软件学报 ISSN 1000-9825, CODEN RUXUEW
2026,37(5):1950−1963 [doi: 10.13328/j.cnki.jos.007544] [CSTR: 32375.14.jos.007544]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

面向免训练视频问答的双重自适应冗余消除*

方承炀, 朱 畅, 姜文晖, 方玉明, 鄢杰斌

(江西财经大学 计算机与人工智能学院, 江西 南昌 330013)

通信作者: 方玉明, E-mail: leo.fangyuming@foxmail.com

摘 要: 近年来, 免训练的视频问答模型因其即插即用的特性, 成为轻量级多模态推理研究的热点. 然而, 包含丰富语义信息的高帧率视频往往具备天然的冗余性, 导致在时间维度上存在信息密度与计算效率之间的平衡问题, 传统的采样策略容易受到噪声帧的干扰. 此外, 在复杂的动态场景中, 背景干扰物和局部身体部位等非目标区域会引入空间特征偏差, 严重影响答案生成的可靠性. 为解决以上两个问题, 提出了双重自适应冗余消除 (dual adaptive redundancy elimination, DARE) 框架, 旨在通过时空冗余协同优化机制, 实现免训练范式下视频语义理解精度与答案质量的系统性提升. 首先, 提出一种基于文本-视觉对齐与帧间语义一致的双关联时间采样方法, 通过双向交互推理筛选视频关键帧序列, 并同步剔除与文本语境冲突的冗余帧. 其次, 引入一种动态空间采样方法, 从与提示相关的热力图候选区域中提取最大连通语义区域, 以消除与问题无关的分散区域的干扰, 增强空间特征表达的紧密相关性. 所提方法在 MSVD-QA、MSRVTT-QA、TGIF-QA 和 ActivityNet-QA 等广泛使用的数据集上进行了实验, 并在零样本 (zero-shot) 设定下与 14 个最新模型进行了对比评估. 实验结果表明, 所提方法在使用更少视频特征序列的情况下实现了更具竞争力的性能. 可视化分析进一步验证了该方法在复杂场景中 (如多人交互和细粒度动作识别) 表现出更准确的时空定位能力. 所提出的双重自适应冗余消除框架通过协同优化时空冗余, 在免训练范式下显著提升了视频问答任务的性能, 能够生成准确且高质量的答案, 展现出其在多模态视频理解中的应用潜力.

关键词: 免训练; 视频问答; 双关联时间采样; 动态空间采样; 零样本

中图法分类号: TP37

中文引用格式: 方承炀,朱畅,姜文晖,方玉明,鄢杰斌.面向免训练视频问答的双重自适应冗余消除.软件学报,2026,37(5):1950–1963. http://www.jos.org.cn/1000-9825/7544.htm

英文引用格式: Fang CY, Zhu C, Jiang WH, Fang YM, Yan JB. Dual Adaptive Redundancy Elimination for Training-free Video Question Answering. Ruan Jian Xue Bao/Journal of Software, 2026, 37(5): 1950–1963 (in Chinese). http://www.jos.org.cn/1000-9825/7544.htm

Dual Adaptive Redundancy Elimination for Training-free Video Question Answering

FANG Cheng-Yang, ZHU Chang, JIANG Wen-Hui, FANG Yu-Ming, YAN Jie-Bin

(School of Computing and Artificial Intelligence, Jiangxi University of Finance and Economics, Nanchang 330013, China)

Abstract: In recent years, training-free video question answering (VQA) models have become a research hotspot for lightweight multimodal reasoning due to their plug-and-play nature. However, although high frame rate videos contain rich semantic information, their inherent redundancy leads to a balance problem between information density and computational efficiency in the temporal dimension, with traditional sampling strategies being susceptible to noise frame interference. Furthermore, in complex dynamic scenes, background clutter

* 基金项目: 国家自然科学基金 (62562035, 62441203, 62311530101, 62461028); 江西省重点研发计划 (20252BCE310034); 江西省自然科学基金 (20242BAB23012, 20252BAC200182); 江西省双千计划 (jxsq2023101092); 江西省职业早期青年科技人才培养项目 (20252BEJ730121); 江西财经大学第十九届学生科研课题 (20241126104806366)
本文由“多媒体智能理解与生成”专题特约编辑孙立峰教授、闵巍庆副研究员、马占宇教授、蒋树强研究员、彭宇新教授、田丰研究员、黄庆明教授推荐.
收稿时间: 2025-05-26; 修改时间: 2025-07-11; 采用时间: 2025-09-05; jos 在线出版时间: 2025-09-23
CNKI 网络首发时间: 2025-12-25

and local body parts, as non-target regions, introduce spatial feature bias, significantly affecting the reliability of answer generation. To address these two issues, this study proposes a dual adaptive redundancy elimination (DARE) framework, which aims to systematically improve the accuracy of video semantic understanding and answer quality in the training-free paradigm through a spatiotemporal redundancy collaborative optimization mechanism. First, a dual-relation temporal sampling method is proposed, based on text-visual alignment and inter-frame semantic consistency. This method selects key frame sequences through bidirectional interactive reasoning, while simultaneously eliminating redundant frames that conflict with the text context. Next, a dynamic spatial sampling method is introduced, which extracts the largest connected semantic region from candidate regions in the prompt-related heatmap, aiming to eliminate scattered non-target regions and enhance the compactness of spatial feature representations. Experiments are conducted on widely used datasets, including MSVD-QA, MSRVTT-QA, TGIF-QA, and ActivityNet-QA. The proposed method is evaluated in a zero-shot setting against 14 state-of-the-art models. The results show that the proposed approach achieves competitive performance with significantly fewer video feature sequences. Visual analysis confirms that the proposed method exhibits more accurate spatiotemporal localization abilities in challenging tasks, such as multi-person interactions and fine-grained action recognition in complex scenes. The proposed DARE-VQA framework achieves significant improvements in video question answering performance by collaboratively optimizing spatiotemporal redundancy. It can generate accurate and high-quality answers within the training-free paradigm, demonstrating its potential in multimodal video understanding.

Key words: training-free; video question answering; dual-relation temporal sampling; dynamic spatial sampling; zero-shot

视频问答技术[1–3]旨在赋予机器对多模态视频内容的语义理解与推理能力, 使其能够回答与视频内容相关的问题. 这一任务已成为衡量多模态人工智能系统认知水平的核心标准之一. 尽管针对图像的视觉问答技术已经相对成熟, 但针对视频的问答技术仍面临诸多独特挑战, 主要在于该技术需要同时建模视频中多帧图像的时间动态变化与空间语义关联. 当前主流方法通常依赖于对大规模预训练视觉-语言模型[4] (如 CLIP[5]模型) 的端到端微调策略来应对上述问题, 虽然这些方法在特定基准测试上取得了一定进展, 但是这种计算密集型的优化范式存在两大瓶颈: 1) 处理长视频序列需要高昂的训练成本 (需要在海量视频帧上优化复杂的时序关联模块), 严重限制了模型的可扩展性; 2) 视频内容分布的高度异构性 (如不同的视频在运动节奏、场景切换等方面高度多样), 显著削弱了模型的跨领域泛化能力.

近年来, 免训练的视频理解模型因其即插即用的特性成为新兴的研究方向. 该范式通过冻结预训练的视觉-语言模型, 并设计轻量化适配策略, 将原始视频帧序列转换为与模型兼容的输入表示, 从而避免了高资源开销的训练过程. 例如 IG-VLM[6]和 FreeVA[7]等模型展示了将视频帧转换为符合高性能视觉-语言模型兼容输入格式并应用大模型的潜力. 这些方法通过重构视频帧表示并结合轻量的时间推理模块, 成功减少了传统视频问答模型的计算开销, 同时保持了与之相当的语义理解能力. 然而, 现有的免训练视频问答方法仍面临两个关键挑战.

挑战 1: 时序对齐失准 (temporal misalignment). 高帧率视频的冗余使得时间采样策略易受噪声干扰. 传统时间采样方法, 如均匀采样或双流特征提取 (如 SF-LLaVA[8]), 无法有效处理视频段落中非均匀的语义分布. 近期如 Free Video-LLM[9]虽引入了基于提示引导的帧筛选方法, 但其未能建立文本语义与视频帧间的关系协同推理机制, 尽管提升了一定的帧选择相关性, 但其缺乏对文本语义与视频帧之间关系的协同建模与联合推理, 难以准确捕捉问题所对应的关键帧. 以图 1 中的场景 1 为例, 面对问题“What is a family having?”, Free Video-LLM 能够筛选出 3 张与问题较相关的帧图, 但其中第 3 张帧明显与语义无关, 表现出帧选择偏差. 这种偏差正是由于缺乏对问题上下文与视频语义之间一致性的建模, 导致模型可能选取非关键帧或遗漏关键动作片段, 从而影响最终的问答准确性.

挑战 2: 空间噪声干扰 (spatial noise interference). 在复杂视频场景中, 背景杂物和局部身体部位等无关区域可能污染视觉特征表示. 现有方法通常采用全局池化的方式应对以上问题, 但是这会导致严重的信息损失, 或使用固定尺寸的提示引导裁剪视频帧区域 (如 Free Video-LLM), 难以动态聚焦于与任务相关的语义区域. 这种缺陷在需要精确空间定位的问题中尤为突出. 如图 1 中的场景 2 所示, 面对问题“How many higher officials meet in a room?”, Free Video-LLM 根据问题与图像标记的相关性选取区域中心, 并在其周围构建固定长宽的矩形框. 然而, 该策略将低相关甚至无关的图像区域一并选入, 另外问题相关的区域形状固定尺寸也并不合理, 导致模型生成错误答案. 这表明缺乏精细空间感知机制会使模型在处理目标计数、关系推理等任务时易受干扰区域误导.

本文提出了双重自适应冗余消除框架来应对上述挑战: 1) 针对时序对齐失准问题, 框架设计双重关联采样策

略. 该策略通过联合建模问题语义与视频帧间的关系图, 实现关键帧序列的精准筛选与动态修正, 在剔除冗余帧的同时严格保持时序连贯性. 这一策略直接修复了图 1 场景 1 中 Free Video-LLM 的帧选择偏差, 通过帧间关系的自适应推理消除错误帧, 最终得到时序连贯的关键帧序列. 2) 针对空间噪声干扰问题, 框架引入动态空间采样策略. 该策略从提示相关热力图中提取最大连通区域, 既能有效消除零散噪声区域的干扰, 又能显著提升空间特征的语义相关性. 此设计优化了图 1 场景 2 中 Free Video-LLM 的固定尺寸裁剪缺陷, 通过自适应聚焦问题相关区域, 规避了固定大小裁剪带来的噪声污染, 为精准空间定位任务提供可靠特征输入.

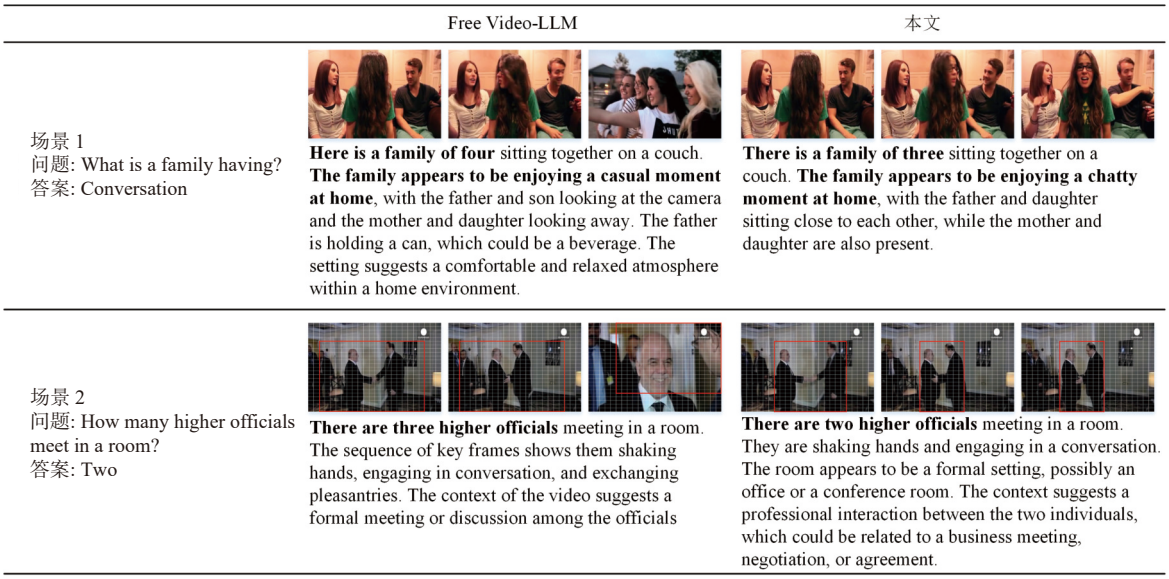


图 1 当前存在的问题与挑战

本文所提方法在 4 个广泛使用且具有挑战性的数据集 (MSVD-QA[10]、MSRVTT-QA[11]、TGIF-QA[12]和 ActivityNet-QA[13]) 上进行了评估. 与 14 个当前最先进的模型进行对比, 该方法在答案准确性和答案质量这两个广泛采用的评估指标上均优于其他对比模型. 可视化分析进一步表明, 该模型能在复杂场景中准确定位关键时空线索. 综上所述, 本文的主要贡献如下.

1) 提出了一种双重关联采样策略, 联合建模问题语义与视频帧间关系, 实现更精准的关键帧筛选与修正. 消融实验验证该方法相较于均匀采样和提示引导采样更加有效.

2) 设计了一种基于最大连通区域提取的动态空间优化机制, 有效消除零散区域的空间噪声干扰, 增强目标区域的语义相关性. 消融实验表明该机制优于全局池化与固定尺寸裁剪策略.

3) 在 4 个公开数据集上进行了大量实验, 结果显示本文方法在答案准确率和答案质量这两个常用指标上超越了 14 个当前最先进模型. 定性分析进一步验证了双重自适应冗余消除框架在复杂视频问答场景中的有效性.

1 相关工作

1.1 图像大语言模型 (Image-LLM)

近年来, 图像大语言模型 (Image-LLM) 取得了显著进展, 主要得益于 CLIP 引入了视觉与语言的共享表示空间以及大规模语言模型 (LLM) 的发展, 如 GPT[14]和 LLaMA[15]的发展. Flamingo[16]通过无缝集成视觉与语言模型, 在少样本任务中展现出强大的能力, BLIP-2[17]则利用轻量级的 Querying Transformer 架构进一步推动了视觉与语言模态的融合. 多模态指令优化进一步提升了图像大语言模型的表现. 例如, LLaVA[18,19]结合 GPT-4 提升了多模态指令数据生成与图文理解能力, InstructBLIP[20]则扩展了 BLIP-2 的功能以适应更广泛的任务范围. MiniGPT-4[21]通

过将视觉编码器与冻结的大语言模型对齐, 增强了图像描述与创意性故事生成的能力. 3D-LLM[22]将三维表示与大语言模型相结合, 为图像语言模型在三维数据处理中的应用打开了新局面. 此外, MM1[23]通过消融实验揭示了影响图像大语言模型性能的关键因素, Ferret[24,25]增强了模型处理物体框/形状模态的能力, 在多模态推理任务中表现更佳. Qwen-VL[26]借鉴了 BLIP-2 中的 Q-Former 结构作为融合层 (fusion layer), 将视觉编码器输出的图像特征与大语言模型的语义空间对齐.

1.2 视频大语言模型 (Video-LLM)

虽然视频与图像同属视觉模态, 但视频作为图像在时间维度上的扩展, 不仅提升了信息密度, 还引入了时序特性. 在图像大语言模型和大语言模型的基础上, 视频大语言模型也迅速发展. FrozenBiLM[27]首创了零样本视频问答方法, 通过使用冻结的双向语言模型, 无需人工标注便达到了当时该任务的先进水平. VideoChat[28]和 Video-LLaMA[29]使用双流架构处理音频与视频信号, 其中 Video-LLaMA 在两个模态中均引入了 Q-Former, 而 VideoChat 则将视频标记特征与字幕感知工具结合, 并利用以视频为中心的多模态指令微调数据集进一步增强模型能力. Video-ChatGPT[30]通过利用预训练视觉编码器提取视频的时空特征, 并将其投射到大语言模型的输入空间, 同时构建了包含 10 万条视频指令问答对的高质量数据集. Chat-UniVi[31]通过动态视觉标记 (token) 统一了图像和视频特征表示; 而 PLLaVA[32]则通过引入了训练后权重融合机制, 以缓解多模态微调过程中模型遗忘的问题. 此外, LLaVA-NeXT-Video[33]通过在视频数据上微调 LLaVA-NeXT[19]提升了其视频理解能力, 其 DPO 版本进一步优化了模型响应, 使其与模型反馈更加一致. TimeChat[34]结合了时间戳感知帧编码器与滑动式视频 Q-Former, 成为面向长视频理解的时间敏感型多模态大语言模型. LLaMA-VID[35]引入了上下文标记与内容标记的双标记机制, 在处理长视频时有效降低了视觉-语言模型的计算开销. Oryx[36]利用可变分辨率编码与动态压缩机制, 能够高效处理不同空间尺度与时间长度的视觉数据, 显著提升了多模态大模型在图像、视频以及多视角 3D 场景理解任务中的适应性与表现能力.

1.3 免训练的视频大语言模型 (training-free Video-LLM)

近年来, 免训练 (training-free) 的视频大语言模型逐渐受到广泛关注, 特别是在如何以最小的调整将现有图像模型架构适配为视频输入方面取得了显著进展. IG-VLM[6]作为该领域的开创性工作, 将视频转换为单个复合图像网格, 使高性能的视觉语言模型 (VLM) 无需在视频数据上训练即可直接应用. FreeVA[7]探索了使用离线图像大语言模型作为免训练视频助手的潜力, 并通过引入无参数的时间聚合模块, 在零样本视频问答任务中取得了强劲表现, 甚至超越了一些经过视频微调的模型. SF-LLaVA[8]采用双流架构, 有效捕捉视频中的空间与时间特征, 在多个视频基准任务上无需额外训练即实现具有竞争力的性能. Free Video-LLM[9]提出了一种提示引导的视觉感知框架, 将空间和时间维度进行解耦, 执行视频帧采样和问题感兴趣区域裁剪, 显著减少视觉标记的数量, 从而降低计算开销. INTP-Video-LLMs[37]则利用视频标记重排序与免训练的上下文窗口扩展方法, 突破固定视频编码器的限制, 增强了模型对长视频内容的理解能力. 此外, 该方法还引入了一种免训练的键值对缓存压缩机制, 在推理阶段有效降低内存开销, 从而实现对长视频的高效处理.

2 方 法

2.1 总 览

本文提出了一种面向免训练视频问答的双重自适应冗余消除方法, 旨在提升免训练大语言模型在视频理解任务中的表现. 整体架构如图 2(a) 所示, 包含以下 4 个核心模块: 1) 视觉编码器, 分别提取视频帧特征和文本特征; 2) 双关联时间采样模块, 利用文本与视频帧的关联性以及视频帧间的内在关系, 实现时间维度的降维采样; 3) 动态空间采样模块, 结合提示信息动态优化视频帧的空间特征表达, 增强区域级别的表征能力; 4) 大语言模型, 接收优化后的视觉标记, 生成与文本标记一致的响应.

为了更清晰地展示方法流程, 首先介绍免训练视频大语言模型的整体框架, 该框架旨在将图像大语言模型 (如 CLIP) 的能力扩展至视频理解任务. 随后, 逐步介绍所提出的双关联时间采样模块和动态空间采样模块两个关键模块的设计理念及其在视频理解中的具体作用.

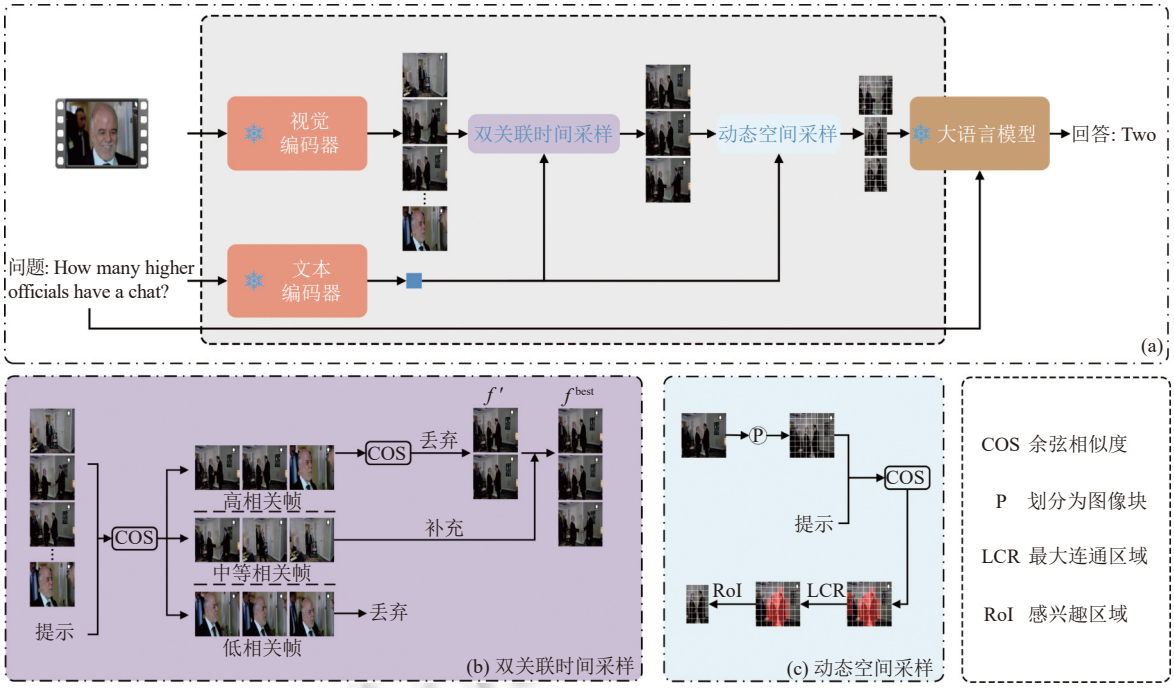


图 2 模型总体框架图

2.2 预备知识: 免训练视频大语言模型

免训练视频大语言模型是一种无需对数据进行进一步微调的技术, 通过基于预训练的图像大语言模型构建视频理解模型. 其核心特点是避免训练过程, 直接利用已有的视觉语言模型进行视频处理.

给定一个视频 $\boldsymbol{V}$, 帧采样器从中均匀选取 T 帧组成图像序列 $\{\boldsymbol{I}_1,\boldsymbol{I}_2,\ldots,\boldsymbol{I}_N\}$. 采样帧随后被送入基于图像的视觉编码器 (如 CLIP-L) 提取视觉特征:

$$\boldsymbol{F}_V = E(\boldsymbol{I}_1,\boldsymbol{I}_2,\ldots,\boldsymbol{I}_N) \tag{1}$$

其中, E 表示视觉编码器, $\boldsymbol{F}_V \in \mathbf{R}^{T\times N\times D}$, T 是从视频中均匀选取的帧数量, N 是每帧视觉标记的数量, D 是标记维度, $\boldsymbol{F}_V$ 组成由所有的帧. 后将视频标注 $\boldsymbol{F}_V$ 和文本标注 $\boldsymbol{F}_T$ 送入 LLM 中以产生响应:

$$\boldsymbol{Y} = LLM([\boldsymbol{F}_V,\boldsymbol{F}_T]) \tag{2}$$

这个简单的流程直观而有效, 被广泛运用[7–9]. 然而, 与图像大语言模型相比, 视频大语言模型的计算复杂度显著增加. 在图像大语言模型中, 模型仅需处理二维图像数据, 其输入维度为 (H,W,C), 计算复杂度主要受空间维度 (H,W) 的影响. 而在视频大语言模型中, 输入数据由多个时间帧组成, 维度扩展为 (T,H,W,C), 这要求模型不仅要处理空间信息, 还需建模时间维度上的依赖关系. 因此, 为了捕捉时间维度上的动态变化, 视频模型通常引入时间建模模块, 如 3D 卷积或时序 Transformer, 这会显著提高计算开销. 例如, 时间维度上的自注意力机制的计算复杂度为 $O(T^2)$, 整体复杂度也从图像模型中的 $O(H^2W^2)$ 增加至 $O(T\cdot H^2W^2)$. 因此, 如何对视频帧进行合理的时间采样和空间降维, 以减少计算量、降低算法复杂度并提升算法效率, 成为提升免训练的视频大语言模型性能的关键.

2.3 双关联时间采样 (dual-relation temporal sampling)

实验动机: 视频通常以高帧速率捕获, 导致相邻帧之间存在大量冗余信息. 对于视频问答任务而言, 需在保留关键信息的前提下减少帧数. 与现有方法 (例如每个视频片段采样一帧[6]、采样稀疏聚合、密集聚合[7]、均匀时间采样[8]、局部聚类[38]和基于提示的时间采样[9]) 不同, 我们提出了一种双关联时间采样策略. 该方法不仅结合了任务或问题的文本上下文, 还能够在推理过程中动态调整图文关联性不匹配的帧. 例如, 在图 1 场景 1 中, 视频展

示了一个家庭的惬意聊天. 对于问题“What is a family having?”, 传统的基于提示的时间采样方法虽然选取了 3 个与问题高度相关的帧, 但第 3 帧并非所需的关键帧. 在这种情况下, 利用视频帧之间的内在语义关系, 从其他候选帧中识别出更合适的替代帧. 该模块通过联合考虑文本与视频帧之间的相关性以及帧与帧之间的语义一致性, 有效解决了时间维度下采样过程中的效率与准确性问题, 从而保证所选信息的完整性与相关性.

设计方案: 在时间维度上, 首先利用提示信息来引导选择与其最相关的时间帧. 具体来说, 使用与视觉编码器 (如 CLIP) 对齐的文本编码器 E 提取问题提示的特征:

$$\boldsymbol{F}_V = E(\boldsymbol{Q}) \tag{3}$$

提示序列 $\boldsymbol{Q}$ 将被表示为一个向量 $\boldsymbol{F}_Q \in \mathbf{R}^D$. 通过视觉编码器处理后, 视频帧的视觉标记不仅捕捉了各帧内部的信息, 还与提示特征处于同一子空间中. 然后, 通过全局平均池化提取视觉特征 $\boldsymbol{F}^{\text{cls}} \in \mathbf{R}^{T\times D}$, 得到每一帧的表示. 接下来, 计算每帧视觉特征与提示特征之间的相关性得分 (即余弦相似度):

$$\cos_i^{\text{IQ}}(\boldsymbol{F}_i^{\text{cls}}, \boldsymbol{F}_Q) = \frac{\boldsymbol{F}_i^{\text{cls}}}{\left\|\boldsymbol{F}_i^{\text{cls}}\right\|_2} \cdot \frac{\boldsymbol{F}_Q}{\left\|\boldsymbol{F}_Q\right\|_2} \tag{4}$$

其中, $i = 1,2,\ldots,T$ 表示帧的时间索引. 根据相似度得分, 将帧 $\boldsymbol{f} \in \mathbf{R}^T$ 按相似度降序排列, 并划分为 k 个与提示最相关帧、t 个中等相关帧以及 $T-k-t$ 个低相关帧 (被丢弃).

在根据提示特征选择出相关帧后, 还需考虑帧与帧之间的关系. 具体地, 计算帧与帧之间的余弦相似度:

$$\cos_{i,j}^{\text{II}}(\boldsymbol{F}_{f_i}^{\text{cls}}, \boldsymbol{F}_{f_j}^{\text{cls}}) = \frac{\boldsymbol{F}_{f_i}^{\text{cls}}}{\left\|\boldsymbol{F}_{f_i}^{\text{cls}}\right\|_2} \cdot \frac{\boldsymbol{F}_{f_j}^{\text{cls}}}{\left\|\boldsymbol{F}_{f_j}^{\text{cls}}\right\|_2} \tag{5}$$

接着, 在前 k 个高相关帧中, 识别出与其他帧相似度最低的帧. 我们计算每个帧与其他 $k-1$ 个帧的相似度总和:

$$\boldsymbol{s}_i^{\text{throw}} = \sum_{j=1,j\neq i}^{k} \cos_{i,j}^{\text{II}}(\boldsymbol{F}_{f_i}^{\text{cls}}, \boldsymbol{F}_{f_j}^{\text{cls}}) = \frac{\boldsymbol{F}_{f_i}^{\text{cls}}}{\left\|\boldsymbol{F}_{f_i}^{\text{cls}}\right\|_2} \cdot \sum_{j=1,j\neq i}^{k} \frac{\boldsymbol{F}_{f_j}^{\text{cls}}}{\left\|\boldsymbol{F}_{f_j}^{\text{cls}}\right\|_2} \tag{6}$$

然后, 丢弃其中 m 个与其他帧间相似度最低的帧, 剩余帧记为 $\boldsymbol{f}'$, 数量为 $k-m$. 接下来, 从中等相关的 t 帧中补充 m 帧 (编号为 $\boldsymbol{f}_i$ $(i = k+1,k+2,\ldots,k+t)$), 要求这些帧不与 $\boldsymbol{f}'$ 中的帧重复. 选择方式如下, 计算候选帧与 $\boldsymbol{f}'$ 中所有帧的相似度总和:

$$\boldsymbol{s}_i^{\text{fill}} = \sum_{j=1}^{k-m} \cos_{i,j}^{\text{II}}(\boldsymbol{F}_{f_i}^{\text{cls}}, \boldsymbol{F}_{f'_j}^{\text{cls}}) = \frac{\boldsymbol{F}_{f_i}^{\text{cls}}}{\left\|\boldsymbol{F}_{f_i}^{\text{cls}}\right\|_2} \cdot \sum_{j=1}^{k-m} \frac{\boldsymbol{F}_{f'_j}^{\text{cls}}}{\left\|\boldsymbol{F}_{f'_j}^{\text{cls}}\right\|_2} \tag{7}$$

最后, 依据 $\boldsymbol{s}_i^{\text{fill}}$ 的得分, 从候选帧中选择 m 个与 $\boldsymbol{f}'$ 不重复的帧, 得到最终选中的关键帧集合 $\boldsymbol{f}^{\text{best}}$. 如图 1 场景 1 中, 使用双关联时间采样策略后, 对于问题“What is a family having?”, 本文的方法可以有效地保证筛选出的视频帧之间保持语义相关性, 同时与问题保持语义一致性.

2.4 动态空间采样 (dynamic spatial sampling)

实验动机: 视频帧中包含丰富的空间信息, 但这些信息往往也包含噪声和冗余. 在视觉问答任务中, 不同的问题需要关注不同的关键区域. 现有方法通常采用最大池化提取全局特征[7], 或基于提示词进行固定大小的空间采样[9], 但这些方式可能无法很好地适应不同的问题需求. 为了解决这一问题, 我们提出了一种动态空间采样策略. 如图 1 场景 2 所示, 对于问题“How many higher officials meet in a room?”, 基于提示词的固定空间采样方法可以在视频帧中突出相关区域, 但左侧分散的区域并不是目标区域, 甚至可能对答案的准确性产生负面影响. 因此, 我们进一步精细化采样策略, 通过剔除无关区域, 以获得更精确的关注区域. 本模块分析与提示高度相关的区域, 提取最大连通区域, 并对视频帧的空间特征表示进行动态优化.

设计方案: 为适应不同任务, 模型需要聚焦于特定空间区域以生成有效回答. 为此, 提出利用提示信息引导视频中的兴趣区域 (region of interest, RoI) 选择. 具体而言, 对于单帧视频的视觉特征标记 $\boldsymbol{F}_{V_i} \in \mathbf{R}^{N\times D}$, 我们将其重塑为空间形式 $\boldsymbol{F}^{\text{IR}} \in \mathbf{R}^{H\times W\times D}$, 其中 $H\times W$ 表示特征图的空间尺寸. 对于空间位置 (h, w) 上的特征向量, 我们计算其与提

示特征的相关性得分 (余弦相似度), 公式如下:

$$\cos_{h,w}^{\mathrm{PQ}}(\boldsymbol{F}_{h,w}^{\mathrm{IR}},\boldsymbol{F}_Q)=\frac{\boldsymbol{F}_{h,w}^{\mathrm{IR}}}{\left\|\boldsymbol{F}_{h,w}^{\mathrm{IR}}\right\|_2}\cdot\frac{\boldsymbol{F}_Q}{\left\|\boldsymbol{F}_Q\right\|_2} \tag{8}$$

其中, $1\leqslant h\leqslant H$ 且 $1\leqslant w\leqslant W$. 根据这些相似度得分, 选取得分最高的前 n 个标记, 其对应的位置集合为: $\{(h_1,w_1),(h_2,w_2),\ldots,(h_n,w_n)\}$, 其中 $n=\alpha HW$, $0\leqslant\alpha\leqslant 1$. 这些位置用于确定兴趣区域的中心坐标及形状.

在特征图上对这些高相关标记的位置分布进行分析, 提取出最大连通区域 e 的中心坐标, 通过该连通区域内所有标记的位置平均计算得出:

$$h_c=\frac{1}{n_e}\sum_{i=1}^{n_e}h_{e_i} \tag{9}$$

$$w_c=\frac{1}{n_e}\sum_{i=1}^{n_e}w_{e_i} \tag{10}$$

其中, n_e 为最大连通区域中的标记数量, e_i 表示第 i 个标记在特征图 $\boldsymbol{F}^{\mathrm{IR}}$ 中的位置. 同时, 兴趣区域的高与宽由该连通区域的最大垂直与水平跨度决定:

$$H'=H_{\max}^{e} \tag{11}$$

$$W'=W_{\max}^{e} \tag{12}$$

其中, $H_{\max}^{e}$、$W_{\max}^{e}$ 分别是连通区域在垂直和水平方向的最大跨度. 根据计算得到的中心坐标和区域大小, 我们从帧的特征图中裁剪出兴趣区域, 并将其重新调整为一组标记向量, 作为视频与提示词对齐的表示.

最后, 这些紧凑的视觉标记与文本提示标记结合, 作为大语言模型的输入. 通过这种方法, 免训练的视频大模型可以使用更少的标记表示视频, 从而显著提升推理效率. 如图 1 场景 2 中, 使用动态空间采样策略后, 模型可以关注不同尺寸的感兴趣区域, 更加精准地获取与问题有关的视觉信息.

3 实　验

3.1 实验准备

1) 基准评测: 本文对提出的方法在多个开放式视频问答基准数据集上进行了全面评估. 开放式视频问答基准数据集包括 MSVD-QA[10]、MSRVTT-QA[11]、TGIF-QA[12]和 ActivityNet-QA[13], 主要用于评估模型生成自由形式答案的能力.

2) 评估指标: 在评估过程中, 本文采用了 GPT 辅助评测方法, 分别从答案的准确性 (答案正确或错误) 和质量 (评分范围为 0–5 分) 两个维度进行量化评估. 为保证评估结果的公平性与一致性, 本文遵循 FreeVA、SF-LLaVA 和 Free Video-LLM 等工作的方法, 统一使用 GPT-3.5-Turbo-0125 作为评估模型.

3) 实验细节: 遵循 SF-LLaVA 和 Free Video-LLM 等工作, 实验中本文采用 LLaVA-v1.6 模型, 分别选用了 7B 和 34B 两种规模模型作为基础框架. 视觉和文本编码器均采用 OpenAI 的 CLIP-L-14 模型, 预训练权重从 HuggingFace 获取. 输入视频被统一调整为 336×336 的分辨率, 每帧图像被划分为 24×24 的标记网格. 为提升模型对长序列的处理能力, 本文将 LLaVA-v1.6 中旋转位置编码的缩放因子设为 2, 以支持最长为 8 192 标记的上下文长度. 所有实验使用 PyTorch 框架实现, 在保持基础模型预训练权重不变的前提下仅对超参数进行了适度调整.

3.2 主要实验结果

开放式视频问答任务的设置方式与 Free Video-LLM[9]保持一致, 模型选取 3 帧双关联时间采样帧特征、5 帧均匀时间采样帧特征及 50 帧密集聚合采样帧特征, 其中双关联时间采样帧特征与均匀时间采样帧还会经动态空间采样处理以降低冗余噪声. 开放式视频问答任务的实验结果如表 1 和表 2 所示. 表中每个单元格中的两个数字分别表示“准确率/分数”. 加粗的数字表示在免训练方法中性能最佳的结果. 表 2 的大语言模型均采用 CLIP-L 视觉编码器. 我们提出的模型在零样本视频问答任务中展现出了显著的性能提升. 具体而言, 通过利用不同规模的大语言模型, 我们的模型在多个基准测试中均优于现有的零样本方法.

表 1 各类视频语言模型在 70 亿参数开放式视频问答任务上的性能对比

类别	模型	大模型规模 (B)	视觉编码器	Open-ended VideoQA (准确率 (%)/分数)			
				MSVD-QA	MSRVTT-QA	ActivityNet-QA	TGIF-QA
训练微调过的模型	Video-LLaMA[29]	7	CLIP-G	51.6/2.5	29.6/1.8	12.4/4.1	—
	Video-LLaVA[39]	7	ViT-L	70.1/3.9	59.2/3.5	45.3/3.3	70.0/4.0
	Vista-LLaMA[40]	7	CLIP-G	65.3/3.6	56.3/3.5	48.3/3.3	—
	VideoChat[28]	7	CLIP-G	56.3/2.8	45.0/2.5	26.5/2.2	34.4/2.3
	VideoChat2[41]	7	CLIP-L	75.3/4.1	59.3/3.9	49.1/3.4	—
	Video-ChatGPT[30]	7	CLIP-L	74.3/3.9	62.3/3.7	50.2/2.3	—
	VideoLLaMA 2[42]	7	CLIP-L	70.9/3.8	—	—	—
	PLLaVA[32]	7	CLIP-L	76.6/4.1	62.0/3.5	56.3/3.5	77.5/4.1
免训练模型	FreeVA[7]	7	CLIP-L	73.8/4.1	60.0/3.5	51.2/3.5	—
	DeepStack-L[43]	7		76.0/4.0	—	49.3/3.1	—
	IG-VLM[6]	7		78.3/4.4	63.7/3.4	54.3/3.4	73.0/4.0
	SF-LLaVA[8]	7		79.1/4.0	65.8/3.4	**55.5**/3.4	78.7/4.2
	Free Video-LLM[9]	7		78.2/4.0	65.6/3.6	54.8/3.4	77.8/4.1
	本文方法	7		**79.4**/4.1	**66.5**/3.6	54.9/3.4	**78.9**/4.1

表 2 各类视频语言模型在 340 亿或更大参数开放式视频问答任务上的性能对比

类别	模型	大模型规模 (B)	Open-ended VideoQA (准确率 (%)/分数)			
			MSVD-QA	MSRVTT-QA	ActivityNet-QA	TGIF-QA
训练微调过的模型	VideoLLaMA 2[42]	46.7	70.5/3.8	—	50.3/3.4	—
	LLaVA-NeXT-Video[33]	34	—	—	58.8/3.4	—
	LLaVA-NeXT-DPO[33]	34	—	—	64.4/3.6	—
免训练模型	LLaVA-NeXT-Image[33]	34	73.8/4.1	60.0/3.5	51.2/3.5	—
	IG-VLM (LLaVA-v1.6)[6]	34	79.6/4.1	62.4/3.5	58.4/3.5	79.1/4.2
	SF-LLaVA-34B[8]	34	79.3/4.1	67.0/3.7	58.8/3.5	80.2/4.3
	Free Video-LLM[9]	34	79.2/4.1	67.2/3.7	58.9/3.5	79.8/4.3
	本文方法	34	**79.7**/4.1	**67.2**/3.7	**59.0**/3.5	**80.9**/4.3

在使用 7B 规模大语言模型时, 模型在视频时序采样和自适应空间采样方面取得显著改进, 这得益于视频帧与提示信息之间以及视频帧与整段视频之间的高效关联建模. 例如, 相较于 Free Video-LLM 方法, 我们的模型在 MSVD-QA 数据集上准确率提升了 1.2%, 达到了 79.4% 的准确率, 同时答案质量得分也从 4.0 提升到了 4.1; 在 MSRVTT-QA 基准测试上准确率提升 0.9%, 达到了 66.5% 的准确率. 此外, 通过对视频帧与提示词及整体视频之间的强关联建模, 我们的方法在快速路径 (fast pathway) 中实现了高效筛选与优化. 与 SF-LLaVA 相比, 该方法在 TGIF-QA 上准确率提升了 0.1%, 达到了 78.9%. 由于我们的方法在零样本视频问答任务中具备强大的知识迁移能力, 因此该方法在不需要额外训练的前提下, 依然表现出与经过训练的视频问答模型相媲美的性能.

在使用 34B 规模 LLM 时, 本文所提的方法在多个任务上进一步超越其他最先进的方法. 在 MSVD-QA 数据集上与 Free Video-LLM 方法对比, 准确率提升了 0.5%, 达到了 79.7%. 在 MSRVTT-QA 数据集上准确率与当前最先进的方法持平. 在 ActivityNet-QA 数据集上准确率达到 59.0%. 在 TGIF-QA 数据集上, 相比于最先进的 SF-LLaVA-34B 方法, 准确率提升了 0.7%.

通过使用不同规模的大模型作为基础框架在多个视频问答数据集上进行公平对比实验, 我们的方法在多个数据集上达到最先进的水平, 这证明了我们所提方法的有效性.

3.3 消融实验

为验证本文模型中各个模块的有效性和高效性, 本文在 MSVD-QA 数据集上进行了消融实验, 所有实验均基

于 7B 规模的 LLaVA-v1.6 模型进行.

3.3.1 双关联时间采样的有效性

为评估双关联时间采样方法在时间维度的有效性, 设计了一系列消融实验, 结果如表 3 所示. 主要对比的几种方法包括: 均匀时间采样 (基线方法)、基于提示引导的时间采样、双关联时间采样、均匀+提示引导采样组合、均匀+双关联采样组合. 现有相关模型的时间采样方案可作为参考, 其中 IG-VLM 采用 6 帧均匀采样, FreeVA 采用 50 帧均匀采样后对每一帧特征进行密集聚合, SF-LLaVA 采用 10 帧均匀采样与 50 帧密集聚合, Free Video-LLM 采用 3 帧均匀采样、5 帧提示引导时间采样和 50 帧密集聚合. 为保证消融实验的公平性, 实验首先在相同的 3 帧采样条件下比较不同方法的性能, 随后再比较不同方法组合后的性能变化情况.

表 3 双关联时间采样策略的消融实验

方法	高相关帧数	中等相关帧数	调整帧数	最终帧数	MSVD-QA (%)
仅均匀时间采样	—	—	—	3	71.7
仅提示引导时间采样	3	—	—	3	75.0
双关联时间采样	3	7	1	3	75.6
均匀+提示引导时间采样	3	7	1	6	76.9
均匀+双关联时间采样	3	7	1	6	77.3

首先, 均匀时间采样方法选取了 3 帧并使用 864 个视觉标记, 在 MSVD-QA 任务中实现了 71.7% 的准确率, 作为对比基线. 接着, 提示引导的时间采样方法同样选取 3 帧和 864 个视觉标记, 但其准确率提高至 75.0%, 表明结合提示信息来引导时间采样过程能够有效提升模型性能. 随后, 对提示引导时间采样方法进行了进一步优化. 在 3 帧中剔除图像关联性最弱的一帧, 并从提示相关性较高的 3 帧和中等相关性的 7 帧中筛选出相似度最高的非重复帧. 该优化方案使得准确率进一步提高至 75.6%, 表明双关联时间采样能够根据问题需求, 选取更合适的视频帧, 从而提升问题解答的准确性. 最后, 将均匀时间采样与双关联时间采样相结合, 选取 6 帧并使用 1 728 个视觉标记, 最终得分提升至 77.3%. 进一步提升了模型性能, 验证了混合策略的有效性.

3.3.2 动态空间采样的有效性

为验证动态空间采样在空间维度的作用, 设计了一系列消融实验, 结果如表 4 所示. 实验包含 4 种配置: 仅使用均匀时间采样、使用双关联时间采样、在双关联时间采样基础上进行提示引导的空间采样和在双关联时间采样基础上进行动态空间采样, 具体结果如下.

表 4 动态空间采样策略的消融实验

方法	RoI比率	最终帧数	平均视觉标记	MSVD-QA (%)
均匀时间采样	—	3	864	71.7
双关联时间采样	—	3	864	75.6
双关联时间采样+提示引导的空间采样	0.6	3	513	74.9
双关联时间采样+动态空间采样	0.6	3	500	75.7

在使用均匀时间采样的方法时, 最终选取了 3 帧视频帧, 一共使用了 864 个视觉标记, 达到了 71.7% 的准确率. 在使用双关联时间采样方法时, 同样选取 3 帧视频帧以及使用 864 个视觉标记, 最终准确率为 75.6%. 在同时使用双关联时间和动态空间采样时, 将 RoI 裁剪比率设为 0.6, 仍选取 3 帧视频帧, 视觉标记减少至 500, 准确率提升至 75.7%. 与“双关联时间采样+提示引导的空间采样”方案相比, 提出的方案不仅减少了 13 个平均视觉标记, 准确率还提升了 0.8%. 结果表明, 双关联时间采样结合动态空间采样不仅显著减少了视觉标记数量, 同时也保持了整体性能, 充分验证了这两种模块的有效性及协同作用.

3.3.3 RoI 比率消融实验

为了进一步分析动态空间采样中 RoI 比率对模型性能的影响, 该实验仅使用双关联时间采样以及动态空间采

样, 使用 3 帧视频帧, 选用 Free Video-LLM 模型结果作为对比基线, 实验结果如图 3 所示. 观察结果如下: 当 RoI 比率为 0.4 时, 视觉标记数约为 360 个, 在 MSVD-QA 数据集上准确率为 73.1%, 而 Free Video-LLM 模型能够达到 74.1% 的准确率, 这是因为 RoI 比率很小的情况下动态空间采样关注视频的位置是变化的, 而 Free Video-LLM 模型以视频帧中心点为筛选区域中心; 当增加 RoI 比率到 0.5 时, 视觉标记数约为 408 个, 准确率提升至 74.4%, 超过了相同设置的 Free Video-LLM 模型性能; 继续增加 RoI 比率到 0.6 及以上时, 尽管视觉标记数量增加, 但是模型性能趋于稳定, 准确率在 75.5%–75.8% 之间浮动. 这表明, 更大的 RoI 比率虽能覆盖更多的空间信息, 但当 RoI 比率超过一定阈值 (如 0.6) 后, 性能提升的边际效益变小. 综上所述, 实验充分表明合理优化 RoI 比率对于提升视频问答性能至关重要, 并进一步验证了我们提出的提示引导采样策略在提升视觉标记效率和模型准确率方面的有效性.

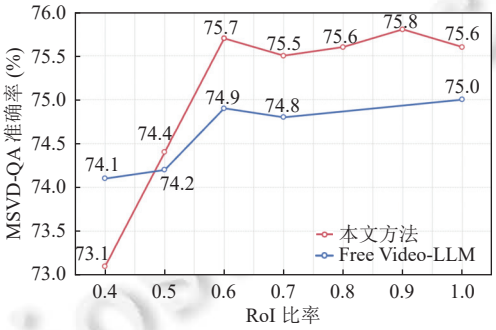


图 3 RoI 比率的影响

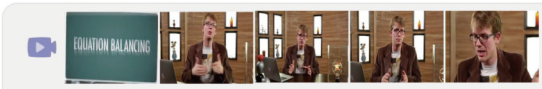
3.4 视频问答可视化分析

前文图 1 展示了本文提出的双关联时间采样与动态空间采样方法, 与 Free Video-LLM 方法在两个典型场景中的对比示例. 在场景 1 中, 利用双关联时间采样策略, 模型能够有效筛选出与提示语高度相关的 3 帧图像, 剔除原方法中误选的第 3 帧语义不符图像, 并从其他相关帧中补充选取更具代表性的图像, 从而提升时间采样的语义一致性. 在场景 2 中, 借助动态空间采样机制, 模型在提示引导下首先定位出相关的图像标记区域, 并进一步基于语义相关性动态划分空间区域, 聚焦于与任务最相关的视觉区域, 从而得到更精确的空间采样结果.

我们还展示了一些来自视觉问答任务复杂场景的可视化案例, 以更直观地体现我们方法的优势, 如后文图 4 所示. 具体来说: 在图 4(a) 中, 本文模型采用了双关联采样策略, 整合了视频帧中的关键信息, 进一步优化了视觉信息与文本信息的对齐, 准确捕捉到了包含黑板的视频帧以及后续视频帧中男子的姿势、办公环境以及正在进行的讲解动作. 在图 4(b) 中, 模型的动态空间采样策略优化了对视频帧中关键区域的关注, 精准提取了手机的手势和方向盘的握法, 并通过对视频上下文的分析, 成功推断出该男子正在演示使用语音识别控制技术进行电话通信. 在图 4(c) 中, 模型高精度地选择了与问题相关的关键帧, 提供了更具上下文准确性的推理, 成功推断出马正在训练场或围栏内奔跑, 突出展现了马的运动细节. 在图 4(d) 中, 模型对情感和动作均作出精准判断, 给出了丰富且细腻的情感描述, 展示了其在处理复杂场景时的优势.

3.5 算法复杂度分析

表 5 呈现了各模型在免训练视频问答任务中的算法复杂度对比结果, 仅包含推理阶段分析, 因该任务无需训练过程. 表中推理速度指标定义为处理所有视频的总时间除以视频总数, 数值越低表明模型推理效率越高. 从结果可见, 在所有免训练视频问答模型中, FreeVA 以 2.72 s 的推理速度表现最优, IG-VLM 以 2.94 s 次之, 而本文方法的推理速度为 3.64 s, 处于中等水平. 这一结果与本文方法的设计理念密切相关, 通过双关联时间采样实现时间维度的精确筛选与修正, 同时借助动态空间采样完成空间维度的冗余降维, 在有效提升任务性能的前提下, 保持了模型推理复杂度的可控性.



问题: what is a man doing?答案: explain

我们的描述: a man is seen sitting at a desk with a laptop, talking and gesturing with his hands. **He appears to be engaged in a discussion or presentation, possibly explaining a concept or demonstrating something on the laptop.** The setting suggests an office or a workspace environment. **(score:4/5)**

Free Video-LLM 的描述: a man is seen in a room with a desk, a laptop, and various decorative items. He appears to be engaged in a conversation or presentation, as he is gesturing with his hands and speaking. The context suggests that the man might be discussing a topic related to the content displayed on the laptop screen. The presence of the desk and the laptop indicates that the setting could be an office or a workspace. **(score:1/5)**

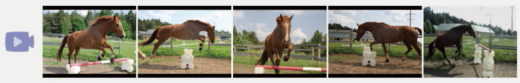
(a) 场景 1: 男子伏案讲解

问题: what does a man using the hands free voice command in his car to make?答案: phone

我们的描述: **a man is seen using a hands-free voice command in his car to make a call**. The sequence of key frames shows the man holding the steering wheel and talking on the phone. The context of the video suggests that the man is using voice command technology to operate his car's infotainment system and make a call without having to physically interact with the system. The video likely demonstrates the convenience and safety benefits of using voice command technology while driving. **(score:5/5)**

Free Video-LLM 的描述: a man is seen using a hands-free voice command in his car to make a sandwich. The sequence of key frames shows the man holding a sandwich in his hand, and the voice command is activated, suggesting that he is using the car's voice command system to prepare the sandwich. The context of the video is a demonstration of the car's voice command system, showcasing its functionality and convenience for the user. **(score:1/5)**

(b) 场景 3: 驾车语音通话

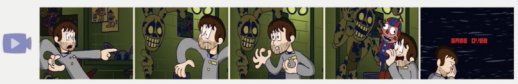


问题: what does a horse run in?答案: field

我们的描述: **the horse is running through a fenced-in area, possibly a corral or a training course.** The horse is seen jumping over a small obstacle, which could be a hurdle or a barrier, indicating that it might be participating in a horse riding or training activity. The setting suggests that the horse is being exercised or trained for equestrian events or activities. **(score:4/5)**

Free Video-LLM 的描述: the horse is running on a track with a red and white pole in the middle. The horse is jumping over the pole, which is part of an obstacle course. The scene captures the horse in action, showcasing its agility and athleticism as it navigates the course. The setting appears to be an outdoor equestrian facility, possibly a training ground or a competition venue. **(score:1/5)**

(c) 场景 3: 马匹场地奔跑



问题: what are cartoon characters doing?答案: interact

我们的描述: there is a scene featuring a man in a room with a cartoon character on the wall. The man is pointing at the character, and the character is smiling. **The overall content of the video seems to be a humorous or lighthearted interaction between the man and the cartoon character.** The context suggests that the man might be engaging with the character in a playful or imaginative way, possibly as part of a storytelling or role-playing scenario. **(score:5/5)**

Free Video-LLM 的描述: there are two cartoon characters, one with a smiling face and the other with a frowning face. The smiling character is pointing to the right, while the frowning character is looking up. They are in a room with a brick wall and a window. The overall content of the video seems to be a humorous or lighthearted scene involving the two characters. **(score:1/5)**

(d) 场景 4: 卡通形象互动

图 4 零样本视频问答可视化及对比

表 5 算法复杂度分析表

模型	推理时间 (s)↓
IG-VLM[6]	2.94
FreeVA[7]	**2.72**
SF-LLaVA[8]	3.75
Free Video-LLM[9]	3.47
本文方法	3.64

注: 加粗字体表示最优结果, "↓"表示数值越低越好

4 总 结

在免训练视频问答 (VQA) 任务中, 存在两个主要挑战: 时序对齐失准和空间噪声干扰. 为了解决这些问题, 我们提出了一种基于双重自适应冗余消除的免训练视频理解框架. 通过构建文本语义与帧间关系的联合推理机制,

实现了关键帧序列的自适应选择和误差修正, 既减少了计算冗余, 又保持了时间动态的完整性. 此外, 提出的动态空间采样策略有效抑制了空间噪声干扰, 增强了目标区域的语义相关性. 我们在 4 个广泛使用的公开数据集上进行了大量实验. 结果表明, 在定量分析中, 该方法在准确率和质量指标上显著优于 14 个最新比较模型. 可视化结果展示了模型在复杂场景中 (如多人交互和细粒度动作) 依然能准确定位时空关键信息. 然而, 现有方法在处理长视频时仍面临视频理解的瓶颈. 未来工作将探索时序跳跃采样策略, 以进一步提升模型对视频叙事结构的理解能力.

References

[1] Hu JX, Meng ZH. Research on video question answering based on video description and reading comprehension. Application Research of Computers, 2021, 38(12): 3781–3785 (in Chinese with English abstract). [doi: 10.19734/j.issn.1001-3695.2021.04.0152]

[2] Yao X, Gao JY, Xu CS. Self-supervised graph contrastive learning for video question answering. Ruan Jian Xue Bao/Journal of Software, 2023, 34(5): 2083–2100 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6775.htm [doi: 10.13328/j.cnki.jos.006775]

[3] Zhao EY, Song N, Nie J, Wang X, Zheng CY, Wei ZQ. Scale-guided fusion inference network for remote sensing visual question answering. Ruan Jian Xue Bao/Journal of Software, 2024, 35(5): 2133–2149 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/7025.htm [doi: 10.13328/j.cnki.jos.007025]

[4] Ying J, Zhang ZD, Gao YH, Yang ZW, Li L, Xiao M, Sun YQ, Yan CG. Survey on vision-language pre-training. Ruan Jian Xue Bao/Journal of Software, 2023, 34(5): 2000–2023 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6774.htm [doi: 10.13328/j.cnki.jos.006774]

[5] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G, Sutskever I. Learning transferable visual models from natural language supervision. In: Proc. of the 38th Int'l Conf. on Machine Learning. PMLR, 2021. 8748–8763.

[6] Kim W, Choi C, Lee W, Rhee W. An image grid can be worth a video: Zero-shot video question answering using a VLM. IEEE Access, 2024, 12: 193057–193075. [doi: 10.1109/ACCESS.2024.3517625]

[7] Wu WH. FreeVA: Offline MLLM as training-free video assistant. arXiv:2405.07798, 2024.

[8] Xu MZ, Gao MF, Gan Z, Chen HY, Lai ZF, Gang HM, Kang K, Dehghan A. SlowFast-LLaVA: A strong training-free baseline for video large language models. arXiv:2407.15841, 2024.

[9] Han K, Guo JY, Tang YH, He W, Wu EH, Wang YH. Free Video-LLM: Prompt-guided visual perception for efficient training-free video LLMs. arXiv:2410.10441, 2024.

[10] Chen D, Dolan W. Collecting highly parallel data for paraphrase evaluation. In: Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland: ACL, 2011. 190–200.

[11] Xu J, Mei T, Yao T, Rui Y. MSR-VTT: A large video description dataset for bridging video and language. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 5288–5296. [doi: 10.1109/CVPR.2016.571]

[12] Jang Y, Song Y, Yu Y, Kim Y, Kim G. TGIF-QA: Toward spatio-temporal reasoning in visual question answering. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 1359–1367. [doi: 10.1109/CVPR.2017.149]

[13] Heilbron FC, Escorcia V, Ghanem B, Niebles JC. ActivityNet: A large-scale video benchmark for human activity understanding. In: Proc. of the 2015 IEEE Conf. on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 961–970. [doi: 10.1109/CVPR.2015.7298698]

[14] Brown TB, Mann B, Ryder N, *et al*. Language models are few-shot learners. In: Proc. of the 34th Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 1877–1901.

[15] Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, Rozière B, Goyal N, Hambro E, Azhar F, Rodriguez A, Joulin A, Grave E, Lample G. LLaMA: Open and efficient foundation language models. arXiv:2302.13971, 2023.

[16] Alayrac JB, Donahue J, Luc P, *et al*. Flamingo: A visual language model for few-shot learning. In: Proc. of the 36th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022. 23716–23736.

[17] Li JN, Li DX, Savarese S, Hoi S. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: Proc. of the 40th Int'l Conf. on Machine Learning. Honolulu: JMLR.org, 2023. 19730–19742.

[18] Liu HT, Li CY, Wu QY, Lee YJ. Visual instruction tuning. In: Proc. of the 37th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2023. 34892–34916.

[19] Liu HT, Li CY, Li YH, Li B, Zhang YH, Shen S, Lee YJ. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge. 2024. https://llava-vl.github.io/blog/2024-01-30-llava-next/

[20] Dai WL, Li JN, Li DX, Tiong AMH, Zhao JQ, Wang WS, Li BY, Fung P, Hoi S. InstructBLIP: Towards general-purpose vision-language

models with instruction tuning. In: Proc. of the 37th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2023. 49250–49267.

[21] Zhu DY, Chen J, Shen XQ, Li X, Elhoseiny M. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. arXiv:2304.10592, 2023.

[22] Hong YN, Zhen HY, Chen PH, Zheng SH, Du YL, Chen ZF, Gan C. 3D-LLM: Injecting the 3D world into large language models. In: Proc. of the 37th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2023. 20482–20494.

[23] McKinzie B, Gan Z, Fauconnier JP, *et al*. MM1: Methods, analysis & insights from multimodal LLM pre-training. arXiv:2403.09611, 2024.

[24] You HX, Zhang HT, Gan Z, Du XZ, Zhang BW, Wang ZR, Cao LL, Chang SF, Yang YF. Ferret: Refer and ground anything anywhere at any granularity. arXiv:2310.07704, 2023.

[25] Zhang HT, You HX, Dufter P, Zhang BW, Chen C, Chen HY, Fu TJ, Wang WY, Chang SF, Gan Z, Yang YF. Ferret-v2: An improved baseline for referring and grounding with large language models. arXiv:2404.07973, 2024.

[26] Bai JZ, Bai S, Yang SS, Wang SJ, Tan SN, Wang P, Lin JY, Zhou C, Zhou JR. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv:2308.12966, 2023.

[27] Yang A, Miech A, Sivic J, Laptev I, Schmid C. Zero-shot video question answering via frozen bidirectional language models. In: Proc. of the 36th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022. 124–141.

[28] Li KC, He Y, Wang Y, Li YZ, Wang WH, Luo P, Wang YL, Wang LM, Qiao Y. VideoChat: Chat-centric video understanding. arXiv: 2305.06355, 2024.

[29] Zhang H, Li X, Bing LD. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In: Proc. of the 2023 Conf. on Empirical Methods in Natural Language Processing: System Demonstrations. Singapore: ACL, 2023. 543–553. [doi: 10.18653/v1/2023.emnlp-demo.49]

[30] Maaz M, Rasheed H, Khan S, Khan F. Video-ChatGPT: Towards detailed video understanding via large vision and language models. In: Proc. of the 62nd Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers). Bangkok: ACL, 2024. 12585–12602. [doi: 10.18653/v1/2024.acl-long.679]

[31] Jin P, Takanobu R, Zhang WC, Cao XC, Yuan L. Chat-UniVi: Unified visual representation empowers large language models with image and video understanding. In: Proc. of the 2024 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024. 13700–13710. [doi: 10.1109/CVPR52733.2024.01300]

[32] Xu L, Zhao YL, Zhou DQ, Lin ZJ, Ng SK, Feng JS. PLLaVA: Parameter-free LLaVA extension from images to videos for video dense captioning. arXiv:2404.16994, 2024.

[33] Zhang YH, Li B, Liu HT, Lee YJ, Gui LK, Fu D, Feng JS, Liu ZW, Li CY. LLaVA-NeXT: A strong zero-shot video understanding model. 2024. https://llava-vl.github.io/blog/2024-04-30-llava-next-video/

[34] Ren SH, Yao LL, Li SC, Sun X, Hou L. TimeChat: A time-sensitive multimodal large language model for long video understanding. In: Proc. of the 2024 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024. 14313–14323. [doi: 10.1109/CVPR52733.2024.01357]

[35] Li YW, Wang CY, Jia JY. LLaMA-VID: An image is worth 2 tokens in large language models. In: Proc. of the 18th European Conf. on Computer Vision. Milan: Springer, 2025. 323–340. [doi: 10.1007/978-3-031-72952-2_19]

[36] Liu ZY, Dong YH, Liu ZW, Hu W, Lu JW, Rao YM. Oryx MLLM: On-demand spatial-temporal understanding at arbitrary resolution. arXiv:2409.12961, 2025.

[37] Shang YZ, Xu BX, Kang WT, Cai M, Li YH, Wen ZH, Dong Z, Keutzer K, Lee YJ, Yan Y. Interpolating Video-LLMs: Toward longer-sequence LMMs in a training-free manner. arXiv:2409.12963, 2024.

[38] Guo D, Yao ST, Wang H, Wang M. Embedding VLAD in Transformer for video question answering. Chinese Journal of Computers, 2023, 46(4): 671–689 (in Chinese with English abstract). [doi: 10.11897/SP.J.1016.2023.00671]

[39] Lin B, Ye Y, Zhu B, Cui JX, Ning MN, Jin P, Yuan L. Video-LLaVA: Learning united visual representation by alignment before projection. In: Proc. of the 2024 Conf. on Empirical Methods in Natural Language Processing. Miami: ACL, 2024. 5971–5984. [doi: 10.18653/v1/2024.emnlp-main.342]

[40] Ma F, Jin X, Wang H, Yang J, Luo P. Vista-LLaMA: Reducing hallucination in video narrator via equal distance to visual tokens. In: Proc. of the 2024 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2024. 13151–13160. [doi: 10.1109/CVPR52733.2024.01249]

[41] Li KC, Wang YL, He Y, Li YZ, Wang Y, Liu Y, Wang Z, Xu JL, Chen G, Luo P, Wang LM, Qiao Y. MVBench: A comprehensive multi-modal video understanding benchmark. In: Proc. of the 2024 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle:

IEEE, 2024. 22195–22206. [doi: 10.1109/CVPR52733.2024.02095]

[42] Cheng ZS, Leng SC, Zhang H, Xin YF, Li X, Chen GZ, Zhu YX, Zhang WQ, Luo ZY, Zhao DL, Bing LD. VideoLLaMA 2: Advancing spatial-temporal modeling and audio understanding in Video-LLMs. arXiv:2406.07476, 2024.

[43] Meng LC, Yang JW, Tian R, Dai XY, Wu ZX, Gao JF, Jiang YG. DeepStack: Deeply stacking visual tokens is surprisingly simple and effective for LMMs. In: Proc. of the 38th Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2025. 23464–23487.

附中文参考文献

[1] 胡锦祥, 孟朝晖. 基于视频描述和阅读理解的视频问答研究. 计算机应用研究, 2021, 38(12): 3781–3785. [doi: 10.19734/j.issn.1001-3695.2021.04.0152]

[2] 姚暄, 高君宇, 徐常胜. 基于自监督图对比学习的视频问答方法. 软件学报, 2023, 34(5): 2083–2100. http://www.jos.org.cn/1000-9825/6775.htm [doi: 10.13328/j.cnki.jos.006775]

[3] 赵恩源, 宋宁, 聂婕, 王鑫, 郑程予, 魏志强. 面向遥感视觉问答的尺度引导融合推理网络. 软件学报, 2024, 35(5): 2133–2149. http://www.jos.org.cn/1000-9825/7025.htm [doi: 10.13328/j.cnki.jos.007025]

[4] 殷炯, 张哲东, 高宇涵, 杨智文, 李亮, 肖芒, 孙垚棋, 颜成钢. 视觉语言预训练综述. 软件学报, 2023, 34(5): 2000–2023. http://www.jos.org.cn/1000-9825/6774.htm [doi: 10.13328/j.cnki.jos.006774]

[38] 郭丹, 姚沈涛, 王辉, 汪萌. 嵌入局部聚类描述符的视频问答 Transformer 模型. 计算机学报, 2023, 46(4): 671–689. [doi: 10.11897/SP.J.1016.2023.00671]

作者简介

方承炀, 博士, 讲师, CCF 专业会员, 主要研究领域为场景文字视觉问答, 多模态内容理解, 计算机视觉.

朱畅, 硕士生, CCF 学生会员, 主要研究领域为多媒体信号处理.

姜文晖, 博士, 副教授, 博士生导师, CCF 专业会员, 主要研究领域为图像内容理解, 跨媒体分析.

方玉明, 博士, 教授, 博士生导师, CCF 杰出会员, 主要研究领域为计算机视觉, 多媒体信号处理, 视觉质量评估.

鄢杰斌, 博士, 讲师, CCF 专业会员, 主要研究领域为计算机视觉.