arXiv:2512.16891v1 [cs.CV] 18 Dec 2025

LinkedOut：从视频大语言模型中提取世界知识表示，用于新一代视频推荐

Haichao Zhang[1, †, ‡] YaoLu[2, †] Lichen Wang[2, †] Yunzhe Li[2] Daiwei Chen[3] YunpengXu[2] YunFu[1]1东北大学 [2]领英 [3]威斯康星大学麦迪逊分校

摘要

视频大语言模型（*VLLM*）通过在互联网规模数据上进行预训练，实现了具备世界知识的视频理解，并已在电影分析和视频问答等任务中展现出应用前景。然而，将视频大语言模型用于视频推荐等下游任务仍面临挑战，因为实际系统需要支持多视频输入、采用轻量级骨干网络、实现低延迟序列推理并快速响应。在实践中，（*1*）仅解码生成会导致序列推理延迟较高；（*2*）典型接口不支持多视频输入；（*3*）将输出限制为语言会丢弃对下游视觉任务至关重要的细粒度视觉细节。我们认为，这些局限源于缺少一种既能保留像素级细节、又能利用世界知识的表示。我们提出 *LinkedOut*，这是一种直接从视频中提取视频大语言模型世界知识的表示，可实现快速推理、支持多视频历史记录，并消除语言瓶颈。*LinkedOut* 使用视频大语言模型从原始帧中提取语义基础扎实、具备知识感知能力的令牌，并通过可提示查询和可选辅助模态进行引导。我们提出一种跨层知识融合混合专家模型，从丰富的视频大语言模型特征中选择适当的抽象层次，从而实现个性化、可解释且低延迟的推荐。据我们所知，*LinkedOut* 是首个基于视频大语言模型的视频推荐方法，无需手工设计标签即可处理原始帧，并在标准基准测试上取得最先进的结果。可解释性研究和消融实验验证了层多样性与逐层融合的优势，展现出一条切实可行的路径，充分利用视频大语言模型的世界知识先验和视觉推理能力，服务于推荐等下游视觉任务。

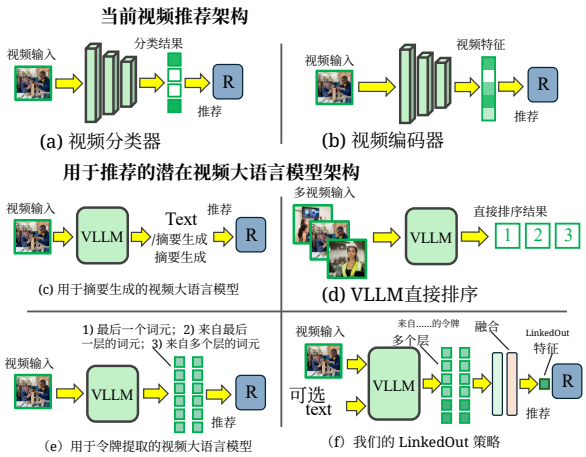


图 1. 我们的 LinkedOut 与其他方法在概念上的差异。现有方法将视频模型用作分类器或特征提取器（即 (a) 和 (b)），并使用不包含大语言模型/世界知识的特定任务数据进行训练。其他方案，例如 (c) 用于内容摘要的视频大语言模型、(d) 视频大语言模型直接进行视频排序，或 (e) 视频大语言模型进行词元级特征提取，既无法充分利用视频大语言模型中的世界知识，在计算上也无法满足推荐在线服务的要求。我们提出的 LinkedOut 方法能够充分利用视频大语言模型中的知识以及像素级视觉信息。最终，视频会被转换为特征向量，以便快速提供在线推荐服务。

1. 引言

视频推荐是观看者大规模发现内容的主要入口 [6]。其对轻量级骨干网络、低延迟顺序推理和快速响应的内在需求，使其成为研究基础模型集成的典型下游场景。尽管表征学习和排序架构取得了显著进展 [16, 26, 32, 34, 48], ，生产系统仍严重依赖从原始视频中提取的手工制作或预先提取的标签（例如类别、标签、反馈）。这种以标签为中心的设计降低了输入维度并简化了在线服务，但也丢弃了大部分

†共同贡献。
‡在领英（微软旗下公司）实习期间完成的工作。

像素中所包含的信息，将语义覆盖范围限制在训练集中已标注的内容，并使个性化在冷启动和长尾场景中变得脆弱。此外，在推荐前将视频概括为文本的处理流程会受到语言瓶颈的限制：它们将系统局限于预选特征和文本抽象，而非完整的视觉令牌空间。具体而言，这类处理流程难以识别细微属性（例如视觉幽默和叙事节奏），也难以利用更广泛的事实或常识背景，并且无法在不进行昂贵重新训练的情况下快速适应快速演变的趋势。

多模态大语言模型近期展现出从互联网规模预训练中获得强大可迁移性的能力，将视觉令牌与自然语言对齐，并编码广泛的世界知识 [1, 19, 21, 30]。在视频领域，对原始帧和转录文本进行联合训练，可解锁具有时间依据的语义和跨模态推理 [2, 25]。这些模型支持提示，因此无需修改主干网络，就能通过自然语言引导模型按需提取特定任务所需的属性和抽象信息。Transformer 层还会因深度不同而呈现专门化：较深层聚合全局背景和更高层次的语义，而较浅层则侧重局部模式。这表明，可以将视频大语言模型中的世界知识和像素级线索直接关联至推荐系统，无需先将视频简化为文本摘要，同时还能自适应地选择最适合每项排序决策的语义粒度。因此，将视频大语言模型应用于视频推荐，是超越手工设计或预提取的以标签为中心的处理流程，并将具备世界知识的视觉理解及潜在推理能力引入推荐系统的一种有前景的途径。

尽管近期许多工作构建了具备世界知识与推理能力的视频大语言模型（VLLM），使其能够用于视频内容理解、视觉问答和视频世界建模等任务，但这些下游用途的范围仍然狭窄。它们主要针对能够容忍当前视频大语言模型局限性的场景。当任务需要输入多个视频、使用轻量级骨干网络、进行低延迟的序列推理并快速响应时，这些局限性会使图1（c–e）所示的直接处理流程变得不切实际。

因此，视频推荐是分析这一问题的典型任务。它必须保留视频大语言模型中的预训练世界知识，但其输出是数据库中的项目索引，而非自然语言，因此不遵循语言分布。由于采用语言接口，当前的视频大语言模型通常依赖文本输出。一些任务（如机器人规划）会对视频大语言模型进行微调，使其生成结构化动作，但这往往会缩窄输出分布，并可能导致原有知识发生灾难性遗忘。此外，当前的视频大语言模型（以及许多大语言模型）在推理时速度较慢，这是因为其采用仅解码器的 Transformer 架构：每个新生成的令牌都必须作为下一步的输入重新馈入模型，从而延长推理过程，因此不适用于实时视频推荐。典型的视频大语言模型接口也缺乏对多视频输入的原生支持。单个视频就可能占用数万个令牌，而较大的令牌预算会进一步增加延迟和计算量，使得设计或后训练能够一次接收多个视频的视频大语言模型变得困难。然而，视频推荐通常需要处理每位用户的一系列历史视频，并编码大型候选池。迄今为止，这些挑战限制了视频大语言模型在此类任务中的应用，因为这类任务同时要求支持多视频输入、轻量级模型、低延迟序列推理和快速响应，同时仍需具备世界知识，以进行视觉推理、内容理解和泛化。

为此，我们提出***LinkedOut***，这是一种具备知识感知能力且可扩展至多种模态的视频推荐框架，它通过视频大语言模型将世界知识从视频像素中提取出来。其总体结构如图2所示。为解决语言输出约束，我们不对视频大语言模型进行微调以使其进入推荐空间，因为这可能导致领域偏移和灾难性遗忘。相反，我们直接从视频大语言模型的中间令牌表示中提取嵌入，因为世界知识分布在各个 Transformer 层中，并与输入词元共同计算。为协调骨干网络所编码的不同抽象层级，我们提出一种跨层知识融合专家混合模型，用于融合多个 Transformer 块的表示。具体而言，*LinkedOut* 在每一层中结合已有令牌和生成令牌，然后应用跨层门控，生成统一的推荐嵌入，将细粒度视觉线索与高层概念知识融合起来 [8, 33]。

为解决推理速度慢并支持大规模部署，*LinkedOut* 采用存储-检索架构。离线流水线首先通过跨层知识融合混合专家预先计算紧凑的 *LinkedOut*特征，并将其存储在嵌入数据库中。在服务时，轻量级推荐模块会检索候选嵌入，并根据用户上下文和提示快速生成排序结果。这种设计既保留了内容感知建模的优势，又能降低延迟，还可在不重新训练在线排序器的情况下离线添加新模态。此外，它还加强了对多视频输入的支持，因为特征提取和繁重的推理都与在线视频大语言模型路径隔离，避免了服务时将多个视频输入视频大语言模型所产生的二次方词元开销。为超越依赖标签或预选特征的传统视频推荐，我们的方法以直接从原始帧中提取的、语义扎实的密集词元嵌入取代脆弱的标签分类体系。可提示的特征聚焦模块会引导

主干网络聚焦于与用户相关的属性或辅助模态（例如转录文本、音频、元数据），或动态上下文（例如热门话题），并通过文本提示实现快速适配，无需重新训练。实证结果表明，*LinkedOut* 在公开视频推荐基准测试中取得了最先进的性能，这得益于视频大语言模型跨层聚合的世界知识。我们的主要贡献如下：

- 我们提出*LinkedOut*，这是一种具备知识感知能力且可扩展模态的视频推荐框架。它通过视频大语言模型从视频像素中提取世界知识特征，从而消除语言瓶颈。据我们所知，这是首个基于视频大语言模型的推荐系统，它整合了视频大语言模型并直接使用原始视频帧，同时避免了仅依赖文本的摘要生成瓶颈，并充分利用预训练视频大语言模型中来自网络规模数据的事实知识和常识知识。
- 我们设计了一种跨层知识融合混合专家，用于从视频大语言模型中间令牌的不同深度中选择并聚合适当抽象层次的表示，在严格的延迟限制下生成融合细粒度视觉线索与高层语义的统一嵌入。与最后一层或最后一个令牌的基线方法相比，该方法表现更优，并提升了在不同类型内容和用户意图下的稳健性。
- 为实现快速推理，我们采用存储与检索流水线，预先离线计算紧凑的*LinkedOut* 特征，并支持基于多视频历史的低延迟在线排序。在公开的视频推荐基准测试中，*LinkedOut* 取得了最先进的结果，这得益于视频大语言模型多个层中汇聚的世界知识先验。

2. 相关工作

2.1. 用于视频理解的多模态大语言模型

网络规模基础模型通过将视觉输入与自然语言监督对齐，极大推动了多模态表征学习的发展。早期工作如 CLIP 通过对比式预训练学习图像—文本对齐，并在各种识别任务中实现了强大的零样本迁移 [30]。沿袭这一范式，Flamingo 将冻结的视觉编码器通过门控交叉注意力与因果语言模型相结合，以支持在广泛的视觉—语言问题上的上下文学习 [1]。BLIP-2 引入轻量级查询变换器，将预训练的视觉编码器与大语言模型连接起来，从而提高数据效率和泛化能力 [19], 而 LLaVA 则扩展视觉指令微调，使大语言模型能够与图像为基础的指令及开放式对话对齐 [21]。

将多模态大语言模型从图像扩展到视频，需要对时间动态和长程依赖关系进行建模。大型旁白视频语料库，例如 HowTo100M 提供了网络规模的视频—文本配对监督，推动了视频—语言表征学习的发展 [25]。Frozen-in-Time 等架构联合编码帧和文本，以支持检索与迁移，为下游视频理解提供了有效基础 [2]。近期的视频大语言模型系统进一步结合逐帧注意力与指令微调，从长视频中提取时间对齐语义和世界知识 [1, 19]。与此同时，视频理解的效率与评估仍是核心挑战：VQToken 提出极端令牌压缩，以在最小令牌预算下实现轻量级视频大语言模型推理 [44], ；Dense Video Understanding 则引入了密集、细粒度的评估设置，以考验时间细节丰富的多模态推理 [45]。

Transformer 模型中一个一致的经验趋势是：较深层会聚合更广泛的上下文并编码更高层次的抽象，而较早层则保留更多局部且细粒度的模式。这促使我们采用层感知表征，以匹配下游任务所需的抽象层次。混合专家提供了一种可扩展机制，可将输入路由至专门化专家，从而提升模型容量，而无需按比例增加计算量 [33], ，这也补充了语言模型领域的混合专家进展，例如 Switch Transformer[8]。我们的工作借鉴这些见解，从视频大语言模型中提取逐层词元表示，并通过跨层混合专家将其融合，生成一种统一嵌入，兼具细粒度视觉线索与更高层次、以知识为基础的语义。不同于以往在不同模块或模型之间进行路由的混合专家方法，我们在单个视频大语言模型内显式融合跨层词元特征，因为不同层编码了不同的语义深度。

2.2. 推荐中的多模态大语言模型

传统推荐系统依赖协同过滤 [9, 12, 32]和序列建模 [13, 16, 43],，其运作主要基于物品 ID。为应对冷启动和长尾限制，内容感知推荐会利用辅助模态。早期方法从预训练网络中提取固定标签或浅层特征 [7]。近期的多模态推荐系统通过图卷积（MMGCN [37]）或基于图的表示（GRCN [39]）整合更丰富的特征。近期的基础模型带来了变革性的替代方案：类似 CLIP 的特征 [30]通过零样本迁移增强检索能力，而 Flamingo [1], 、BLIP-2 [19],和 LLaVA [21]等多模态大语言模型，则通过自然语言提示实现灵活的条件控制，并提供网络规模的世界知识。具体到视频推荐，目前已有三种范式：第一，使用冻结的预训练视频模型提取固定特征 [27]；第二，端到端联合训练视频编码器与推荐目标

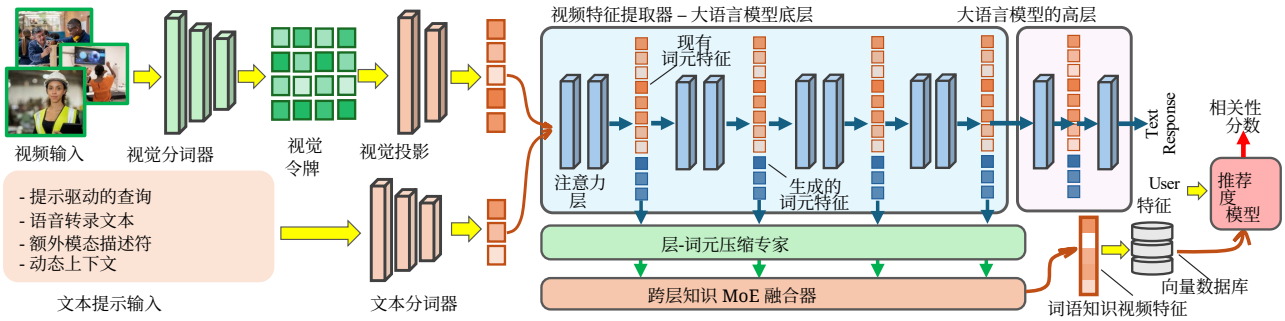


图2. LinkedOut 框架。原始视频帧经过词元化并投射到预训练视频大语言模型中。可选文本查询（例如转录文本、元数据和动态上下文）也会一并词元化并注入模型，以引导信息提取。随着令牌从较低层流向较高的 Transformer 层，跨模态注意力利用已有的视觉/文本令牌以及大语言模型生成的知识令牌来丰富序列。层-词元压缩专家会压缩令牌序列，而层级知识融合专家混合模型则通过跨层门控，自适应地选择合适的语义深度，从而生成统一的世界知识感知视频嵌入。在服务时，用户特征和视频嵌入会传递给轻量级推荐模型，以生成相关性分数。这种可提示设计支持灵活的模态融合和选择性微调，从而实现个性化、可解释且对冷启动具有鲁棒性的视频推荐。

标 [27]；第三，通过生成字幕或自动语音识别将视频概括为文本，再应用基于文本的大语言模型。

然而，现有方法通常存在三方面局限：（1）冻结编码器将多模态特征视为静态描述符，难以适应推荐特定语义和用户上下文；（2）端到端训练虽然性能更佳，但计算成本高得令人难以承受；（3）基于文本的摘要生成引入了语言瓶颈，丢弃像素级视觉信息，同时依赖人工设计的提示或固定层，而这些提示或层可能造成抽象程度不足或过高。LinkedOut 通过直接从视频像素中提取知识、采用自适应逐层融合为每个实例选择恰当的语义粒度，并将离线嵌入计算与在线推理解耦以提升效率，从而弥合基础模型能力与生产级推荐系统约束之间的差距。

3. LinkedOut 方法

3.1. 预备知识与可解释性分析

视频大语言模型的预备知识。典型的视频大语言模型（VLLM）首先使用各自对应的分词器对每种输入模态（例如视频帧、文本提示、音频）进行分词。随后，将生成的词元投影到共同的嵌入空间并拼接成一个词元序列，再输入大语言模型，以利用模型中存储的世界知识进行推理[23, 29, 35, 40]。大语言模型本身由多个 Transformer 层组成。世界知识分布在这些预训练的 Transformer 层中。在每一层中，词元表示会经过自注意力和前馈模块更新，生成新的词元嵌入，并在自回归的下一词元预表达方式。

旧词元与新词元。解码期间，每个 Transformer 层首先对所有模态的全部输入词元应用（多头）自注意力或交叉注意力。完成这一轮处理后，模型开始自回归地生成词元：每个新预测出的词元都会追加到已有词元之后，并作为下一步的输入。键值缓存 [28] 可以加快这一过程，但整体解码仍是顺序进行的。为便于说明，在给定层中，我们将索引位于原始输入长度范围内的词元称为旧词元，将索引超出原始输入长度的词元称为新词元。

逐层词元表示与可解释性。近期研究表明，不同 Transformer 层往往捕获不同层次的知识 [4, 17, 22, 31]。例如，在一个包含 24 层的大语言模型中，我们可以粗略地将这些层分为早期、中期和后期阶段：早期层通常编码局部视觉模式（如边缘和纹理）及基础词元嵌入 [36, 46]；中间层通常捕获跨模态对齐 [15, 41]（例如，区域与短语之间的对应关系）、物体部件、关系和浅层语义分组；而后期层则更多与全局场景理解、抽象概念落地以及高层次多模态推理相关 [5, 20, 47]。因此，与仅依赖最后一层相比，从多个深度提取并融合表征更为合理。

LinkedOut 表示定义。 由于不同层编码了不同层次的视觉和语义知识，而最终的文本输出仅仅是 VLLM 内部推理的投影，因此我们的目标是提取多个层中的旧令牌和新令牌，从而保留并集中模型内部的世界知识嵌入 [5, 23, 35]。我们将*LinkedOut* 形式化为一种内容感知的知识基础表征，它将预训练视频大语言模型重新用作特征提取器，并

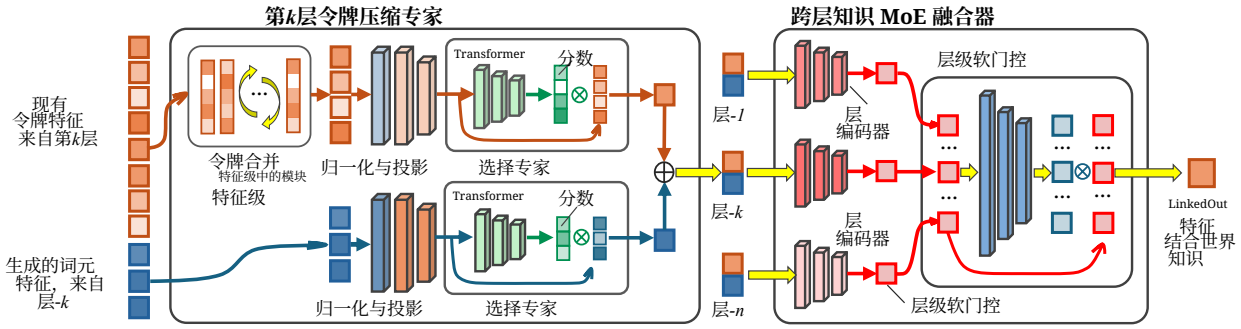


图3。跨层知识融合 MoE 概览。*令牌压缩专家* 接收现有的令牌特征和生成的令牌特征。令牌合并模块在特征级聚合冗余信号，随后进行归一化和投影。轻量级选择专家预测分数，以重新加权令牌；随后将两条流合并，为层 k生成紧凑且语义忠实的表示。*跨层知识 MoE 融合器* 将每一层视为一个专家。层编码器对压缩后的令牌进行概括，层级软门控生成上下文条件化的混合权重，自适应地融合各层信息。最终得到的 LinkedOut 特征将细粒度视觉细节与世界知识融合起来，提供单一的、可直接用于检索的嵌入，在提升效率的同时保留与推荐相关的语义。

通过提示暴露世界知识。设一个视频 v被采样为帧 $\{x_t\}_{t=1}^{T}$，并且可选的辅助通道（例如，自动语音识别转录文本、元数据、标签）被 s。自然语言提示 p引导主干网络关注与任务相关的属性。视频大语言模型 $\mathcal{L}$ 包含 L个 Transformer 模块，并为视觉和文本令牌组成的联合序列生成隐藏状态 $\{H^{(\ell)}\}_{\ell=1}^{L}$。下面我们将介绍如何收集这些逐层令牌特征；跨层知识 MoE 融合器将在第3.3.2节中介绍。图 3展示了两个核心模块：令牌压缩专家和跨层混合专家。

3.2. 原始世界知识令牌提取

由于世界知识编码在预训练 Transformer 权重中，并应用于每一层的词元嵌入，因此我们直接从大语言模型的中间层提取词元表示。原始世界知识令牌提取模块以从冻结（或选择性调优）的视频大语言模型 [1, 19, 21]中获得的稠密中间令牌特征，取代预定义或人工构造的标签。具体而言，每个帧 x_t 由视觉分词器 $g(\cdot)$ 分词为图像块，经轻量级适配器 ϕ （由主干网络的视觉-文本桥接模块初始化）投影至大语言模型的嵌入空间，并与文本令牌拼接：

$$Z = \big[\phi(g(x_1)), \ldots, \phi(g(x_T)),\ \mathrm{Tok}(p, s)\big]. \qquad (1)$$

主干网络 $\mathcal{L}$ 随后处理 Z 并生成逐层隐藏状态 $H^{(\ell)} \in \mathbb{R}^{N_\ell \times d}$，其中 N_ℓ 是层 ℓ的令牌长度，d是模型宽度。

3.3. 跨层知识融合 MoE

跨层知识融合 MoE 分两步从视频大语言模型中构建 *LinkedOut* 表示。首先，层级令牌压缩专家 将每层中的旧令牌和新令牌压缩为紧凑的特征。其次，跨层知识混合专家融合器 根据数据为各层分配权重，并融合其特征，从而生成统一且具备知识感知能力的物品嵌入。这两个组件在设计上紧密耦合；移除任何一个都会破坏*LinkedOut*的构建。

3.3.1. 层级令牌压缩专家

每层的令牌数量可能很多，这会带来沉重的计算负担，也使下游模块难以确定每层的贡献。为了让逐层特征更加紧凑且具有可比性，我们为每层分配一个令牌压缩专家，记为 $C^{(\ell)}$（见图3左侧面板）。每个专家通过独立分支处理旧令牌和新令牌，因为旧令牌主要编码累积上下文，而新令牌则捕获新近生成的知识。具体而言，旧分支先进行令牌合并，再进行压缩；新分支则直接压缩令牌。最后，我们将两个分支的输出拼接起来，形成层表示。

$$\mathbf{e}^{(\ell)} = \mathrm{concat}\Big(C^{(\ell)}_{\mathrm{old}}\big(\mathrm{Merge}^{(\ell)}_{\mathrm{old}}(H^{(\ell)}_{\mathrm{old}})\big),\ C^{(\ell)}_{\mathrm{new}}\big(H^{(\ell)}_{\mathrm{new}}\big)\Big).$$

这里，$C^{(\ell)}$ 旧和 $C^{(\ell)}_{\mathrm{new}}$ 是带有可学习查询的轻量级注意力池化模块；旧分支采用更激进的合并方式 [3] 来减少冗余。我们将此方法应用于每隔 N 层（例如，每隔 2 层或 4 层），实验表明，这既能保留世界知识随深度逐步演进的过程，又能使表示保持紧凑。

3.3.2. 跨层知识 MoE 融合器

视频大语言模型中不同深度的 Transformer 会将信息从局部模式组织到全局抽象。我们已经

表1. MicroLens-50K 和 MicroLens-100K 的统计数据。

统计数据	MicroLens-50K	MicroLens-100K
# 用户	50,000	100,000
# 项目（视频）	19,220	19,738
# 总交互次数	359,708	719,405
# 训练样本	239,511	478,355
# 验证样本	50,000	100,000
# 测试样本	50,000	100,000
# 细粒度标签	15,580	15,580
平均序列长度	7.19	7.19
每位用户的交互次数	5–15	5–15
视频时长	< 400 秒	< 400 秒

表2. MicroLens-50K 和 MicroLens-100K 的对比。据我们所知，它们是唯一公开可用的视频推荐原始视频数据集。

	MicroLens-50K				MicroLens-100K			
模型	**HR @10**	**HR @20**	**NDCG @10**	**NDCG @20**	**HR @10**	**HR @20**	**NDCG @10**	**NDCG @20**
YouTube [6]	0.0375	0.0632	0.0178	0.0245	0.0461	0.0747	0.0229	0.0301
VBPR [10]	0.0544	0.0888	0.0273	0.0361	0.0624	0.1002	0.0314	0.0410
MMGCN [38]	0.0403	0.0670	0.0197	0.0264	0.0214	0.0374	0.0103	0.0143
LightGCN [12]	0.0365	0.0534	0.0284	0.0345	0.0372	0.0618	0.0177	0.0239
GRCN [39]	0.0631	0.0982	0.0328	0.0415	0.0282	0.0497	0.0131	0.0185
LayerGCN [50]	0.0627	0.0994	0.0320	0.0414	0.0730	0.1120	0.0382	0.0480
BM3 [51]	0.0565	0.0918	0.0281	0.0372	0.0601	0.0975	0.0305	0.0401
Freedom [49]	0.0656	0.1028	0.0334	0.0429	0.0654	0.1016	0.0337	0.0431
MGCN [42]	0.0708	0.1089	0.0363	0.0459	0.0717	0.1096	0.0371	0.0467
MHCR [24]	0.0736	0.1102	0.0383	0.0477	0.0798	0.1187	0.0420	0.0519
LinkedOut（我们的方法）	**0.0761**	**0.1127**	**0.0399**	**0.0491**	**0.1015**	**0.1477**	**0.0548**	**0.0664**

在每一层的旧令牌和新令牌中集中存储了知识，但仍需学习哪些层对给定的下游实例最有用。

为了针对每个视频条目选择合适的语义粒度，我们使用混合专家（MoE）融合逐层特征 [8, 33]（见图3右侧面板）。最后的每个 N 提取层都会提供一个压缩后的向量$\tilde{\mathbf{e}}^{(\ell)} \in \mathbb{R}^{d_c}$（经过令牌压缩和提示混合后）。每层专家随后将该向量映射到公共空间：

$$\mathbf{h}^{(\ell)} = E^{(\ell)}\big(\tilde{\mathbf{e}}^{(\ell)}\big) \in \mathbb{R}^{d_z}, \tag{2}$$

其中，$E^{(\ell)}$ 是一个带有残差连接的轻量级多层感知机。随后，条目条件门控为各层分配软权重，且不会产生循环依赖：

$$\pi = \mathrm{Softmax}\big(G(\mathbf{z}_v)\big) \in \mathbb{R}^N, \quad \mathbf{z}_v = \sum_{\ell=1}^{N} \pi_\ell \, \mathbf{h}^{(\ell)}, \tag{3}$$

其中 $G(\cdot)$是一个门控多层感知机。由此得到的嵌入$\mathbf{z}_v$是排序器使用的统一、具备世界知识感知能力的条目表示。我们对所有提取的层采用稠密软权重。如有需要，也可以采用带负载均衡的稀疏 top-k 变体；不过，鉴于专家和层的数量不多，这里没有必要这样做。

表3. MicroLens-100k 基准测试中不同视频推荐模型类型的性能比较。

模型	**Type**	**HR @10**	**NDCG @10**	**HR @20**	**NDCG @20**
DSSM [14]	IDRec (CF)	0.0394	0.0193	0.0654	0.0258
LightGCN [12]		0.0372	0.0177	0.0618	0.0239
NFM [11]		0.0313	0.0159	0.0480	0.0201
DeepFM [9]		0.0350	0.0170	0.0571	0.0225
NextItNet [43]	IDRec (SR)	0.0805	0.0442	0.1175	0.0535
GRU4Rec [13]		0.0782	0.0423	0.1147	0.0515
SASRec [16]		0.0909	0.0517	0.1278	0.0610
YouTubeID [7]	VID Rec	0.0461	0.0229	0.0747	0.0301
YouTubeID+V [7]		0.0392	0.0188	0.0648	0.0252
MMGCNID [37]		0.0141	0.0065	0.0247	0.0092
MMGCNID+V [37]		0.0214	0.0103	0.0374	0.0143
GRCNID [39]		0.0282	0.0131	0.0497	0.0185
GRCNID+V [39]		0.0306	0.0144	0.0547	0.0204
DSSMID+V [14]		0.0279	0.0137	0.0461	0.0183
SASRecID+V [16]		0.0799	0.0415	0.1217	0.0520
NextItNetV [43]	视频 Rec	0.0862	0.0468	0.1246	0.0562
GRU4RecV [13]		0.0954	0.0517	0.1377	0.0623
SASRecV [16]		0.0948	0.0515	0.1364	0.0619
LinkedOut（我们的方法）	视频 LLMRec	**0.1015**	**0.0548**	**0.1477**	**0.0664**

3.4. 存储-检索架构

我们采用存储-检索架构，以实现快速的视频大语言模型推理。如图1（c–f）所示，在服务时直接应用视频大语言模型速度太慢。然而，在实际的推荐系统中，视频特征提取是在离线运行并定期更新的，与在线推理路径分开。由于 $\tilde{\mathbf{e}}^{(\ell)}$ 已针对每个条目（以及每个提示词库）预先离线计算，在线阶段只需评估门控网络和少量选定的专家。门控权重 π 表明哪些抽象层次会影响最终得分，而提示词混合则揭示了哪些属性受到强调，从而生成可读性强的解释。

通过使用少量规范提示对所有目录视频运行提取器，可以获得离线物品嵌入。生成的逐层向量按物品 ID 存储在特征库中，从而支持低延迟检索和在线排序，无需重新编码视频。提取器的计算量随帧数 T 和所选层数 N线性增长。令牌压缩将每个物品的内存占用从 $O(TNd)$ 降至 $O(d_c)$，其中 d是原始令牌维度，d_c 是压缩后的世界知识维度。这既加快了编码速度，也使存储与检索流水线能够在大规模场景下实际应用。

4. 实验

4.1. 数据集

我们在 MicroLens-50K 和 MicroLens-100K 上进行评估 [27]。据我们所知，这两个大规模、内容驱动的微视频推荐基准测试是唯一提供原始视频的公开数据集，可用于内容感知

表4. 组件有效性分析的消融研究。

变体	HR@10	NDCG@10	HR@20	NDCG@20
最后一层最后一个词元	0.0763	0.0404	0.1121	0.0494
平均池化+混合专家	0.0888	0.0471	0.1291	0.0572
最后一个词元+混合专家	0.0958	0.0518	0.1394	0.0628
LinkedOut（完整）	**0.1015**	**0.0548**	**0.1477**	**0.0664**

建模。表1 总结了关键统计数据。每个视频都包含丰富的多模态内容，例如原始帧（时长大多不足 400 秒）、音轨、标题、封面图像和用户评论，因此这些数据集适合用于评估基础模型驱动的推荐。交互数据由带时间戳的评论构成，涵盖 15,580 个细粒度标签（例如食品、体育、旅行、娱乐），支持序列建模。我们遵循官方的训练与测试协议 [27]。

4.2. 基准测试

我们将 LinkedOut 与最先进的视频推荐基线方法进行比较，这些方法大致可分为三类：(1) IDRec [27] 方法仅依赖物品 ID，不利用视频内容。其中包括协同过滤（CF）方法：DSSM [14] 通过双塔匹配学习用户/物品嵌入；LightGCN[12]在用户-物品二部图上应用图卷积；NFM [11] 通过神经双线性交互对二阶特征交互进行建模；DeepFM [9]将因子分解机与深度网络相结合。序列推荐（SR）模型包括采用循环单元的 GRU4Rec [13]、采用自注意力的 SASRec [16]，以及采用空洞卷积的 NextItNet [43]。(2) VIDRec 方法在 IDRec 的基础上加入冻结的预提取视频特征作为辅助信息，并以后缀“+V”表示（例如 YouTubeID+V[7], MMGCNID+V [37], GRCNID+V [39], DSSMID+V、SASRecID+V）。这些方法将视频嵌入与 ID 嵌入拼接，同时保持视频编码器冻结。YouTubeID [7] 是一个包含候选生成和排序的两阶段系统；MMGCN [37] 按模态分别传播用户-物品信号；GRCN [39]利用基于图的表示改进协同过滤。(3) VideoRec [27] 方法通过端到端（E2E）学习联合训练视频编码器和推荐系统，以可学习的视频表示取代 ID 嵌入（GRU4RecV [13], SASRecV [16], NextItNetV[43]）。这些 E2E 模型计算开销较大，但通过同时优化视频理解和推荐，取得了更优的性能 [27]。

4.3. 实现细节

视频大语言模型提取。 我们使用 LLaVA-OneVision [18] 7B 作为视频大语言模型的主干网络，并在特征提取期间冻结其参数。我们附加 PyTorch 钩子，以捕获生成过程中的词元嵌入：随着模型解码，先前生成的词元会先经过解码器处理，而新词元则逐个生成。为了减少计算量，我们每隔四个 Transformer 层记录一次视频大语言模型内部的词元表示。

推荐模型训练。 我们使用 8×NVIDIA H100 GPU进行混合精度训练，批次大小为 256 个序列。每个用户的历史记录最多包含 10 个视频。除非另有说明，我们使用 AdamW 训练 100 个 epoch，并为不同参数组设置专属学习率：视频编码器（未冻结时）为 0.0001，文本和图像编码器（池化器与分类器头部分开设置）为 0.0001，推荐头为 0.00001。权重衰减设为 0.1，并将全局梯度范数裁剪至 5。图像调整至 224×224。我们采用阶梯式调度，并以 epoch 为粒度使用预热调度器（间隔 $= 1$ epoch，$\alpha = 1.0$），默认预热步数为零。总损失结合了对齐项和均匀性项，按批次报告，并与推荐目标联合优化。

4.4. 性能分析

按视频推荐模型类别进行性能比较。表3 比较了 LinkedOut 与 MicroLens-100K 上四类最先进基线方法的表现。LinkedOut 取得了最佳性能，HR@10 达到 0.1015，NDCG@10 达到 0.0548，大幅超越所有基线方法。与最强的基于 ID 的序列推荐模型（SASRec，HR@10：0.0909）相比，LinkedOut 的相对性能提升了 11.7%，表明基础模型驱动的视频理解能够提供超越协同过滤模式的更丰富语义。与最佳端到端 VideoRec 模型（GRU4RecV，HR@10：0.0954）相比，LinkedOut 的相对性能提升了 6.4%，验证了我们的逐层 MoE 融合和词元压缩能够比标准视频编码器更有效地提取具有判别性的特征。值得注意的是，大多数 VIDRec 方法（例如 MMGCNID+V、GRCNID+V）的表现都不如其仅使用 ID 的对应模型，这表明将冻结的视频特征与 ID 嵌入进行简单拼接会引入噪声，而非有用信号。相比之下，LinkedOut 的自适应层选择和词元级聚合能够精细控制哪些语义层级参与推荐，避免 VIDRec 方法中观察到的特征退化。所有指标（HR@10、NDCG@10、HR@20、NDCG@20）均取得稳定提升，证实 LinkedOut 能够更有效地捕捉相关性和排序质量。

跨数据集比较。表2 比较了*Linked-Out* 在两个公开原始视频数据集 MicroLens-50KandMicroLens-100K 上与近期基线方法的表现。*LinkedOut*在两个数据集上的四项指标（HR@10、HR@20、NDCG@10、NDCG@20）均有所提升，表明其在不同数据集规模和内容复杂度下的表现始终如一——

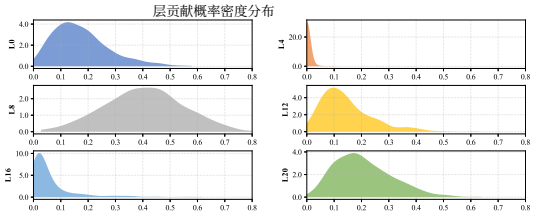


图4. 通过核密度估计（KDE）得到的逐层混合专家门控权重概率密度分布。每个子图展示一个 Transformer 层（L0、L4、L8、L12、L16、L20）的归一化贡献值分布。分布的宽度和峰值位置揭示了利用模式：L8 的分布最宽，中心位于 0.4；而 L4 和 L16 的分布更集中，贡献值也相对较小。

在不同的数据集规模和内容复杂度下的性能。

4.5. 消融研究

组件有效性分析。为验证设计选择的重要性，我们针对各个组件开展了消融研究。表4报告了 MicroLens-100K 上的结果。首先，我们评估了一个仅使用最后一层最后一个词元且不进行混合专家融合的基线。这种更简单的方法与标准大语言模型提取实践相似，导致性能显著下降（HR@10：0.0763 对比 0.1015），表明逐层融合和词元级聚合都至关重要。其次，我们考察了结合混合专家的均值池化方法：该方法采用逐层融合，但在每层对所有词元进行简单的平均池化。尽管这一变体受益于多层知识（HR@10：0.0888），但它对所有词元一视同仁，无法突出推荐相关特征。第三，我们测试了结合混合专家的最后一个词元方法：该方法采用逐层融合，但每层仅使用最后一个词元。这一变体取得了更好的性能（HR@10：0.0958），说明最后一个词元包含更多任务相关信息，但由于忽略了丰富的中间词元表示，其性能仍不及完整模型。这些消融结果证实，LinkedOut 的完整设计对于达到最先进水平至关重要。每个组件都对最终结果作出了有意义的贡献。

逐层贡献分析。为了了解哪些语义层级对视频推荐最有价值，我们分析了所有验证样本中学得的混合专家门控权重。我们从两个互补的角度可视化层贡献：图 4 展示了通过核密度估计 (KDE) 得到的各个采样层（L0、L4、L8、L12、L16、L20）的概率密度分布；图 5 展示了中位数、均值和分布离散程度的箱线图统计数据，并标注了平均贡献百分比。我们的主要发现是，中间层，
尤其是 L8，占比达 40.9%，显著优于早期层（L0：16.4%）和后期层（L20：20.5%），而 L4 和 L16 的贡献最小——

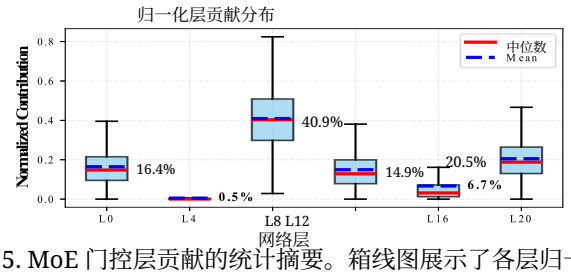


图5. MoE 门控层贡献的统计摘要。箱线图展示了各层归一化贡献值的分布，红线表示中位数，蓝色虚线表示均值。标注的百分比为平均贡献值，其中 L8 占主导地位，为 40.9%；其次是 L20（20.5%）和 L0（16.4%），而 L4（0.5%）和 L16（6.7%）的贡献相对较小。表5. 基于 *VLLM* 的推荐方案的推理时间比较。我们报告了直接通过 VLLM 在线服务的延迟，以及采用存储与检索设计时的延迟；传统的非 VLLM 模块仅作为系统级延迟的参考，并非语义基线。目标是使基于 VLLM 的方法达到实际可用的服务延迟。

基于 VLLM?	描述	Type	Time
✗	传统视频推荐 /仅供参考/ 图1 (a,b) 0.864 毫秒		
✓	基于 VLLM 的直接方案	图1（c、d）5510 毫秒	
✓	存储与检索 VLLM + 视频推荐模块（LinkedOut）	图1（f）5.964 毫秒	

仅为 0.5% 和 6.7%。核密度估计分布进一步表明，L8 在不同视频中呈现出广泛而稳定的模式，而 L4 和 L16 的分布则更为集中。这证实，视频推荐得益于自适应的多层级语义特征；与视觉语言模型中通常认为只依赖最深层不同，中间层表征能够平衡视觉细节与抽象概念。

效率分析。为展示我们的存储-检索架构在计算方面的优势，我们在单个 NVIDIA H100 GPU 上比较了三种方法的推理延迟（表5）。不使用 VLLM 的传统视频推荐（图1 a、b）每次推理耗时 0.864 毫秒，而直接进行实时 VLLM 推理（图1c、d）会产生高达 5.51 秒的延迟，因而无法部署。LinkedOut 的存储与检索范式（图1 f）将离线特征提取（批量处理 100 个视频时每个视频耗时 5.02 秒，仅需支付一次）与在线推理解耦；其中，混合专家融合（5.1 毫秒）和推荐排序（0.846 毫秒）总计每次查询仅需 5.964 毫秒，几乎 1000×比直接使用 VLLM 更快，同时保留丰富的语义理解能力。这一效率提升验证了基础模型驱动的视频推荐在大规模场景下的可行性。

5. 结论

我们提出了 LinkedOut，这是一种具备知识感知能力的视频推荐框架，能够将世界知识与原始像素

通过预训练视频大语言模型连接起来。通过直接从帧中提取具有语义依据的词元表示，LinkedOut 消除了传统方法中的语言瓶颈和手工标注约束。我们的可提示特征聚焦模块无需重新训练即可灵活引导提取过程，而跨层混合专家则能自适应地融合不同 Transformer 深度的表征，为每个实例选择恰当的语义粒度。在 MicroLens-50K 和 MicroLens-100K 上进行的实验表明，该方法达到了最先进的性能，并在个性化、冷启动和长尾场景中持续取得提升。消融研究证实了逐层融合、可提示提取和自适应语义深度选择的重要性。这些结果展示了一条切实可行的路径，推动下一代推荐系统充分利用视频大语言模型的世界知识和像素级视觉语义，实现稳健、可扩展的视频推荐。

参考文献

[1] Jean-Baptiste Alayrac、Jeff Donahue、Pauline Luc、Antoine Miech、Iain Barr、Yana Hasson 等。《Flamingo：一种用于少样本学习的视觉语言模型》，2022年。2、3、5[2] Max Bain、Arsha Nagrani、G¨ul Varol 和 Andrew Zisser-man。《Frozen in Time：用于端到端检索的视频和图像联合编码器》。载于 IEEE/CVF 国际计算机视觉会议论文集，2021年。2、3[3] Daniel Bolya、Cheng-Yang Fu、Xiaoliang Dai、Peizhao Zhang、Christoph Feichtenhofer 和 Judy Hoffman。《令牌合并：让你的视觉Transformer更快》。载于第十一届国际学习表征会议，2023年。5[4] Haoran Chen、Junyan Lin、Xinghao Chen、Yue Fan、Jian-feng Dong、Xin Jin、Hui Su、Jinlan Fu 和 Xiaoyu Shen。《多模态语言模型向浅层查看时表现更佳》。载于2025年自然语言处理实证方法会议论文集，第6688–6706页，2025年。4 – .[5] Haoran Chen、Junyan Lin、Xinhao Chen、Yue Fan、Xin Jin、Hui Su、Jianfeng Dong、Jinlan Fu 和 Xiaoyu Shen。《重新思考多模态大语言模型中的视觉层选择》。arXiv预印本 arXiv:2504.21447，2025年。4[6] Paul Covington、Jay Adams 和 Emre Sargin。《用于YouTube推荐的深度神经网络》。载于第十届ACM推荐系统会议论文集，第191–198页。ACM，2016年。1、6[7] Paul Covington、Jay Adams 和 Emre Sargin。《用于YouTube推荐的深度神经网络》。载于第十届ACM推荐系统会议论文集，第191–198页。ACM，2016年。3、6、7[8] William Fedus、Barret Zoph 和 Noam Shazeer。《Switch Transformer：以简单高效的稀疏性扩展至万亿参数模型》，2021年。2、3、6[9] Huifeng Guo、Ruiming Tang、Yunming Ye、Zhenguo Li 和 Xiuqiang He。《DeepFM：一种基于因子分解机的点击率预测深度神经网络》。载于第二十六届国际人工智能联合会议论文集，第1725–1731页。AAAI出版社，2017年。3、6、7[10] Ruining He 和 Julian McAuley。VBPR：基于隐式反馈的视觉贝叶斯个性化排序。载于《AAAI 人工智能大会论文集》，2016年。6[11] Xiangnan He 和 Tat-Seng Chua。用于稀疏预测分析的神经因子分解机。载于《第40届国际 ACM SIGIR 信息检索研究与发展大会论文集》，第355–364页。ACM，2017年。6, 7[12] Xiangnan He、Kuan Deng、Xiang Wang、Yan Li、Yongdong Zhang 和 Meng Wang。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大会论文集》，第639–648页。ACM，2020年。3, 6, 7[13] Bal´azs Hidasi、Alexandros Karatzoglou、Linas Baltrunas 和 Domonkos Tikk。基于循环神经网络的会话推荐。载于《国际学习表征会议》（ICLR），2016年。3, 6, 7[14] Po-Sen Huang、Xiaodong He、Jianfeng Gao、Li Deng、Alex Acero 和 Larry Heck。利用点击数据为网络搜索学习深度结构化语义模型。载于《第22届 ACM 国际信息与知识管理大会论文集》，第2333–2338页。ACM，2013年。6, 7[15] Qidong Huang、Xiaoyi Dong、Pan Zhang、Yuhang Zang、Yuhang Cao、Jiaqi Wang、Weiming Zhang 和 Nenghai Yu。通过模态集成率解读大型视觉语言模型中的跨模态对齐。载于《IEEE/CVF 国际计算机视觉大会论文集》，第218–227页，2025年。4[16] 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。1, 3, 6, 7[17] Michael Lepori、Alexa Tartaglini、Wai Keen Vong、Thomas Serre、Brenden M Lake 和 Ellie Pavlick。超越感知之门：视觉Transformer能够表征物体间的关系。《神经信息处理系统进展》，37：131503–131544，2024年。4[18] Bo Li、Yuanhan Zhang、Dong Guo、Renrui Zhang、Feng Li、Hao Zhang、Kaichen Zhang、Peiyuan Zhang、Yan wei Li、Zi-wei Liu 等。LLaVA-OneVision：轻松实现视觉任务迁移。arXiv 预印本 arXiv:2408.03326，2024年。7[19] Junnan Li、Dongxu Li、Silvio Savarese 和 Steven C. H. Hoi。BLIP-2：利用冻结图像编码器和大语言模型引导语言-图像预训练，2023年。2, 3, 5[20] Zhihang Lin、Mingbao Lin、Luxi Lin 和 Rongrong Ji。通过移除视觉令牌提升多模态大语言模型，实现快速推理。载于《AAAI 人工智能大会论文集》，第5334–5342页，2025年。4[21] Haotian Liu、Chunyuan Li、Qingyun Wu 和 Yong Jae Lee。视觉指令微调，2023年。2, 3, 5[22] Nelson F. Liu、Matt Gardner、Yonatan Belinkov、Matthew E. Peters 和 Noah A. Smith。上下文表征的语言学知识与可迁移性。载于《

2019年北美计算语言学协会分会会议：人类语言技术，第1卷（长篇与短篇论文），第1073–1094页，明尼阿波利斯，明尼苏达州，2019年。计算语言学协会。4[23] Y. Liu 等。面向多模态大语言模型的令牌高效长视频理解（storm）。arXiv 预印本 arXiv:2503 04130，2025 4 . . [24] Sisuo Lyu、Xiuze Zhou 和 Xuming Hu。用于冷启动微视频推荐的基于多视图超图的对比学习模型。载于《ICASSP 2025—2025年 IEEE 国际声学、语音与信号处理会议》（ICASSP），第1–5页。IEEE，2025年。6[25]Antoine Miech、Dimitri Zhukov、Jean-Baptiste Alayrac、Makarand Tapaswi、Ivan Laptev 和 Josef Sivic。HowTo100M：通过观看一亿个带旁白的视频片段学习文本-视频嵌入。载于《IEEE/CVF 国际计算机视觉大会论文集》，第2630–2640页，2019年。2, 3[26] Maxim Naumov、Dheevatsa Mudigere、Hao-Jun Michael Shi、Jianyu Huang、Narayanan Sundaraman 等。用于个性化和推荐系统的深度学习推荐模型。arXiv 预印本 arXiv:1906 00091，2019 1 . . [27] 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。arXiv 预印本 arXiv:2309.15379，2023年。3, 4, 6, 7[28] Reiner Pope、Sholto Douglas、Aakanksha Chowdhery、Jacob Devlin、James Bradbury、Jonathan Heek、Kefan Xiao、Shiv-ani Agrawal 和 Jeff Dean。高效扩展Transformer推理。《机器学习与系统会议论文集》，5：606–624，2023年。4[29] Rui Qian、Xiaoyi Dong、Pan Zhang、Yuhang Zang、Shuan-grui Ding、Dahua Lin 和 Jiaqi Wang。利用大语言模型进行流式长视频理解。《神经信息处理系统进展》，37 119360，2024 119336 4 : – .[30]Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sas-try、Amanda Askell、Pam Mishkin、Jack Clark、Gretchen Krueger 和 Ilya Sutskever。通过自然语言监督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》，第8748–8763页。PMLR，2021年。2, 3[31] Maithra Raghu、Thomas Unterthiner、Simon Kornblith、Chiyuan Zhang 和 Alexey Dosovitskiy。视觉Transformer看起来像卷积神经网络吗？《神经信息处理系统进展》，34 12128，2021 12116 4 : – . [32] Steffen Rendle。因子分解机。载于《2010年 IEEE 国际数据挖掘会议》，第995–1000页。IEEE，2010年。1, 3[33] Carlos Riquelme、Basil Mustafa、Joan Puigcerver、Rodolphe Jenatton、Michalis Titsias、Edward Lockhart 等。利用稀疏专家混合扩展视觉能力。载于《神经信息处理系统进展》，2021年。2, 3, 6 [34] Fei Sun、Jun Liu、Jian Wu、Changhua Pei、Xiao Lin、Wenwu Ou 和 Peng Jiang。BERT4Rec：利用 Transformer 双向编码器表示进行序列推荐。载于第28届ACM信息与知识管理国际会议论文集，第1441–1450页。ACM，2019年。 1[35] Yunlong Tang、Jing Bi、Siting Xu、Luchuan Song、Susan Liang、Teng Wang、Daoan Zhang、Jie An、Jingyang Lin、Rongyi Zhu 等。大语言模型的视频理解：综述。《IEEE电路与系统视频技术汇刊》，2025年。 4[36] Ian Tenney、Dipanjan Das 和 Ellie Pavlick。BERT重新发现了经典自然语言处理流程。arXiv预印本 arXiv:1905.05950，2019年。 4[37] 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国际多媒体会议论文集，第1437 1445页，2019年 3，6，7 – 。 [38] 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国际多媒体会议论文集，第1437–1445页，2019年。 6[39] Yinwei Wei、Xiang Wang、Liqiang Nie、Xiangnan He 和 Tat-Seng Chua。用于隐式反馈多媒体推荐的图细化卷积网络。载于第28届ACM国际多媒体会议论文集，第3541–3549页。ACM，2020年。 3，6，7[40] Jiayang Wu、Wensheng Gan、Zefeng Chen、Shicheng Wan 和 Philip S Yu。多模态大型语言模型：综述。载于2023年 IEEE国际大数据会议（BigData）论文集，第2247–2256页。IEEE，2023年。 4[41] Heeji Yoon、Jaewoo Jung、Junwan Kim、Hyungyu Choi、Heeseong Shin、Sangbeom Lim、Honggyu An、Chaehyun Kim、Jisang Han、Donghyun Kim 等。多模态大型语言模型的视觉表征对齐。arXiv预印本 arXiv:2509.07979，2025年。 4[42] Penghang Yu、Zhiyi Tan、Guanming Lu 和 Bing-Kun Bao。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，第6576–6585页，2023年。6[43] Fajie Yuan、Xiangnan He、Qingquan Jiang、Jiawei Xiong、Shengyu Zhao 和 Zhenhua Qin。用于下一项目推荐的简单卷积生成网络，2019年。 3，6，7[44] Haichao Zhang 和 Yun Fu。VQToken：面向视频大语言模型极端令牌缩减的神经离散令牌表示学习。载于第三十九届神经信息处理系统会议论文集，2025年。 3[45] Haichao Zhang、Wenhao Chai、Shwai He、Ang Li 和 Yun Fu。利用门控残差令牌化实现密集视频理解。arXiv预印本 arXiv:2509.14199，2025年。 3[46] Zhi Zhang、Srishti Yadav、Fengze Han 和 Ekaterina Shutova。多模态大型

语言模型中的跨模态信息流。载于计算机视觉与模式识别会议论文集，第19781–19791页，2025年。4[47] Shiyu Zhao、Zhenting Wang、Felix Juefei-Xu、Xide Xia、Miao Liu、Xiaofang Wang、Mingfu Liang、Ning Zhang、Dim- itris N Metaxas 和 Licheng Yu。通过搜索最优视觉令牌缩减方案加速多模态大型语言模型。载于计算机视觉与模式识别会议论文集，第29869–29879页，2025年。4[48] Guorui Zhou、Xiaoqiang Zhu、Chenru Song、Ying Fan、Han Zhu、Xiao Ma、Yanghui Yan、Junqi Jin、Han Li 和 Kun Gai。用于点击率预测的深度兴趣网络。载于*ACM SIGKDD*知识发现与数据挖掘国际会议论文集，第1059–1068页。ACM，2018年。1[49] Xin Zhou 和 Zhiqi Shen。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，第935–943页，2023年。6[50] Xin Zhou、Donghui Lin、Yong Liu 和 Chunyan Miao。用于推荐的层次精炼图卷积网络。载于*2023*年*IEEE*第*39*届国际数据工程会议（*ICDE*）论文集，第1247–1259页。IEEE，2023年。6[51] Xin Zhou、Hongyu Zhou、Yong Liu、Zhiwei Zeng、Chunyan Miao、Pengwei Wang、Yuan You 和 Feijun Jiang。多模态推荐的自举潜在表征。载于*2023*年*ACM Web*会议论文集，第845–854页，2023年。6