。

（iii）优化病态。

定理 4（冗余子空间上的快速收敛）。在梯度流 $\frac{d\mathbf{V}}{dt} = -\nabla_{\mathbf{V}}\mathcal{L}$ 其中 $\mathcal{L} = -\sum_{(u,\,i,\,j)}\log\sigma(y_{ui} - y_{uj})$，将梯度分解为：

$$\nabla_{\mathbf{V}}\mathcal{L} = \nabla_{\mathbf{V}_\parallel}\mathcal{L} + \nabla_{\mathbf{V}_\perp}\mathcal{L} \tag{23}$$

收敛速率满足：

$$\|\mathbf{V}_\parallel(t) - \mathbf{V}_\parallel^*\| \leq e^{-\mu_\parallel t}\|\mathbf{V}_\parallel(0) - \mathbf{V}_\parallel^*\| \tag{24}$$

$$\|\mathbf{V}_\perp(t) - \mathbf{V}_\perp^*\| \leq e^{-\mu_\perp t}\|\mathbf{V}_\perp(0) - \mathbf{V}_\perp^*\| \tag{25}$$

其中 $\mu_\parallel \gg \mu_\perp$，$\mu_\parallel \propto \sigma_1^2$和 $\mu_\perp \propto \sigma_{k+1}^2$。

证明。BPR 损失关于$\mathbf{V}$ 的海森矩阵为：

$$\mathbf{H} = \nabla_{\mathbf{V}}^2\mathcal{L} = \sum_{(u,i,j)} \sigma'(y_{ui} - y_{uj})\mathbf{e}_i\mathbf{e}_i^\top \tag{26}$$

投影到冗余子空间：$\mathbf{H}_\parallel = \mathbf{U}_k^\top \mathbf{H}\mathbf{U}_k$。由于冗余特征与协同信号高度相关：

$$\lambda_{\min}(\mathbf{H}_\parallel) \propto \|\mathbf{U}_k^\top \mathbf{E}\|^2 \propto \sigma_1^2 \tag{27}$$

而 $\lambda_{\min}(\mathbf{H}$ □)

$\propto \sigma_{k+1}^2 \ll \sigma_1^2$。根据线性收敛理论，$(\mathbf{H})$，从而得到该结论。

A.3 零空间投影可证明地减少冗余

定理 5（奇异值抑制）。应用投影矩阵$\mathbf{P}_v = \mathbf{I} - \lambda\mathbf{U}_k\mathbf{U}_k^\top$和 $\mathbf{P}_t = \mathbf{I} - \lambda\mathbf{W}_k\mathbf{W}_k^\top$后，新的协方差变为：

$$\tilde{\mathbf{C}} = \frac{1}{N}(\mathbf{V}\mathbf{P}_v)^\top(\mathbf{T}\mathbf{P}_t) = \sum_{i=1}^{d}\tilde{\sigma}_i\mathbf{u}_i\mathbf{w}_i^\top \tag{28}$$

其中：

$$\tilde{\sigma}_i = \begin{cases}(1-\lambda)^2\sigma_i & i \leq k\\ \sigma_i & i > k\end{cases} \tag{29}$$

证明。展开投影后的协方差：

$$\tilde{\mathbf{C}} = \mathbf{P}_v^\top\mathbf{C}\mathbf{P}_t \tag{30}$$

$$= (\mathbf{I} - \lambda\mathbf{U}_k\mathbf{U}_k^\top)\mathbf{U}\Sigma\mathbf{W}^\top(\mathbf{I} - \lambda\mathbf{W}_k\mathbf{W}_k^\top) \tag{31}$$

$$= \mathbf{U}\Sigma\mathbf{W}^\top - \lambda\mathbf{U}_k\mathbf{U}_k^\top\mathbf{U}\Sigma\mathbf{W}^\top - \lambda\mathbf{U}\Sigma\mathbf{W}^\top\mathbf{W}_k\mathbf{W}_k^\top \tag{32}$$

$$+ \lambda^2\mathbf{U}_k\mathbf{U}_k^\top\mathbf{U}\Sigma\mathbf{W}^\top\mathbf{W}_k\mathbf{W}_k^\top \tag{33}$$

将$\mathbf{U} = [\mathbf{U}_k, \mathbf{U}_{\bar{k}}]$ 分块，并利用正交归一性：

$$\mathbf{U}_k^\top\mathbf{U} = \begin{bmatrix}\mathbf{I}_k \\ \mathbf{0}\end{bmatrix}, \quad \mathbf{W}^\top\mathbf{W}_k = \begin{bmatrix}\mathbf{I}_k \\ \mathbf{0}\end{bmatrix} \tag{34}$$

代入：

$$\tilde{\mathbf{C}} = \mathbf{U}\begin{bmatrix}(1-\lambda)^2\Sigma_k & \mathbf{0} \\ \mathbf{0} & \Sigma_{\bar{k}}\end{bmatrix}\mathbf{W}^\top \tag{35}$$

$$= \sum_{i=1}^{k}(1-\lambda)^2\sigma_i\mathbf{u}_i\mathbf{w}_i^\top + \sum_{i=k+1}^{d}\sigma_i\mathbf{u}_i\mathbf{w}_i^\top \tag{36}$$

□

算法1： CLEAR 训练过程

输入： 交互 $\mathbf{R}$，原始特征 $\mathbf{V}^{\text{raw}}$、$\mathbf{T}^{\text{raw}}$，超参数 k, λ, γ，更新间隔 τ

输出： 优化后的模型参数 Θ

对于 训练轮次 $e = 1$ **到** 最大_训练轮次 **执行**

  如果 e 取模 $\tau = 0$ 或 $e = 1$ **则**

    $\mathbf{V} \leftarrow \phi_v(\mathbf{V}^{\text{raw}}), \mathbf{T} \leftarrow \phi_t(\mathbf{T}^{\text{raw}})$;

    $\bar{\mathbf{V}} \leftarrow \mathbf{V} - \frac{1}{N}\mathbf{1}\bar{\mathbf{V}}$, $\bar{\mathbf{T}} \leftarrow \mathbf{T} - \frac{1}{N}\mathbf{1}\bar{\mathbf{T}}$;

    $\mathbf{C} \leftarrow \frac{1}{N}\bar{\mathbf{V}}^{\top}\bar{\mathbf{T}}$;

    $\mathbf{U}, \Sigma, \mathbf{W}^{\top} \leftarrow$ 奇异值分解(detach(C));

    $\mathbf{P}_v \leftarrow \mathbf{I} - \lambda \mathbf{U}_{:,1:k}\mathbf{U}_{1:k,:}^{\top}$, $\mathbf{P}_t \leftarrow \mathbf{I} - \lambda \mathbf{W}_{1:k,:}^{\top}\mathbf{W}_{1:k,:}$;

  end

  对于每个小批次 (u, i^+, i^-) **执行**

    $\tilde{\mathbf{V}} \leftarrow \mathbf{V}\mathbf{P}_v$, $\tilde{\mathbf{T}} \leftarrow \mathbf{T}\mathbf{P}_t$;

    $\mathbf{H}_v, \mathbf{H}_t \leftarrow$ 图卷积网络($\bar{\mathbf{V}}, \bar{\mathbf{T}}, \mathbf{A}^{\text{drop}}$);

    $\mathbf{E}_u \leftarrow$ 融合($\mathbf{H}_v[:N_u], \mathbf{H}_t[:N_u], \mathbf{w}_u$);

    $\mathbf{E}_i \leftarrow \mathbf{H}_v[N_u:] + \mathbf{H}_t[N_u:]$ + 物品图($[\mathbf{H}_v, \mathbf{H}_t][N_u:], \mathbf{A}_{\text{mm}}$);

    $\hat{y}_{ui^+} \leftarrow \mathbf{E}_u, \mathbf{E}_{i^+}$, $\hat{y}_{ui^-} \leftarrow \mathbf{E}_u, \mathbf{E}_{i^-}$;

    $L \leftarrow -\log \sigma(\hat{y}_{ui^+} - \hat{y}_{ui^-}) + \gamma(\|\mathbf{H}_v[u]\|_2^2 + \|\mathbf{H}_t[u]\|_2^2 + \|\mathbf{w}_u\|_2^2)$;

    $\Theta \leftarrow \Theta - \eta \nabla_{\Theta} L$;

  end

end

A.4 关于任务无关冗余的可辨识性

假设1（语义层级假设）。我们假设多模态特征呈现层级结构：

$$\mathbf{v} = \mathbf{v}_{coarse} + \mathbf{v}_{fine} \tag{37}$$

$$\mathbf{t} = \mathbf{t}_{coarse} + \mathbf{t}_{fine} \tag{38}$$

其中：

• 粗粒度：$\mathbf{v}$粗粒度 $\approx \mathbf{t}$粗粒度（例如，产品类别、主导颜色） • 细粒度：$\mathbf{v}$细粒度 $\not\approx \mathbf{t}$细粒度（例如，纹理细节、风格描述符）此外，$\mathbb{E}[\|\mathbf{v}\text{粗粒度}\|^2] \gg \mathbb{E}[\|\mathbf{v}\text{细粒度}\|^2]$ 但判别能力满足：

$$FDR(\mathbf{v}_{fine}) \gg FDR(\mathbf{v}_{coarse}) \tag{39}$$

定理6（奇异值分解优先捕获粗粒度冗余）。在假设*1*下，将 $\mathbf{C} = \mathbf{C}_{coarse} + \mathbf{C}_{fine}$分解如下：

$$\mathbf{C}_{coarse} = \frac{1}{N}\bar{\mathbf{V}}_{coarse}^{\top}\bar{\mathbf{T}}_{coarse}, \quad \mathbf{C}_{fine} = \frac{1}{N}\bar{\mathbf{V}}_{fine}^{\top}\bar{\mathbf{T}}_{fine} \tag{40}$$

则最大奇异值满足：

$$\sigma_1(\mathbf{C}) \geq \sigma_1(\mathbf{C}_{coarse}) - \|\mathbf{C}_{fine}\|_2 \tag{41}$$

且当 k 足够小时：

$$span(\mathbf{U}_k) \approx span(\mathbf{U}_{coarse}) \tag{42}$$

其中，$\mathbf{U}_{coarse}$ 是$\mathbf{C}_{coarse}$的主奇异向量。

证明。根据奇异值的 Weyl不等式：

$$|\sigma_i(\mathbf{C}) - \sigma_i(\mathbf{C}_{\text{coarse}})| \leq \|\mathbf{C}_{\text{fine}}\|_2 \tag{43}$$

由于 $\|\mathbf{C}_{\text{coarse}}\|_F \gg \|\mathbf{C}_{\text{fine}}\|_F$（粗粒度成分占主导能量），因此有：

$$\sigma_1(\mathbf{C}_{\text{coarse}}) \gg \|\mathbf{C}_{\text{fine}}\|_2 \tag{44}$$

根据 Davis-Kahan 正弦定理，子空间距离满足：

$$\|\sin\Theta(\mathbf{U}_k, \mathbf{U}_{\text{coarse},k})\|_F \leq \frac{\|\mathbf{C}_{\text{fine}}\|_F}{\sigma_k(\mathbf{C}_{\text{coarse}}) - \sigma_{k+1}(\mathbf{C})} \tag{45}$$

当谱隙 $\sigma_k(\mathbf{C}_{\text{coarse}}) - \sigma_{k+1}(\mathbf{C})$ 较大且 $\|\mathbf{C}_{\text{fine}}\|_F$ 较小时，这些子空间几乎重合。□

推论7（粗粒度信息的任务无关性）。在推荐中，粗粒度类别信息的效用有限，原因如下：

$$P(\text{user } u \text{ likes item } i | category(i)) \approx P(\text{user } u \text{ likes item } i) \tag{46}$$

也就是说，仅知道类别几乎无法预测个性化偏好。细粒度模态特定属性（由 $\mathbf{v}_{\perp}$、$\mathbf{t}_{\perp}$ 表示）对于区分类别内的物品至关重要。

定理8（作为风险最小化的软投影）。定义权衡：

$$\min_{\lambda,k} \underbrace{\mathbb{E}[\mathcal{L}_{rec}(\tilde{\mathbf{V}}, \tilde{\mathbf{T}})]}_{task\ loss} + \beta \underbrace{\|\tilde{\mathbf{C}}\|_F^2}_{redundancy\ penalty} \tag{47}$$

软投影参数 $\lambda \in [0,1]$ 实现了一种连续松弛：

$$\lambda = 0: \ no\ suppression, \quad \lambda = 1: \ hard\ projection \tag{48}$$

通过改变 λ，我们可以沿着冗余消除与信息保留之间的帕累托前沿移动。

证明。将 $\tilde{\sigma}_i = (1-\lambda)^2\sigma_i$ 代入 $i \leq k$：

$$\|\tilde{\mathbf{C}}\|_F^2 = \sum_{i=1}^{k}(1-\lambda)^4\sigma_i^2 + \sum_{i>k}\sigma_i^2 \tag{49}$$

$$= (1-\lambda)^4\mathcal{R}_k + \text{const} \tag{50}$$

对 λ求导：

$$\frac{\partial}{\partial\lambda}\|\tilde{\mathbf{C}}\|_F^2 = -4(1-\lambda)^3\mathcal{R}_k < 0 \tag{51}$$

因此，增大 λ 会单调地减少冗余。通过验证，最优的λ^* 可在此与任务损失之间取得平衡。□

B 时间与空间复杂度

为了更好地研究时间与空间复杂度，我们在算法1中提供了 CLEAR 的伪代码。

C 实验设置详情

C.1 数据集

• **Baby**：该数据集包含与婴儿用品的用户交互，包括婴儿车、玩具和婴儿护理产品等物品。

• **Sports**：该数据集包含与运动与户外产品的用户交互，涵盖装备、衣服和配饰。

• **服装**：该数据集包含用户与服装、鞋和珠宝商品的交互，代表了一个时尚导向领域。

表4：效率分析。我们统计了每个模型每个训练轮次的内存占用和运行时间。

数据集	指标	通用协同过滤模型			多模态模型			
		BPR	LightGCN	VBPR	MMGCN	LATTICE	FREEDOM	CLEAR
Baby	内存 (GB)	1.59	1.69	1.89	2.69	4.53	2.13	3.11
	时间 (秒/训练轮次)	0.86	1.81	1.04	6.35	3.42	2.58	2.30
Sports	内存 (GB)	2.00	2.24	2.71	3.91	19.93	3.34	4.12
	时间（秒/训练轮次）	1.24	3.74	1.67	20.88	13.59	4.52	4.77
服装	内存（GB）	2.16	2.43	3.02	4.24	28.22	4.15	4.43
	时间（秒/训练轮次）	1.52	4.58	2.02	26.13	23.96	5.74	5.34

C.2 基线方法

为了全面评估所提出方法的有效性，我们将其与 13 种具有代表性的基线方法进行比较，这些方法可分为两组：

通用协同过滤方法。这些方法仅利用用户-物品交互，不使用多模态信息：

• **MF-BPR** [23]：一种经典的矩阵分解方法，通过优化贝叶斯个性化排序损失来实现隐式反馈推荐。

• **LightGCN** [9]：一种简化的图卷积网络，移除了特征变换和非线性激活，在协同过滤中实现了最先进性能。

• **LayerGCN** [42]：一种经过层级优化的图卷积网络模型，通过在信息传播过程中优化层表示来解决过平滑问题。

多模态推荐方法。这些方法整合视觉特征和文本特征，以提升推荐效果：

• **VBPR** [8]：一种开创性方法，通过贝叶斯个性化排序将视觉特征融入矩阵分解。

• **MMGCN** [30]：一种多模态图卷积网络，通过构建模态专属二部图来捕捉细粒度的用户偏好。

• **DualGNN** [28]：一种双图神经网络，同时利用用户-物品二部图和用户共现图来建模模态特定用户偏好。

• **LATTICE** [38]：一种从多模态特征中挖掘潜在物品结构，并在学习得到的物品-物品图上执行图卷积的方法。

• **SLMRec** [26]：一种自监督学习框架，通过设计特征丢弃和掩码增强来揭示隐藏的多模态模式。

• **BM3** [44]：一种简化的自监督方法，通过基于丢弃的扰动为多模态推荐自举潜在表示。

• **MMSSL** [29]：一种多模态自监督学习方法，利用对抗扰动和跨模态对比学习来解耦共性与特定特征。

• **FREEDOM** [43]：一种冻结物品-物品图结构，并通过度敏感边丢弃实现多模态推荐去噪的方法。

• **MGCN** [35]：一种多视图图卷积网络，可净化模态特征，并设计行为感知融合器以自适应地学习模态的重要性。

• **LGMRec** [7]：一种利用超图神经网络深入提取各种模态信息的方法。

• **MENTOR** [31]：一种利用多层次自监督跨模态对齐来解决标签稀疏性问题和模态对齐问题的方法。