MealRec：基于层次扩散模型的微视频推荐多粒度序列建模

Xinxin Dong
d
ongxinxin@nudt.edu.cn
国防科技大学 中国长沙

Haokai Ma*
haokai.ma1997@gmail.com
新加坡国立大学 新加坡

Yuze Zheng
zhengyuzezzz
@gmail.com 国防科技大学
中国长沙

Yongfu Zha
zhayongfu@nudt.edu.cn
国防科技大学 中国长沙

Yonghui Yang
yyh.hfut@gmail.com 新加坡
国立大学 新加坡

Xiaodong Wang
xdwang@nudt.edu.cn
国防科技大学
技术
长沙，中国

arXiv:2603.01926v1 [cs.IR] 2 Mar 2026

摘要

微视频推荐旨在从用户交互过的微视频中的协同信息和上下文信息中捕捉用户偏好，从而预测合适的视频。然而，多模态内容中固有的噪声以及不可靠的隐式反馈常常阻碍这一目标，削弱行为与潜在兴趣之间的对应关系。尽管传统研究主要通过行为增强建模和以内容为中心的多模态分析来处理此类场景，但这些范式可能无意中引发两个不容忽视的挑战：提取与偏好无关的视频表示，以及固有的模态冲突。为解决这些问题，我们提出了一种基于层次扩散模型的微视频多粒度序列建模方法（***MealRec***），从视频内和视频间两个视角，在偏好建模过程中同时考虑时间相关性。具体而言，我们首先提出时间引导的内容扩散（TCD），在视频内时间信息的引导下，结合个性化协同信号来优化视频表示，突出重要内容并抑制冗余。为实现语义连贯的偏好建模，我们进一步设计了无噪声条件偏好去噪（NPD），通过盲去噪从受损状态中恢复有信息量的用户偏好。在来自两个平台的四个微视频数据集上开展的大量实验和分析，证明了我们的 *MealRec* 的有效性、普适性和鲁棒性，并进一步揭示了所提出的 TCD 和 NPD 的有效机制。源代码和相应的数据集将在论文录用后公开。

CCS 概念

• **信息系统** → **推荐系统**。

关键词

微视频推荐；序列推荐；多模态分析，扩散模型

ACM 参考文献格式：

Xinxin Dong、HaokaiMa、Yuze Zheng、Yongfu Zha、Yonghui Yang 和 Xiaodong Wang。2018。*MealRec*：面向微视频推荐的基于层次扩散模型的多粒度序列建模。收录于第*32*届知识发现与数据挖掘 *SIGKDD* 会议论文集（*ACM KDD '26*）。ACM，美国纽约州纽约市，10页。https://doi. org/XXXXXXX.XXXXXXX

1 引言

随着短视频平台的普及，微视频已成为在线内容消费的主要形式。有效的微视频推荐对于提升用户参与度和增强平台黏性至关重要。与传统推荐任务 [1–3], 不同，微视频推荐必须同时刻画每位用户交互视频序列中的动态协同依赖，以及每个视频帧序列内部的视觉演化，从而给学术界和工业界带来了额外挑战 [4–7]。

针对这些具体挑战，现有的视频推荐研究大致分为两种范式：行为增强建模和以内容为中心的多模态分析。前者主要利用用户的辅助行为信号（即观看时长、评论和分享），从协同视角细致地建模用户偏好。然而，这种范式严重依赖平台特定日志的可用性，因此难以迁移至更通用的微视频推荐场景；在这些场景中，只能获取隐式行为序列和视频内容。对于后者，先前的研究通常从模态视角建模用户偏好，例如通过模态特定偏好捕获 [8], 片段级兴趣建模 [9],，或采用参数高效微调策略适配多模态基础模型 [10]。然而，这些方法通常将语料库中的每个视频概括为静态内容表示，忽视帧之间的时间动态，因而无法

*通讯作者。

允许免费为个人或课堂使用而制作本作品全部或部分内容的数字副本或纸质副本，前提是这些副本不得用于营利或商业利益，也不得分发，并且副本须保留本声明及首页的完整引文。对于本作品中由作者以外的其他方所有的组成部分，必须尊重其版权。注明出处后可以摘录。若要以其他方式复制、再版、发布到服务器或分发至列表，则必须事先获得具体许可和/或支付费用。请向 permissions@acm.org 申请许可。*ACM KDD '26*，韩国济州岛© 2018 版权归所有者/作者所有。出版权已授权给 ACM。ACM ISBN 978-1-4503-XXXX-X/2018/06 https://doi.org/XXXXXXX.XXXXXXX

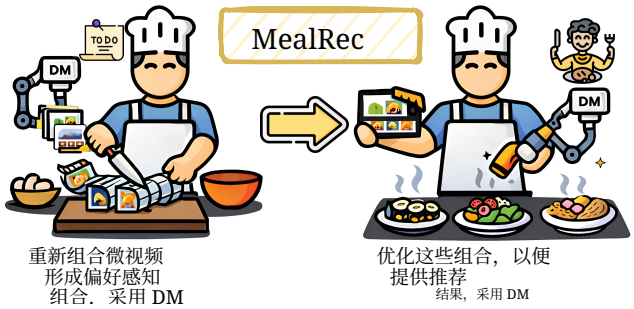

图1：拟议的*MealRec* 如同一位准备餐食的厨师：它首先利用“厨具”（扩散模型）依次切分并重新组合“食材”（微视频），形成“感知偏好的组合”（视频表征），然后再使用另一种“工具”（扩散模型）对这些“组合”进行专业精修，最终端上美味的“餐食”（推荐结果）。

有效地使内容与用户偏好相匹配，导致推荐效果不佳。

受扩散模型（DM）在时空动态建模和不确定性去噪方面的启发，许多研究者已探索其在视频理解 [11],多模态推荐 [12] 和序列推荐 [13] 任务中的有效性。顺应这一趋势，近期研究开始转向更具挑战性的多模态序列推荐任务。其中，DMMD4SR [12] 采用多层级扩散模型和不确定性引导融合，以缓解预训练多模态特征中的兴趣无关噪声；而M^{3}BSR [14] 则将迭代去噪与跨模态融合相结合，用于偏好建模。值得注意的是，扩散模型已被证明在视频理解方面行之有效：对时空表征进行迭代去噪，能够捕捉帧间依赖和动态视觉模式。这些成功启发了我们提出一种潜在的解决方案：我们能否联合融合扩散模型在多模态序列推荐和视频理解中的优势，以提升微视频推荐效果？

尽管这一做法合理，但仍面临不容忽视的挑战：偏好无关的视频表示提取以及固有的模态冲突。因此，我们提出*MealRec*，这是一个分层扩散框架，将微视频推荐构建为多粒度序列建模过程。如图1所示，借鉴厨师烹饪的过程，所提出的*MealRec* 引入了时间引导内容扩散（TCD）和无噪声条件偏好去噪（NPD），以重新构建微视频推荐。具体而言，我们首先在视频内层级提出TCD模块，将帧级时间引导和个性化协同信号同时融入微视频内容提取，从而缓解偏好无关的视频表示提取问题。受到近期研究发现的启发，即无需显式噪声调度即可从受损表示中推断扩散模型的噪声水平 [15],，我们在视频间层级提出NPD模块。在此，NPD通过盲去噪进行多模态融合，在视频和文本表示的引导下，直接从受损状态中恢复富含信息的偏好表示，而无需进行时间步条件预测 ϵ，从而无缝缓解固有的模态冲突。

我们在来自两个不同平台的四个微视频数据集上进行了大量实验，以证明*MealRec*的优越性。我们还开展了多项消融研究、可视化、噪声鲁棒性分析、骨干网络分析、重要参数分析和复杂度分析，以验证所提出的TCD和NPD的有效性。贡献总结如下：

•我们提出了*MealRec*，这是一种用于微视频推荐的分层扩散框架，可联合建模视频内时序动态和视频间偏好演化。据我们所知，我们是首个探索微视频推荐中多粒度序列建模问题的研究。

• 所提出的*MealRec* 将 TCD 与定制的 U-Net 估计器相结合，从视频内帧中细化偏好引导的视频表示；同时通过盲去噪，利用 NPD 根据视频间交互重建用户兴趣。

• 在四个真实世界数据集上进行的大量实验，证明了*MealRec* 的有效性、通用性、鲁棒性和轻量化特性。

2 相关工作

2.1 微视频推荐

微视频天然包含丰富的多模态信息，包括视觉帧、音轨和文本描述。如何有效利用这些异构模态，已成为微视频推荐中的核心挑战。MMGCN [8]率先将图卷积网络应用于多模态推荐，通过构建特定于模态的用户-物品交互图实现这一目标。M3CSR [16] 通过预训练编码器预处理多模态特征，并获取可训练的类别ID；MLLM-MSR [17] 则利用多模态大语言模型进行多模态物品摘要和用户偏好推断。然而，这些方法通常将视频视为视频级的原子单元，忽视了单个视频不同片段时刻发生的细粒度兴趣漂移。为解决这一局限，另一类研究侧重于建模单个视频内的细粒度用户兴趣。FRAME [5] 将视频分解为片段，并采用图卷积网络对用户-片段交互进行建模，从正向和负向两个角度捕捉不同时间片段中的兴趣变化。SINE [18]通过多任务学习，在片段级联合优化正向反馈和被动负反馈，以识别主要子兴趣，从而能够建模用户在视频中的注意力变化。SegRec [9] 融合异构模态，进一步推进了片段级偏好建模，以捕捉动态兴趣模式。尽管这些方法能够捕捉兴趣的时间动态，却未能充分利用视频中的丰富内容信息。相比之下，我们的方法通过在帧级建模用户偏好，弥合了这一差距；它将内容感知的多模态理解与细粒度兴趣建模相结合，从而实现更精准的偏好表示。

2.2 序列推荐

序列推荐通过捕捉交互历史中的时间依赖关系，对用户的动态偏好进行建模。早期方法依赖马尔可夫链 [19] 和循环神经网络 [20],，而基于 Transformer 的模型已成为主流范式

，采用自注意力机制。SASRec [21] 引入了用于序列建模的单向自注意力，而 BERT4Rec [22] 则采用双向注意力和掩码物品预测。后续研究进一步利用时间间隔 [23], 、个性化 [24], 和对比学习 [25]等因素改进序列建模。

近期研究利用多模态信息，在稀疏交互之外丰富物品表示。MMSR [26] 提出了自适应多模态融合，MP4SR [27] 则引入了采用对比目标的多模态预训练。除多模态融合外，基于大语言模型的推荐也因其在个性化、潜在推理和对话策略优化方面的潜力而受到关注 [28–30]。然而，现有的多模态序列推荐方法通常将多模态特征视为辅助信息，仅简单地将其与 ID 嵌入拼接或融合，而未能充分利用模态语义。此外，这些方法主要依赖粗粒度的物品级标签，而非细粒度内容线索，因此难以充分刻画用户对多模态内容的动态偏好；视频推荐领域的相关探索仍然有限。

2.3 用于推荐的扩散模型

扩散模型凭借其在建模复杂数据分布和去噪受损信号方面的卓越能力，近年来已成为推荐系统中一种强大的生成范式 [31–34]。DiffRec [35] 首次将去噪扩散概率模型应用于学习用户-物品交互的生成过程，展现出其在处理噪声交互方面优于传统生成模型的优势。后续研究将扩散模型拓展到了各种推荐场景。在协同过滤领域，先前工作利用高阶连接性和图结构信号探索扩散建模，以更好地捕捉用户-物品交互模式 [36, 37]。在序列推荐中，DiffuRec [38] 将物品表示为分布而非固定向量，以捕捉用户的多样化偏好；DiffuASR [39] 则利用扩散模型进行数据增强，以缓解数据稀疏性和长尾用户问题。PDRec [40] 提出了一种插件式扩散框架，通过时间间隔扩散模型推断用户的动态偏好。近期，扩散模型也被引入多模态序列推荐，以联合建模交互序列与多模态物品内容 [12]。尽管取得了这些进展，现有基于扩散模型的推荐方法通常对基于ID的交互或粗粒度物品级特征进行建模，未能充分利用细粒度多模态语义。尤其是，如何将视频内部的语义和时间线索有效融入序列推荐的扩散过程中，仍鲜有探索。

3 预备知识

3.1 扩散模型

扩散模型是一类生成模型，通过逆转逐步噪声污染过程来学习合成数据。本节简要介绍前向扩散过程、反向去噪过程和训练目标。

前向过程。 前向过程通过在 T时间步中逐渐添加高斯噪声来破坏数据。给定初始数据样本 $\mathbf{x}_0 \sim q(\mathbf{x}_0)$，前向过程定义为一个马尔可夫链：

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}), \tag{1}$$

其中 $\{\beta_t\}_{t=1}^T$ 是预先设定的方差调度。该过程的一个关键性质是，我们可以在任意时间步 t直接从$\mathbf{x}_0$ 以闭式形式采样$\mathbf{x}_t$：

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}), \tag{2}$$

其中 $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$。这使得我们能够通过$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon} \sim \mathbf{0}$,高效采样，其中 $\mathbf{N}(\mathbf{I})$。

反向过程。 反向过程旨在通过学习去噪估计器 $f_\theta(\cdot)$，从噪声中恢复原始数据。给定加噪输入$\mathbf{x}_t$ 和可选的条件信息$\mathbf{c}$，该估计器可以设计为预测原始数据$\mathbf{x}_0$ 或噪声分量 $\boldsymbol{\epsilon}$：

$$\hat{\mathbf{x}}_0 = f_\theta(\mathbf{x}_t, \mathbf{c}, t) = f_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \mathbf{c}, t), \tag{3}$$

其中$\mathbf{c}$表示任务特定的上下文引导。在推理过程中，通过从$\mathbf{x}_T \sim \mathbf{N}(\mathbf{0}, \mathbf{I})$开始迭代去噪来生成样本，以获得$\mathbf{x}_0$。

训练目标。根据预测目标的不同，扩散模型可以采用不同的目标进行训练。对于直接数据预测，我们最小化原始数据与估计数据之间的均方误差（MSE）：

$$\mathcal{L}_{\mathbf{x}_0} = \mathbb{E}_{\mathbf{x}_0,\boldsymbol{\epsilon},t}\left[\|\mathbf{x}_0 - f_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \mathbf{c}, t)\|^2\right]. \tag{4}$$

3.2 基础推荐器

由于用户偏好本质上会随时间变化，且观看行为按时间顺序记录，因此我们采用序列推荐器作为骨干网络。具体而言，我们使用 SASRec [21]作为基础编码器。它基于 Transformer 架构，并利用自注意力对用户行为序列中的复杂物品间转移进行建模。给定由物品嵌入与位置编码组合而成的输入嵌入矩阵 $\mathbf{D} \in \mathbb{R}^{L\times d}$，其中 L 为序列长度， d 为潜在维度，SASRec 通过多头自注意力计算上下文化表示：

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\right)\mathbf{V}, \tag{5}$$

其中，$\mathbf{Q} = \mathbf{D}\mathbf{W}^Q$、$\mathbf{K} = \mathbf{D}\mathbf{W}^K$和$\mathbf{V} = \mathbf{D}\mathbf{W}^V$分别表示通过可学习投影矩阵$\mathbf{W}^Q$、$\mathbf{W}^K$、$\mathbf{W}^V \in \mathbb{R}^{d\times d}$得到的查询、键和值矩阵。随后应用逐位置前馈网络，以增强表示能力。该模块生成 L 上下文表示 d维状态，每个状态都编码其对应位置的表示。

为展示我们框架的通用性，我们评估了三种具有代表性的序列推荐骨干网络：SAS- Rec [21], CL4SRec [25], 和 TedRec [41], ，涵盖自注意力、对比学习和多模态融合。这些结果验证了我们的方法不依赖于特定的基础编码器。详细分析见第5.4节。

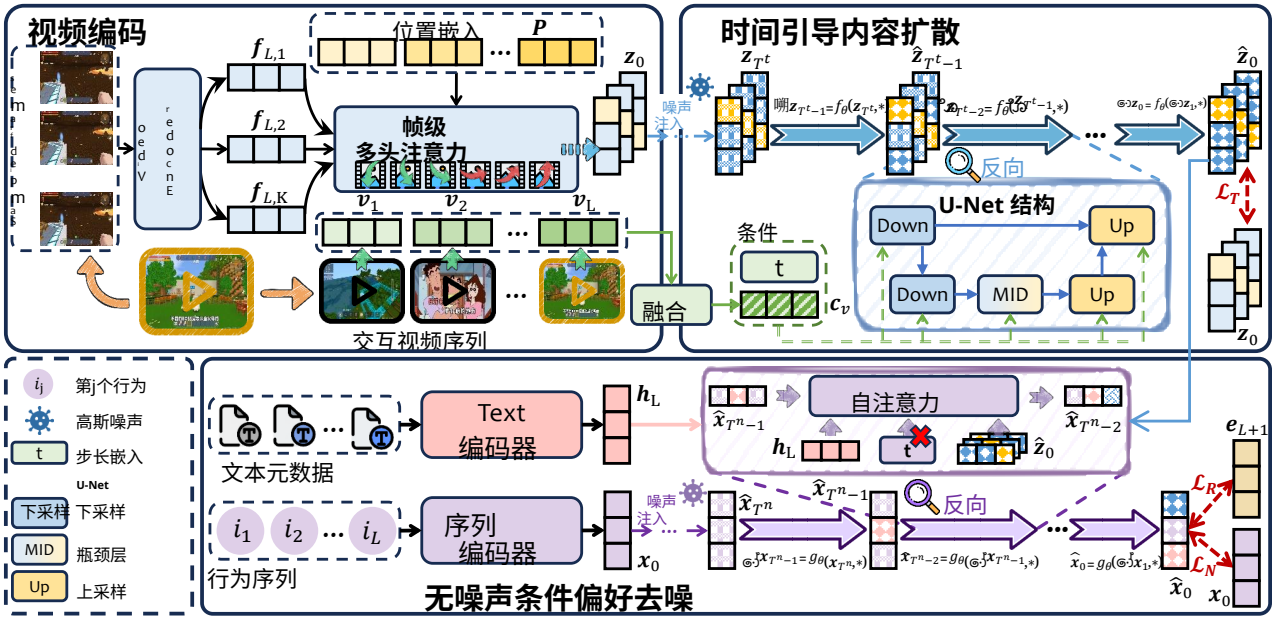

图2：我们提出的*MealRec*整体结构。

4 MealRec

4.1 问题定义

本文聚焦于微视频序列推荐任务。令 U 和 I 分别表示用户集合和物品集合。对于每个用户 $u \in$U，我们将其历史交互序列表示为 $S_u = [i_1, i_2, \ldots, i_L]$，其中 $i_k \in$I 表示第 k 个交互物品，L 表示序列长度。每个物品 i 都由一个多模态特征元组 $M_i = (\mathbf{v}_i, \mathbf{F}_i, \mathbf{h}_i)$ 表示，其中$\mathbf{v}_i \in \mathbb{R}^{d_v}$是概括视频整体内容的全局视觉嵌入，$\mathbf{F}_i = [\mathbf{f}_{i,1}, \mathbf{f}_{i,2}, \ldots, \mathbf{f}_{i,K}] \in \mathbb{R}^{K \times d_v}$是由 K 个采样帧级特征构成的序列，用于捕捉细粒度的时间动态；而$\mathbf{h}_i \in \mathbb{R}^{d_h}$是从标题和描述中提取的文本嵌入。给定用户的交互序列 S_u 及其相关多模态特征，我们的目标是预测最符合用户不断变化的偏好的下一个物品 i_{L+1}。

4.2 整体框架

为了解决微视频推荐中多个时间尺度上的噪声问题，我们提出了*MealRec*，一种用于序列建模的分层去噪框架。如图2所示，*Meal-Rec* 采用了由两个不同粒度组件构成的分层去噪架构。在视频内 层面，TCD通过建模视频内的时间依赖关系来优化帧级表示，从而生成偏好引导的视频特征。随后，这些优化后的特征被用于用户建模。在视频间 层面，NPD对交互序列进行去噪，从含噪的行为历史中恢复干净的用户偏好表示。最后，根据去噪后的用户偏好表示对候选物品进行排序，生成推荐结果。

4.3 时间引导内容扩散

现有方法通常将视频压缩为静态嵌入，忽略了帧间细粒度的时间依赖关系。为解决这一局限，TCD显式建模帧级时间动态，以捕捉微视频中不断变化的视觉内容。TCD模块将时间依赖建模与基于扩散的去噪相结合，并根据用户的历史观看行为对帧序列表示进行优化。如图2所示，TCD由两个部分组成：(1)视频内时间依赖关系建模，用于捕捉帧间的时间关系；(2) 基于扩散的特征去噪与优化，在偏好引导下逐步提升表示质量。

4.3.1 视频内时序动态建模。 为捕捉视频内时序结构，TCD 对用户最近一次交互的视频中的帧级依赖关系进行建模 i_L（$S_u = [i_1, \ldots, i_L]$ 中的第 L 个物品），该视频提供了当前用户意图的最新证据。我们将该视频表示为采样帧特征序列 $\mathbf{F}_L = [\mathbf{f}_{L,1}, \ldots, \mathbf{f}_{L,K}] \in \mathbb{R}^{K \times d_v}$。为保留时间顺序，我们使用可学习的位置嵌入 $\mathbf{P} \in \mathbb{R}^{K \times d_v}$ 增强帧特征，并应用时间注意力编码器：

$$\mathbf{z}_0 = \text{TempAttnEnc}(\mathbf{F}_L + \mathbf{P}), \tag{6}$$

其中$\mathbf{z}_0 \in \mathbb{R}^{K \times d_v}$表示经过时间上下文建模的帧表示。为融入用户从交互历史中体现的不断变化的偏好，我们基于通过指数衰减权重聚合得到的时近性感知视觉上下文向量$\mathbf{c}_v$，对后续的扩散细化过程进行条件控制：

$$\alpha_k = \frac{\exp(-\gamma(L-k))}{\sum_{j=1}^{L}\exp(-\gamma(L-j))}, \qquad \mathbf{c}_v = \sum_{k=1}^{L} \alpha_k \, \mathbf{v}_k, \tag{7}$$

其中，$\gamma > 0$ 控制衰减率，$\{\mathbf{v}_k\}_{k=1}^{L}$ 是历史物品的全局视觉特征。上下文向量$\mathbf{c}_v$ 充当条件信号，引导扩散过程着重关注近期观看模式。

4.3.2 基于扩散的时序上下文去噪。我们采用第 3.1 节中的标准前向扩散过程，对$\mathbf{z}_0$进行破坏，并根据预定义的调度$\{\beta_t\}_{t=1}^{T}$得到噪声潜变量$\mathbf{z}_t$。随着 t 增加，$\mathbf{z}_0$中的时序结构会逐渐被破坏。为在偏好引导下逆转这种破坏，我们采用适用于时序序列的 U-Net架构，对条件去噪网络 f_θ进行参数化。在每个反向步骤

$$t \in \quad T \; T-\{,,,\},$$

1 1

网络都会预测干净的潜变量：

$$\hat{\mathbf{z}}_0^{(t)} = f_\theta(\mathbf{z}_t, t, \mathbf{c}_v), \tag{8}$$

其中，$\mathbf{c}_v$ 通过交叉注意力注入，以提供用户偏好引导。利用该预测，我们通过反向后验计算前一时刻的去噪潜变量：

$$\mathbf{z}_{t-1} = \boldsymbol{\mu}_\theta(\mathbf{z}_t, t, \mathbf{c}_v) + \sigma_t \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{9}$$

其中，$\boldsymbol{\mu}_\theta$根据标准去噪扩散概率模型（DDPM）公式，由 $\hat{\mathbf{z}}_0^{(t)}$ 推导得到。该迭代细化过程如下：

$$\mathbf{z}_T \xrightarrow{f_\theta(\mathbf{z}_T, T, \mathbf{c}_v)} \mathbf{z}_{T-1} \xrightarrow{f_\theta(\mathbf{z}_{T-1}, T-1, \mathbf{c}_v)} \cdots \xrightarrow{f_\theta(\mathbf{z}_1, 1, \mathbf{c}_v)} \hat{\mathbf{z}}_0, \tag{10}$$

从而得到用于后续用户建模的偏好对齐帧表示 $\hat{\mathbf{z}}_0 \in \mathbb{R}^{K \times d_v}$。因此，我们采用扩散重建目标优化 f_θ，以促进在$\mathbf{c}_v$条件下对干净潜变量的忠实恢复。形式化地，我们最小化：

$$\mathcal{L}_{\mathrm{T}} = \mathbb{E}_{\mathbf{z}_0, t}\left[\|\mathbf{z}_0 - f_\theta(\mathbf{z}_t, t, \mathbf{c}_v)\|^2\right]. \tag{11}$$

4.4 无噪声条件偏好去噪

尽管视频内时序建模能够捕捉单个视频内部的动态变化，却忽略了对用户交互序列中依赖关系的建模。实际上，行为日志常常受到噪声行为的污染（例如误点和探索性浏览），这可能会掩盖潜在的偏好信号。受近期研究发现的启发，即受损的高维观测能够揭示噪声特征，[15], 我们提出了一种无时间步估计器，利用多模态内容先验进行盲偏好恢复。

我们将用户的交互序列 $\mathrm{S}_u = [i_1, \ldots, i_L]$建模为其潜在偏好的噪声表现形式。令$\mathbf{E} \in \mathbb{R}^{|\mathcal{I}| \times d}$表示物品嵌入矩阵。我们将序列编码器应用于嵌入后的序列 $[\mathbf{e}_1, \ldots, \mathbf{e}_L]$，以获得初始偏好表示：

$$\mathbf{x}_0 = \mathrm{SeqEnc}([\mathbf{e}_1, \ldots, \mathbf{e}_L]) \in \mathbb{R}^d. \tag{12}$$

按照第 3.1 节所述，我们应用标准前向扩散过程对$\mathbf{x}_0$进行扰动，并获得步骤 t处的噪声潜变量$\mathbf{x}_t$，其中噪声水平由预定义的调度方案控制。

为偏好重建提供丰富的语义引导，我们以最新交互物品 i_L 的两种互补内容信号作为去噪过程的条件：(1) 经 TCD 优化的视觉表示 $\hat{\mathbf{z}}_0 \in \mathbb{R}^{K \times d_v}$，其蕴含了与偏好对齐的时间动态；(2) 从标题和描述中提取的文本特征 $\mathbf{h}_L \in \mathbb{R}^{d_h}$。

我们使用自注意力机制对去噪网络 g_θ进行参数化，使每种模态都能关注其他模态，从而促进语义对齐和噪声识别。关键的是，g_θ在不访问扩散时间步 t的情况下运行，促使

表1：数据集的详细统计信息。

数据集	#用户	#物品	#交互	平均长度	稀疏度
Microlens-small	8,694	8,265	50,838	5.85	99.93%
Microlens-big	86,073	18,700	597,104	6.94	99.96%
Shortvideo-small	3,460	84,192	240,843	69.61	99.92%
Shortvideo-big	7,584	122,845	533,381	70.33	99.94%

模型仅根据各模态之间的语义一致性来推断噪声特征。

$$\mathbf{x}_T \xrightarrow{g_\theta(\mathbf{x}_T, \mathbf{c}_m)} \mathbf{x}_{T-1} \xrightarrow{g_\theta(\mathbf{x}_{T-1}, \mathbf{c}_m)} \mathbf{x}_{T-2} \xrightarrow{g_\theta(\mathbf{x}_{T-2}, \mathbf{c}_m)} \cdots \xrightarrow{g_\theta(\mathbf{x}_1, \mathbf{c}_m)} \hat{\mathbf{x}}_0, \tag{13}$$

值得注意的是，g_θ 不依赖时间步，因此仅依靠 $\hat{\mathbf{z}}_0$ 与$\mathbf{h}_L$之间的跨模态语义一致性进行盲去噪。此外，多模态条件$\mathbf{c}_m$ 在整个反向过程中保持固定，使去噪轨迹锚定于 i_L 的内容，并减轻表征漂移。

因此，我们通过重建目标优化 g_θ，该目标鼓励在固定多模态条件下从噪声输入中准确恢复$\mathbf{x}_0$：

$$\mathcal{L}_{\mathrm{N}} = \mathbb{E}_{\mathbf{x}_0, t}\left[\|\mathbf{x}_0 - g_\theta(\mathbf{x}_t, \hat{\mathbf{z}}_0, \mathbf{h}_L)\|^2\right]. \tag{14}$$

4.5 优化目标

我们通过对项目集合 I 使用交叉熵，优化判别式下一项预测目标：

$$\mathcal{L}_{\mathrm{R}} = -\sum_{(u,i) \in \mathcal{D}} \log \frac{e^{\hat{y}_{ui}}}{\sum_{k \in \mathcal{I}} e^{\hat{y}_{uk}}}. \tag{15}$$

我们使用混合目标对该框架进行端到端训练，该目标结合了判别式排序损失和两个去噪重建损失：

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{R}} + \lambda_{\mathrm{T}} \mathcal{L}_{\mathrm{T}} + \lambda_{\mathrm{N}} \mathcal{L}_{\mathrm{N}}, \tag{16}$$

其中，λ_{T} 和 λ_{N} 用于平衡去噪项。$\mathrm{L_R}$、$\mathrm{L_T}$和 $\mathrm{L_N}$ 分别在式 (15)、式 (11)和式 (14)中定义。

4.6 讨论

4.6.1 复杂度分析。 我们进一步分析 MealRec 的训练计算复杂度，并讨论其相较于标准基于扩散的推荐模型的效率。MealRec 的训练开销主要来自序列建模和细粒度视觉重建。令 L、d和 K分别表示序列长度、隐藏维度和采样的视频帧数量。序列编码器和基于注意力的 NPD 估计器的复杂度均为 $\mathrm{O}(L^2 d)$，因为它们使用自注意力来捕获用户历史中的时间依赖关系。基于 U-Net 的 TCD 模块对每个物品的 K 帧进行时间卷积，复杂度为 $\mathrm{O}(LKd^2)$。由于 TCD 帧处理可在各物品间完全并行，且 NPD 采用单时间步采样进行训练，因此总体训练复杂度为 $\mathrm{O}(L^2 d + LKd^2)$，使 MealRec 的计算开销与标准多模态序列推荐系统相当。

4.6.2 与现有方法的比较。现有微视频推荐方法主要分为两种范式。（i）基于预训练特征的方法会离线提取视频特征，并

表2：MealRec在来自两个平台的四个微视频数据集上与十种基线方法的性能比较。*表示MealRec相较于基线方法取得了显著提升（使用配对t检验，$p < 0.01$）。

算法	Microlens-small				Microlens-big				Shortvideo-small				Shortvideo-big			
	H@10	N@10	H@20	N@20	H@10	N@10	H@20	N@20	H@10	N@10	H@20	N@20	H@10	N@10	H@20	N@20
SASRec（18）	0.0781	0.0459	0.1000	0.0515	0.0714	0.0363	0.1082	0.0465	0.0121	0.0073	0.0171	0.0085	0.0133	0.0064	0.0224	0.0087
CL4SRec（22）	0.0789	0.0478	0.1020	0.0536	0.0735	0.0382	0.1122	0.0479	0.0142	0.0082	0.0214	0.0100	0.0178	0.0085	0.0280	0.0110
SSDRec (24)	0.0710	0.0377	0.0942	0.0436	0.0558	0.0265	0.0867	0.0343	0.0090	0.0056	0.0116	0.0062	0.0127	0.0077	0.0161	0.0085
DiQDiff (25)	0.0614	0.0399	0.0682	0.0416	0.0734	0.0403	0.1091	0.0492	0.0083	0.0049	0.0147	0.0065	0.0129	0.0065	0.0217	0.0087
MoRec (23)	0.0681	0.0322	0.1028	0.0410	0.0540	0.0265	0.0893	0.0354	0.0140	0.0080	0.0216	0.0099	0.0145	0.0072	0.0198	0.0086
TedRec (24)	0.0837	0.0463	0.1056	0.0518	0.0743	0.0387	0.1113	0.0481	0.0153	0.0078	0.0249	0.0102	0.0185	0.0089	0.0301	0.0118
IISAN (24)	0.0793	0.0397	0.1024	0.0457	0.0759	0.0390	0.1147	0.0488	0.0154	0.0086	0.0228	0.0104	0.0140	0.0079	0.0200	0.0094
DMMD4SR (25)	0.0805	0.0431	0.1011	0.0483	0.0796	0.0385	0.1182	0.0482	0.0139	0.0068	0.0234	0.0091	0.0182	0.0089	0.0281	0.0113
MMGCN (19)	0.0369	0.0150	0.0774	0.0253	0.0202	0.0093	0.0348	0.0130	0.0116	0.0053	0.0173	0.0068	0.0117	0.0066	0.0193	0.0084
IISAN-Verse (25)	0.0830	0.0478	0.1139	0.0536	0.0769	0.0414	0.1114	0.0501	0.0175	0.0089	0.0236	0.0104	0.0154	0.0083	0.0222	0.0100
MealRec（我们的方法）	**0.0897***	**0.0498***	**0.1195***	**0.0573***	**0.0809**	**0.0431***	**0.1225***	**0.0536***	**0.0194***	**0.0090**	**0.0292***	**0.0115***	**0.0207***	**0.0104***	**0.0335***	**0.0136***
相对重要性	*7.17%*	*4.18%*	*4.92%*	*6.90%*	*1.63%*	*4.11%*	*3.64%*	*6.99%*	*10.86%*	*1.12%*	*17.27%*	*10.58%*	*11.89%*	*16.85%*	*11.30%*	*15.25%*

将其作为序列建模的固定输入。尽管这类方法效果不错，但无法根据用户的个性化兴趣调整视觉表示，而且往往忽略细粒度视频内时序线索。（ii）端到端特征学习方法将视觉编码器与推荐目标联合优化，这通常需要大量训练时计算和存储开销。与这些范式相比，我们的框架旨在缓解这两个问题。具体而言，TCD 执行偏好引导的扩散细化，生成具有时间上下文且与偏好对齐的视觉表示；NPD 则通过由多模态内容先验引导的迭代去噪来恢复用户偏好。这种设计将偏好重建与含噪交互解耦，并为推荐生成稳健的用户表示。

5 实验

我们进行了大量实验与分析，以回答以下研究问题：

研究问题1：*MealRec* 在微视频推荐方面与最先进的基线方法相比表现如何？（参见 第5.2节）**RQ2**：*MealRec* 提出的各个组件如何影响其推荐性能？（参见 第5.3节）**RQ3**：*MealRec* 如何助力视频表示和用户偏好建模？（参见 第5.6节）**RQ4**：使用其他基础序列编码器时，*MealRec* 是否仍然有效？（参见第5.5节）**RQ5**：在不同噪声水平和超参数设置下，*MealRec* 的表现如何？其计算复杂度与有竞争力的基线方法相比又如何？（参见第5.4节）

5.1 实验设置

5.1.1 数据集。我们在来自两个平台的四个真实世界微视频数据集上开展实验：Microlens-small 和 Microlens-big [42]，，以及 Shortvideo-small 和 Shortvideo-big [43]。为提高数据可靠性，我们采用 4-core过滤，并为每位用户构建按时间顺序排列的交互序列。我们遵循留一法设置，将最后一次和倒数第二次交互分别用于测试和验证。数据集的主要统计信息汇总于表1。

5.1.2 基线方法。为评估*MealRec*的有效性，我们将其与多种基线方法进行比较，并将这些方法分为三类：序列推荐方法（即 SAS-Rec [21], CL4SRec [25], SSDRec [44], 和 DiQDiff [13]）、多模态序列推荐方法（即 MoRec [45], TedRec [41],IISAN [46], 和 DMMD4SR [12]），以及视频推荐方法（即 MMGCN [8] 和 IISAN-Verse [10]）。

5.1.3 实现细节。所有实验均使用 Python 3.10.18 在单块 NVIDIA A100GPU 上进行。对于视觉输入，我们遵循 [42] 中的帧提取策略，以每秒 1 帧的速度从每个视频的中间部分均匀采样，得到每个视频的五帧。使用预训练的 VideoMAE [47] 对这些帧进行编码，以获取用于多模态建模的帧级特征和全局视觉表示。对于文本，我们在 MicroLens 数据集上使用 BERT [48] 提取文本特征，并在 Shortvideo 数据集上使用 GloVe [49]提取文本特征。

我们采用 SASRec、CL4SRec 和 TedRec 作为替代序列骨干网络来实例化 *MealRec*，同时保持架构的其余部分不变。为确保公平比较，所有变体均采用相同的骨干网络配置。具体而言，序列编码器由两个自注意力模块组成，每个模块包含两个注意力头，最大序列长度设为 10。我们使用 Adam 训练模型，学习率为 1×10^{-3}，批次大小为 2048。在 TCD 模块中，时间衰减因子设为 $\gamma = 0.9$。训练期间，我们从 $\{5, 10, 15, 20, 25\}$ 中均匀采样扩散步 t，并为 TCD 和 NPD 使用相同的扩散步数，以减少超参数调优，即 $T^t = T^n = T$。损失权重 λ_T 和 λ_N 通过网格搜索在 $\{0.01, 0.05, 0.1, 0.3, 0.5, 0.7, 0.9\}$ 范围内进行调优。

5.2 性能比较（研究问题1）

表2报告了四个数据集上的性能比较，评估指标为命中率（H@k）和归一化折损累积增益（N@k），其中 k∈ {10，20}。此处，最佳和次佳结果分别以**粗体** 和下划线标出。总体而言，*MealRec* 在所有数据集和指标上始终取得最佳结果，证明了其在微视频推荐方面的有效性。我们将这些提升归因于其对细粒度视频内动态和序列级偏好去噪的统一建模。

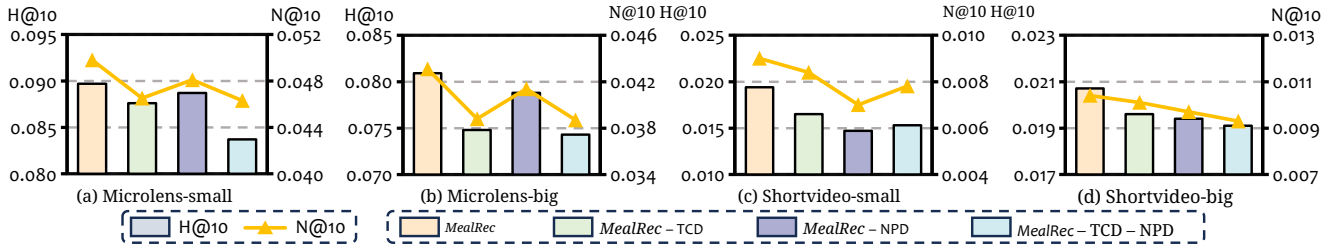

图3：MealRec 在四个数据集上的消融分析。总体而言，MealRec 中的每个组件都有效。

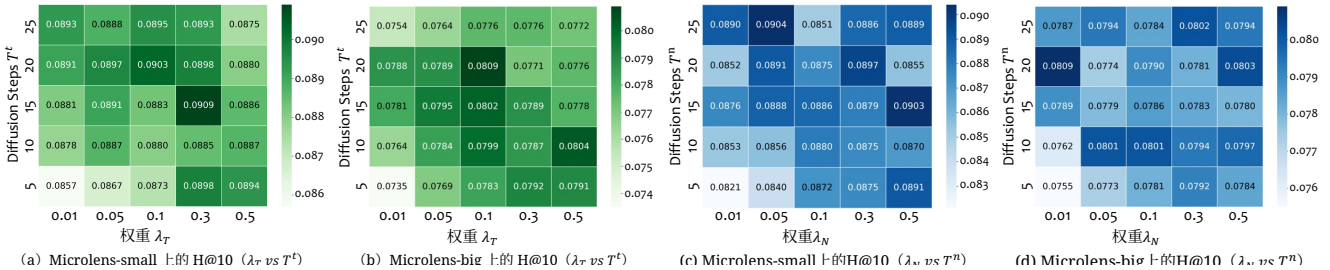

图4：参数敏感性分析。(a) 和 (b)：TCD 的损失权重 λ_T 与扩散步 T^t 的比较；(c) 和 (d)：NPD 的损失权重 λ_T 与扩散步 T^t 的比较。颜色越深表示分数越高。

与**仅基于序列的**基线方法（例如，SASRec、CL4SRec、SSDRec 和 DiQDiff）相比，*MealRec* 通过融合视频和文本内容，取得了持续的性能提升。这些提升凸显了多模态语义对微视频推荐的价值：细粒度视觉细节和描述性线索为用户兴趣提供了丰富的信息。通过融入内容感知表示，*MealRec* 能够更好地区分视觉主题相似的物品，并提高识别真正符合用户偏好的视频的准确性。

与多模态序列基线方法（例如，MoRec、TedRec、IISAN 和 DMMD4SR）相比，MealRec 通过利用细粒度的视频内内容信息进一步提升了性能。现有方法主要依赖物品级表示，因此往往忽视单个视频中丰富的时间动态。我们的结果凸显了在微视频推荐中，对超越粗粒度特征的视频内结构进行建模的重要性。通过融入帧级时间表示，MealRec 能更好地解耦用户兴趣并提高推荐准确率。

与面向视频的基线方法（例如，MMGCN 和 IISAN-Verse）相比，结果凸显了联合建模序列依赖和视频内时序结构的重要性。MealRec 的表现始终优于这些方法，表明其鲁棒性和泛化能力有所提升。这些提升主要源于分层扩散：它在保留微视频时间动态的同时，对偏好表示进行去噪。

5.3 消融研究（RQ2）

为考察所提出的各个模块是否有效，我们在 TedRec 上开展消融研究，以量化扩散组件的影响。如图3所示，我们评估了若干变体，以考察各组件的作用。我们可以观察到

即：1) **MealRec-TCD** 将精炼后的视频表示替换为视频编码器提取的初始特征（即禁用基于TCD的视觉扩散），以检验基于扩散的

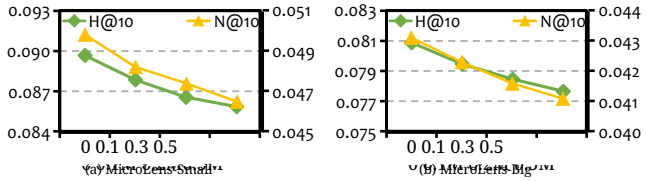

图5：在预先提取的视觉特征上添加不同水平的高斯噪声（σ）时，*MealRec* 的性能。

视觉精炼是否必要。结果表明，TCD在大多数数据集上都带来了稳定的提升，凸显了其从细粒度视频动态中捕捉用户观看兴趣的有效性。在Shortvideo-small上的轻微性能下降表明，该设置下的用户序列可能噪声较大，说明交互序列仍需进一步去噪。2) **MealRec-NPD**移除了NPD扩散模块，改为使用多层感知机融合多模态特征。所有数据集上的性能下降表明，NPD在缓解用户交互序列中的噪声、稳定偏好演变方面发挥着重要作用。3) **MealRec-TCD-NPD** 禁用了两个扩散模块，导致所有变体中性能下降幅度最大。该结果表明，视觉扩散与偏好去噪具有互补性，对细粒度视频动态和去噪后的偏好序列进行联合建模，是提升*MealRec*性能的关键。

5.4 鲁棒性分析（RQ5）

5.4.1 不同噪声水平下的鲁棒性分析。为分析 MealRec 在视频帧表示受到扰动时的鲁棒性，我们在 Microlens-small 和 Microlens-big 数据集上，向视频编码器生成的视频嵌入中注入不同强度（0、0.1、0.3 和 0.5）的标准高斯噪声。图 5 展示了 H@10 和 N@10 随噪声水平变化的趋势。随着噪声强度逐渐增大，H@10

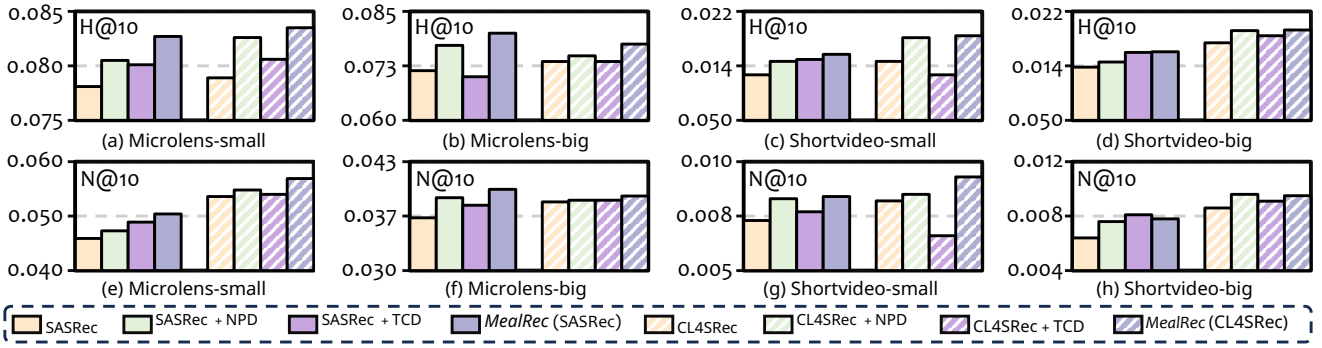

图 6：*MealRec* 的通用性分析。我们展示了 *MealRec* 在 SASRec 和 CL4SRec 上所有消融变体的结果。

表 3：TedRec（骨干模型）、IISAN-Versa（SOTA 基线）与我们的*MealRec*之间的计算复杂度比较。这里的"#Tra."、"#Eval."和"#Mem."分别表示每轮的训练时间、评估时间和内存使用量。

数据集	算法	H@20	#Tra.（秒）	#Eval.（秒）	#内存（MB）
MicroLens -小款	TedRec	0.1056	1.23	1.47	2,716
	IISAN-Versa	0.1139	180.53	12.63	40,894
	MealRec	0.1195	1.87	2.31	3,280
MicroLens -big	TedRec	0.0893	16.85	14.71	3,574
	IISAN-Versa	0.1114	2493.07	107.40	41,336
	MealRec	0.1225	24.94	22.08	4,012

和 N@10 仅略有下降，没有出现明显的性能退化。这种鲁棒性部分得益于 MealRec 中的分层扩散，它利用时间新近性感知的用户引导来优化视觉表示，并补充协同信息。

5.4.2 参数敏感性分析。 图4展示了不同数据集和扩散组件上的一致趋势。对于 TCD 和 NPD，*MealRec* 在广泛的超参数设置下均表现稳定，表明其对损失权重和扩散步数的选择不敏感。具体而言，中等损失权重与中等至较大的扩散步数相结合时，能够持续取得较强的性能，这表明重建强度与去噪深度之间实现了良好的平衡。此外，不同数据集上相对平滑的性能表现进一步证明，所提出的框架能够稳健应对超参数变化。这一特性对于实际部署十分理想，因为在实际场景中通常难以进行大量调参。

5.4.3 复杂度比较。 如图3所示，复杂度比较揭示了推荐性能与计算成本之间明显的权衡关系。尽管 IISAN-Versa 相较于 TedRec 带来了显著提升，但其计算开销也大幅增加，同时提高了运行时间和内存消耗，从而限制了其在大规模部署中的实用性。这一问题在规模更大的 MicroLens-big 数据集上表现得更加明显。相比之下，*MealRec* 在各数据集上均取得了最佳性能，同时保持了与轻量级骨干网络相当的效率。与 TedRec 相比，MealRec 在

仅适度增加训练时间和内存占用，且其开销仍远低于IISAN-Versa。总体而言，这些

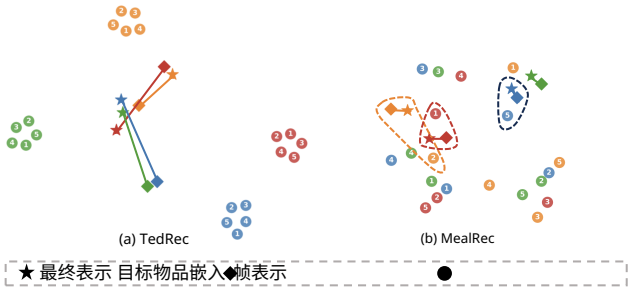

图7：TedRec（a）和*MealRec*（b）上的分布可视化。我们使用颜色区分用户，同时使用形状和数字分别区分同一视频中的偏好表示、目标视频嵌入和帧表示。

结果表明，*MealRec* 在不牺牲计算效率的情况下提升了推荐准确率，在性能与实际可部署性之间取得了良好平衡。

5.5 通用性分析（RQ4）

除 TedRec 之外，我们进一步在另外两个骨干网络（即SASRec 和 CL4SRec）上实例化 *MealRec*，以评估其通用性。

如图6所示，*MealRec* 在所有数据集上均持续优于全部骨干网络和消融变体，表明其在不同架构和数据特征下具有较强的通用性。我们进一步观察到，TCD 与 NPD 的贡献因数据集而异。具体而言，在 MicroLens 数据集上，TCD 带来的增益大于 NPD，这表明对细粒度时序动态进行建模所提供的偏好信号比仅进行去噪更具信息量。将 TCD 与 NPD 结合使用，相较任一组件单独使用都能带来额外提升，说明这两个模块具有互补优势。最后，在更大规模的数据集上，改进更为显著，表明所提出的蒸馏机制能够受益于更丰富的监督，并且非常适合大规模推荐场景。

5.6 可视化分析（RQ3）

为评估视频表示在引导偏好建模方面的有效性，我们在图7中通过 t-SNE 可视化了 TedRec 与 *MealRec* 的表示。

与 TedRec 相比，*MealRec* 使最终用户表示与目标物品嵌入之间的距离更小，表明偏好–目标对齐有所改善。此外，

MealRec 通过区分不同帧特征所体现的用户偏好，在帧表示之间形成了偏好感知结构。具体而言，兴趣相关帧更靠近用户表示，这表明时间帧建模与用户偏好密切相关，也证实了我们方法的有效性。

6 结论

本研究从多粒度序列建模的角度探讨微视频推荐，其中，有效的预测需要同时捕捉视频内时序动态和视频间偏好演化。我们提出了*MealRec*，这是一种层次扩散框架，旨在应对以往研究中的两个关键挑战：与偏好无关的视频表示提取，以及固有的模态冲突。具体而言，时间引导内容扩散（TCD）通过融入帧级时间引导和个性化协同信号来优化视频内表示；无噪声条件偏好去噪（NPD）则通过盲去噪进行信息融合，直接从受损状态中恢复有信息量的用户偏好，而无需进行以时间步为条件的 ϵ-预测。在四个数据集上的大量实验验证了该方法的有效性、通用性和鲁棒性。未来的工作包括推动*MealRec* 与视觉编码器进行端到端优化，以及融入互补的用户反馈和更丰富的语义线索，以加强偏好 elicitation。

参考文献

[1] Haokai Ma、Zhuang Qi、Xinxin Dong、Xiangxian Li、Yuze Zheng、Xiangxu Meng 和 Lei Meng。用于冷启动推荐的跨模态内容推断与特征增强。载于国际神经网络联合会议（*IJCNN 2023*），黄金海岸，澳大利亚，*2023*年*6*月*18*日至*23*日，第1–8页。IEEE，2023年。[2] Haokai Ma、Ruobing Xie、Lei Meng、Xin Chen、Xu Zhang、Leyu Lin 和 Jie Zhou。探索跨域推荐中的伪困难负样本。载于第*17*届*ACM*推荐系统会议论文集（*RecSys 2023*），新加坡，新加坡，*2023*年*9*月*18*日至*22*日，第502–514页，2023年。[3] Haokai Ma、Ruobing Xie、Lei Meng、Fuli Feng、Xiaoyu Du、Xingwu Sun、Zhanhui Kang 和 Xiangxu Meng。推荐中的负采样：综述与未来方向。*ACM*信息系统汇刊，2026年。[4] Yongxin Ni、Yu Cheng、Xiangyan Liu、Junchen Fu、Youhua Li、Xiangnan He、Yongfeng Zhang 和 Fajie Yuan。大规模内容驱动的微视频推荐数据集。载于第*34*届*ACM*国际信息与知识管理会议，第6486–6491页，2025年。[5] Yu Shang、Chen Gao、Jiansheng Chen、Depeng Jin、Meng Wang 和 Yong Li。学习用于微视频推荐的细粒度用户兴趣。载于第*46*届国际*ACM SIGIR*信息检索研究与发展会议，第433–442页，2023年。[6] Han Xu、Taoxing Pan、Zhiqiang Liu 和 Xiaoxiao Xu。面向短视频推荐的互信息感知知识蒸馏。载于第*31*届*ACM SIGKDD*知识发现与数据挖掘会议第*1*卷，第2725–2734页，2025年。[7] Xu Zhao、Ruibo Ma、Jiaqi Chen、Weiqi Zhao、Ping Yang 和 Yao Hu。通过指数高斯混合网络进行视频观看时长预测的多粒度分布建模。载于第*19*届*ACM*推荐系统会议，第309–318页，2025年。[8] Yinwei Wei、Xiang Wang、Liqiang Nie、Xiangnan He、Richang Hong 和 Tat-Seng Chua。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，第1437–1445页，2019年。[9] Zhiyu He、Zhixin Ling、Jiayu Li、Zhiqiang Guo、Weizhi Ma、Xinchen Luo、Min Zhang 和 Guorui Zhou。个性化推荐中的短视频片段级用户动态兴趣建模。载于第*48*届国际*ACM SIGIR*信息检索研究与发展会议，第1880–1890页，2025年。[10] Junchen Fu、Xuri Ge、Xin Xin、Alexandros Karatzoglou、Ioannis Arapakis、Kaiwen Zheng、Yongxin Ni 和 Joemon M Jose Joemon。序列推荐中多模态基础模型的高效有效适配。*IEEE* 知识与数据工程汇刊，2025年。[11] Jonathan Ho、Tim Salimans、Alexey Gritsenko、William Chan、Mohammad Norouzi 和 David J Fleet。视频扩散模型。《神经信息处理系统进展》，35:8633–8646，2022年。[12] Weihai Lu 和 Li Yin。DMMD4SR：基于扩散模型的序列推荐多层级多模态去噪。载于第33届ACM国际多媒体会议论文集，第6363–6372页，2025年。[13] Wenyu Mao、Shuchang Liu、Haoyang Liu、Haozhe Liu、Xiang Li 和 Lantao Hu。面向基于扩散的序列推荐的差异化量化引导。载于2025年ACM Web会议论文集，第425–435页，2025年。[14] Xiaoxi Cui、Weihai Lu、Yu Tong、Yiheng Li 和 Zhejun Zhao。基于条件扩散的特征去噪的多模态多行为序列推荐。载于第48届国际ACM SIGIR信息检索研究与发展会议论文集，第1593–1602页，2025年。[15] Qiao Sun、Zhicheng Jiang、Hanhong Zhao 和 Kaiming He。去噪生成模型是否需要噪声条件？arXiv预印本 arXiv:2502.13129，2025年。[16] Gaode Chen、Ruina Sun、Yuezihan Jiang、Jiangxia Cao、Qi Zhang、Jingjian Lin、Han Li、Kun Gai 和 Xinghua Zhang。面向冷启动短视频推荐的多模态建模框架。载于第18届ACM推荐系统会议论文集，第391–400页，2024年。[17] Yuyang Ye、Zhi Zheng、Yishan Shen、Tianshu Wang、Hengruo Zhang、Peijun Zhu、Runlong Yu、Kai Zhang 和 Hui Xiong。利用多模态大语言模型实现多模态序列推荐。载于AAAI人工智能会议论文集，第39卷，第13069–13077页，2025年。[18] Yunzhu Pan、Chen Gao、Jianxin Chang、Yanan Niu、Yang Song、Kun Gai、Depeng Jin 和 Yong Li。理解并建模短视频序列推荐中的被动负反馈。载于第17届ACM推荐系统会议论文集，第540–550页，2023年。[19] Steffen Rendle、Christoph Freudenthaler 和 Lars Schmidt-Thieme。为下一篮子推荐分解个性化马尔可夫链。载于第19届万维网国际会议论文集，第811–820页，2010年。[20] Balázs Hidasi、Alexandros Karatzoglou、Linas Baltrunas 和 Domonkos Tikk。基于循环神经网络的会话推荐。arXiv预印本 arXiv:1511.06939，2015年。[21] Wang-Cheng Kang 和 Julian McAuley。自注意力序列推荐。载于2018年IEEE国际数据挖掘会议（ICDM），第197–206页。IEEE，2018年。[22] Fei Sun、Jun Liu、Jian Wu、Changhua Pei、Xiao Lin、Wenwu Ou 和 Peng Jiang。BERT4Rec：基于Transformer双向编码器表示的序列推荐。载于第28届ACM国际信息与知识管理会议论文集，第1441–1450页，2019年。[23] Jiacheng Li、Yujie Wang 和 Julian McAuley。时间间隔感知的序列推荐自注意力。载于第13届国际网络搜索与数据挖掘会议论文集，第322–330页，2020年。[24] Liwei Wu、Shuqing Li、Cho-Jui Hsieh 和 James Sharpnack。SSE-PT：通过个性化Transformer实现序列推荐。载于第14届ACM推荐系统会议论文集，第328–337页，2020年。[25] Xu Xie、Fei Sun、Zhaoyang Liu、Shiwen Wu、Jinyang Gao、Jiandong Zhang、Bolin Ding 和 Bin Cui。序列推荐的对比学习。载于2022年IEEE第38届国际数据工程会议（ICDE），第1259–1273页。IEEE，2022年。[26] Hengchang Hu、Wei Guo、Yong Liu 和 Min-Yen Kan。序列推荐系统中的自适应多模态融合。载于第32届ACM国际信息与知识管理会议论文集，第843–853页，2023年。[27] Lingzi Zhang、Xin Zhou、Zhiwei Zeng 和 Zhiqi Shen。通过对比学习实现序列推荐的多模态预训练。ACM推荐系统汇刊，3(1):1–23，2024年。[28] Xiaoyan Zhao、Juntao You、Yang Zhang、Wenjie Wang、Hong Cheng、Fuli Feng、See-Kiong Ng 和 Tat-Seng Chua。NextQuill：用于增强大语言模型个性化的因果偏好建模。ICLR，2026年。[29] Yang Zhang、Wenxin Xu、Xiaoyan Zhao、Wenjie Wang、Fuli Feng、Xiangnan He 和 Tat-Seng Chua。面向基于大语言模型的推荐的强化潜在推理。ICLR，2026年。[30] Xiaoyan Zhao、Ming Yan、Yang Zhang、Yang Deng、Jian Wang、Fengbin Zhu、Yilun Qiu、Hong Cheng 和 Tat-Seng Chua。通过专家网络优化对话式推荐系统的强化策略。arXiv预印本 arXiv:2509.26093，2025年。[31] Haokai Ma、Ruobing Xie、Lei Meng、Xin Chen、Xu Zhang、Leyu Lin 和 Jie Zhou。面向跨域推荐的三元组序列学习。ACM信息系统汇刊，42(4)，2024年。[32] Haokai Ma、Ruobing Xie、Lei Meng、Yimeng Yang、Xingwu Sun 和 Zhanhui Kang。SeedRec：基于义原扩散的序列推荐。载于第三十三届国际人工智能联合会议论文集，IJCAI 2024，韩国济州，2024年8月3日至9日，第2270–2278页，2024年。[33] Haokai Ma、Yimeng Yang、Lei Meng、Ruobing Xie 和 Xiangxu Meng。用于推荐的多模态条件扩散模型。载于ACM Web会议2024伴随论文集，WWW '24，第1733–1740页，美国纽约，

2024年。计算机协会。[34] Yongfu Zha、Xinxin Dong、Haokai Ma、Yonghui Yang 和 Xiaodong Wang。Align-for-fusion：通过双向扩散协调三元组偏好，实现跨域序列推荐，2025年。[35] Wenjie Wang、Yiyan Xu、Fuli Feng、Xinyu Lin、Xiangnan He 和 Tat-Seng Chua。扩散推荐模型。载于第*46*届国际*ACM SIGIR*信息检索研究与发展会议论文集，第832–841页，2023年。[36] Yu Hou、Jin-Duk Park 和 Won-Yong Shin。基于扩散模型的协同过滤：揭示高阶连接性的潜力。载于第*47*届国际*ACM SIGIR*信息检索研究与发展会议论文集，第1360–1369页，2024年。[37] Yunqin Zhu、Chao Wang、Qi Zhang 和 Hui Xiong。用于协同过滤的图信号扩散模型。载于第*47*届国际*ACM SIGIR*信息检索研究与发展会议论文集，第1380–1390页，2024年。[38] Zihao Li、Aixin Sun 和 Chenliang Li。DiffuRec：用于序列推荐的扩散模型。*ACM*信息系统汇刊，42(3):1–28，2023年。[39] Qidong Liu、Fan Yan、Xiangyu Zhao、Zhaocheng Du、Huifeng Guo、Ruiming Tang 和 Feng Tian。序列推荐的扩散增强。载于第*32*届*ACM*国际信息与知识管理会议论文集，第1576–1586页，2023年。[40] Haokai Ma、Ruobing Xie、Lei Meng、Xin Chen、Xu Zhang、Leyu Lin 和 Zhanhui Kang。用于序列推荐的插件式扩散模型。载于*AAAI*人工智能会议论文集，第38卷，第8886–8894页，2024年。[41] Lanling Xu、Zhen Tian、Bingqian Li、Junjie Zhang、Daoyuan Wang、Hongyu Wang、Jinpeng Wang、Sheng Chen 和 Wayne Xin Zhao。推荐系统的序列级语义表示融合。载于*CIKM*。ACM，2024年。[42] Yongxin Ni、Yu Cheng、Xiangyan Liu、Junchen Fu、Youhua Li、Xiangnan He、Yongfeng Zhang 和 Fajie Yuan。一种内容驱动的微视频推荐

大规模数据集。*arXiv*预印本 *arXiv:2309.15379*，2023年。[43] Yu Shang、Chen Gao、Nian Li和Yong Li。一个包含移动短视频平台行为、属性与内容的大规模数据集。载于*2025*年*ACM Web*会议配套论文集，第793–796页，2025年。[44] Chi Zhang、Qilong Han、Rui Chen、Xiangyu Zhao、Peng Tang和Hongtao Song。SSDRec：面向序列推荐的自增强序列去噪。载于*2024*年*IEEE*第*40*届数据工程国际会议（*ICDE*），第803–815页。IEEE，2024年。[45] Zheng Yuan、Fajie Yuan、Yu Song、Youhua Li、Junchen Fu、Fei Yang、Yunzhu Pan和Yongxin Ni。推荐系统下一步何去何从？重新审视基于ID与基于模态的推荐模型。载于第*46*届国际*ACM SIGIR*信息检索研究与发展会议论文集，第2639–2649页，2023年。[46] Junchen Fu、Xuri Ge、Xin Xin、Alexandros Karatzoglou、Ioannis Arapakis、Jie Wang和Joemon M Jose。IISAN：利用解耦参数高效微调高效适配序列推荐的多模态表示。载于第*47*届国际*ACM SIGIR*信息检索研究与发展会议论文集，第687–697页，2024年。[47] Zhan Tong、Yibing Song、Jue Wang和Limin Wang。VideoMAE：掩码自编码器是自监督视频预训练的数据高效学习器。神经信息处理系统进展，35:10078–10093，2022年。[48] Jacob Devlin、Ming-Wei Chang、Kenton Lee和Kristina Toutanova。BERT：面向语言理解的深度双向Transformer预训练。载于*2019*年计算语言学协会北美分会会议论文集：人类语言技术，第*1*卷（长文与短文），第4171–4186页，2019年。[49] Jeffrey Pennington、Richard Socher和Christopher D Manning。GloVe：词表示全局向量。载于*2014*年自然语言处理实证方法会议论文集（*EMNLP*），第1532–1543页，2014年。