

中图法分类号: TP391 文献标识码: A 文章编号: 1006-8961(XXXX)XX-0001-16

论文引用格式: Fang Chengyang, Zhang Nan, Jiang Wenhui, Liu Xuelin, Fang Yuming. Text Semantic Collaborative Filtering and Region-Aware Scene Text Video Question Answering[J/OL]. Journal of Image and Graphics, XXXX: 1-16. DOI: 10.11834/jig.260072. (方承炀, 章楠, 姜文晖, 刘学林, 方玉明. 文本语义协同筛选与区域感知的场景文字视频问答[J/OL]. 中国图象图形学报, XXXX: 1-16. DOI: 10.11834/jig.260072.) [DOI: 10.11834/jig.260072]

文本语义协同筛选与区域感知的场景文字视频问答

方承炀, 章楠, 姜文晖, 刘学林, 方玉明

江西财经大学计算机与人工智能学院, 江西省南昌市 330032

摘要: 目的 场景文字视频问答是多模态理解领域的关键任务, 现有多模态大语言模型在处理长视频时受输入帧数限制, 直接输入全部视频帧会导致计算开销大, 而均匀采样减少视频帧会导致关键文本信息丢失。为了解决以上问题, 本文提出一种文本语义协同筛选与区域感知的场景文字视频问答方法, 以提升场景文字视频问答的准确性和效率。**方法** 首先, 提出基于文本-语义协同的关键帧选择方法, 利用多模态大语言模型生成的候选答案作为语义锚点, 通过计算其与场景文本的 Levenshtein 编辑距离精准选择关键帧, 并辅以对比语言-图像预训练(contrastive language-image pre-training, CLIP)补充视觉语义相关帧, 以解决视频中细粒度文本信息稀疏容易丢失的问题; 同时, 设计文字区域感知方法, 通过基于掩膜的亮度衰减策略构建视觉“聚光灯”效应, 抑制无关背景噪声使模型聚焦文本区域, 结合推理的置信度差异动态优选答案, 从而提升场景文字视频问答的准确率。**结果** 在 M4-ViteVQA (multi-category multi-frame multi-resolution multi-modal benchmark for video text visual question answering)、RoadText-VQA (road text video question answering) 和 NewsVideoQA (news video question answering) 三个公开场景文字视频问答数据集上进行充分实验, 所提方法在准确率(accuracy, Acc)指标上分别取得 65.55%、72.03% 和 75.54% 的得分, 相比均匀采样基线分别提升了 3.04%、0.76% 和 1.08%, 优于现有方法。消融实验进一步验证了关键帧选择模块和文字区域感知模块的有效性。**结论** 本文提出的方法有效解决了场景文字视频问答中计算效率与信息完整性的矛盾, 通过文本-语义协同选帧策略增强了对细粒度文本的捕获能力, 并通过视觉引导抑制背景噪声, 从而提升了场景文字视频问答的准确性与鲁棒性。

关键词: 场景文字视频问答; 关键帧选择; Levenshtein 距离; 视觉注意力引导; 置信度融合

Text Semantic Collaborative Filtering and Region-Aware Scene Text Video Question Answering

Fang Chengyang, Zhang Nan, Jiang Wenhui, Liu Xuelin, Fang Yuming

School of Computer Science and Artificial Intelligence, Jiangxi University of Finance and Economics, Nanchang 330032, China

Abstract: Objective Scene text video question answering is a critical task in multimodal understanding that aims to reason over both dynamic visual content and fine-grained textual information embedded in videos to answer natural language ques-

收稿日期: 2026-01-31; 修回日期: 2026-09-18

* 通信作者: 姜文晖 wenhui@jxufe.edu.cn

基金项目: 国家自然科学基金项目(62562035, U24A20220, 62132006, 62562034); 江西省自然科学基金(20242BAB23012, 20252BAC200182, 20262BAC240066, 20252BAC230003); 江西省双千计划项目(jxsq2023101092); 江西省职业早期青年科技人才项目(20252BEJ730121)

Supported by: National Natural Science Foundation of China (62562035, U24A20220, 62132006, 62562034); Natural Science Foundation of Jiangxi Province (20242BAB23012, 20252BAC200182, 20262BAC240066, 20252BAC230003); Double Thousand Plan of Jiangxi Province (jxsq2023101092); Early-Career Young Scientists and Technologists Project of Jiangxi Province (20252BEJ730121)

© 中国图象图形学报版权所有

tions. Unlike general video question answering tasks that primarily rely on coarse visual semantics, scene text video question answering requires precise perception and temporal reasoning over scene text, which is often sparse, transient, and highly sensitive to frame selection. With the rapid development of multimodal large language models, such as Qwen2.5-VL, recent approaches have demonstrated promising capabilities in unified vision-language reasoning. However, these models are constrained by limited visual input capacity when processing long videos. Directly feeding all video frames incurs prohibitive computational cost, while conventional uniform frame sampling strategies frequently discard frames containing crucial textual evidence, leading to incomplete or incorrect reasoning. Moreover, scene text in real-world videos is often affected by motion blur, occlusion, illumination variation, and cluttered backgrounds, making it difficult for models to reliably attend to text regions. Existing methods typically treat frame selection and answer prediction as independent processes, without explicitly leveraging semantic feedback from candidate answers to guide visual sampling or attention allocation. As a result, the mismatch between computational efficiency and fine-grained textual reasoning remains a major bottleneck for scene text video question answering. To address these challenges, this paper proposes an answer-driven scene text video question answering framework that tightly integrates scene text cues with visual semantics. The objective is to simultaneously reduce redundant visual computation and enhance the preservation of critical textual information by using candidate answers as semantic anchors, while guiding the model's attention toward text-relevant regions through visual modulation. The proposed approach aims to improve both accuracy and robustness in scene text-centric video question answering under limited visual input budgets. **Method** The proposed framework consists of three tightly coupled components: (1) a text-semantic collaborative keyframe selection strategy driven by candidate answers, (2) a visual semantic supplement mechanism based on contrastive language-image pre-training (CLIP) similarity to ensure coverage of non-textual cues, and (3) a text-region-aware visual guidance module that enhances textual focus and suppresses background noise during reasoning. First, to overcome the limitations of uniform frame sampling, we introduce an answer-driven keyframe selection method that exploits semantic feedback from the model itself. Given a video and a question, a multimodal large language model is first prompted to generate multiple candidate answers based on a coarse visual input. These candidate answers serve as semantic anchors that reflect potential reasoning paths. For each video frame, optical character recognition is applied to extract scene text, forming frame-level textual representations. The relevance between each frame and the candidate answers is then quantified using the Levenshtein edit distance, which effectively captures fine-grained character-level similarity and is particularly well-suited for noisy optical character recognition outputs. Frames whose extracted text exhibits high similarity to candidate answers are selected as keyframes, ensuring that frames containing discriminative textual evidence are preserved. This design explicitly aligns visual sampling with downstream answer semantics, enabling precise retention of fine-grained text information while significantly reducing the number of processed frames. Second, to compensate for cases where critical visual cues are not directly reflected in optical character recognition text—such as logos, symbols, or context-dependent objects—we incorporate a CLIP-based visual semantic supplementation mechanism. By computing cross-modal similarity between video frames and the question-answer semantics in a shared embedding space, the framework identifies visually relevant frames that may lack explicit textual content but are nonetheless essential for reasoning. These frames are merged with the text-selected keyframes, forming a compact yet semantically comprehensive frame set that balances textual precision and visual completeness. Third, to further enhance the model's ability to focus on text regions during reasoning, we propose a text-region-aware visual guidance module. Based on optical character recognition detected text bounding boxes, a mask-guided luminance attenuation strategy is applied to the visual input, creating a “spotlight” effect that highlights text regions while gradually suppressing irrelevant background areas. Unlike hard masking or cropping, this soft attenuation preserves global context while biasing attention toward text-bearing regions. The modulated frames are then fed into the multimodal model for answer reasoning. To improve robustness, the framework evaluates the confidence distribution of generated answers across different visual inputs and dynamically selects the final answer based on confidence divergence, further reducing the impact of noisy frames or ambiguous cues. Together, these components form an integrated answer-driven reasoning pipeline that tightly couples frame selection, visual guidance, and answer validation. The method is fully compatible with existing multimodal large language models and does not require additional annotations or retraining, making it both efficient and scalable. **Result** Extensive experiments were conducted on three public

benchmarks for scene text-centric video question answering: multi-category multi-frame multi-resolution multi-modal benchmark for video text visual question answering (M4-ViteVQA), road text video question answering (RoadTextVQA), news video question answering (NewsVideoQA). These datasets cover a wide range of real-world scenarios, including traffic scenes, news broadcasts, and complex outdoor environments, where scene text plays a critical role in question answering. The proposed method achieved Accuracy scores of 65.55%, 72.03%, and 75.54% on M4-ViteVQA, RoadTextVQA, and NewsVideoQA, respectively. Compared with uniform frame sampling baselines under the same visual input budget, the proposed approach improved performance by 3.04%, 0.76%, and 1.08%, demonstrating consistent gains across diverse datasets. In addition, the method outperformed several existing state-of-the-art scene text video question answering approaches, highlighting the effectiveness of answer-driven frame selection and text-aware visual guidance. Ablation studies further validate the contribution of each module. Removing the answer-driven keyframe selection leads to notable performance degradation, confirming that semantic alignment between candidate answers and scene text is crucial for preserving discriminative frames. Moreover, disabling the text-region-aware visual guidance module results in increased attention dispersion and incorrect reasoning, particularly in cluttered backgrounds, underscoring its role in suppressing irrelevant visual noise. Qualitative analysis reveals that the proposed method successfully focuses on frames containing critical textual evidence, such as street signs, captions, and subtitles, while ignoring redundant or misleading frames. Visualization of attention patterns shows that the spotlight-based guidance effectively directs the model's attention toward text regions, leading to more accurate and explainable reasoning outcomes. **Conclusion** This paper presents an answer-driven scene text video question answering framework that effectively addresses the long-standing trade-off between computational efficiency and fine-grained textual reasoning in scene text-centric video question answering. By leveraging candidate answers as semantic anchors, the proposed text-semantic collaborative keyframe selection strategy ensures that frames containing critical textual information are retained under limited visual input constraints. The integration of CLIP-based visual semantic supplementation further enhances robustness by preserving essential non-textual cues, while the text-region-aware visual guidance module suppresses background noise and promotes focused reasoning. Extensive experiments across multiple benchmarks demonstrate that the proposed method consistently improves accuracy and robustness over uniform sampling and existing approaches, without requiring additional annotations or model retraining. These results highlight the importance of incorporating answer-level semantic feedback and text-aware visual modulation into multimodal reasoning pipelines. Overall, this work provides a scalable and effective solution for scene text video question answering and offers new insights into answer-driven visual sampling and attention guidance for multimodal large language models. The proposed framework has strong potential for real-world applications such as intelligent transportation systems, news video analysis, and multimedia information retrieval, where accurate understanding of dynamic scene text is essential.

Key words: scene text video question answering; key frame selection; Levenshtein distance; visual attention guidance; confidence fusion

论文引用格式: Fang Chengyang, Zhang Nan, Jiang Wenhui, Liu Xuelin, Fang Yuming. Text Semantic Collaborative Filtering and Region-Aware Scene Text Video Question Answering[J/OL]. Journal of Image and Graphics. DOI: 10.11834/jig.260072. (方承炀, 章楠, 姜文晖, 刘学林, 方玉明. 文本语义协同筛选与区域感知的场景文字视频问答[J/OL]. 中国图象图形学报. DOI:10.11834/jig.260072.)

0 引言

场景文字视频问答旨在让模型在理解视频动态

视觉内容的基础上,结合其中蕴含的细粒度文字信息(Gao等,2023),如路牌、标语、新闻字幕等,来回答自然语言提问(Li等,2023;Maaz等,2024)。相比静态图像,视频包含更丰富的时序信息,但也带来了更高的计算和推理复杂性(Tong等,2022;Wang等,2022b)。作为多模态理解与自然语言处理交叉领域的前沿方向,该任务在智能交通系统、新闻视频检索以及视障辅助等实际应用中具有巨大的潜力。

现有的视频问答方法通常采用基于多模态大语言模型的架构,如大语言与视觉助手视频模型(Large Language and Vision Assistant for Video, LLaVA-Video)、通义千问视觉语言模型(qwen vision-

language model, Qwen-VL) 系列(Bai 等, 2023)、Otter (Otter: A Multi-Modal Model with In-Context Instruction Tuning) (Li 等, 2025a) 以及 VideoChat (VideoChat: Chat-Centric Video Understanding) (Li 等, 2025b), 将视频帧编码为视觉 token (Yan 等, 2022), 作为语言模型的输入上下文 (Dai 等, 2023; Ye 等, 2023)。虽然这些多模态大语言模型在推理方面展现了强大能力, 但受限于模型输入限制, 通常只能对视频进行有限的采样。在这种通用框架下, 场景文字视频问答任务依然面临两个关键挑战。

首先, 视频帧采样策略需要在降低计算开销与保留细粒度信息之间进行权衡 (Lin 等, 2024)。视频包含海量帧, 直接输入全量视频帧会超出模型的处理能力 (Li 等, 2024)。然而, 场景文本在视频中往往具有稀疏性、瞬时性和细粒度特征。传统的主流方法多采用均匀采样策略, 由于缺乏对文本内容的感知, 并未充分利用视频中的文本信息 (Tom 等, 2023), 这种“盲目”采样极易丢弃包含关键文字证据的帧。如图 1 左侧所示, 当全量视频帧由于计算开销过大无法处理时, 简单的均匀采样导致关键的文本相关帧被忽略, 从而使模型无法感知到对于回答问题至关重要的细粒度文本信息。针对视频帧采样导致细粒度文本丢失的挑战, 部分研究尝试引入场景文本 (Fang 等, 2025) 或视频描述作为额外模态来

解决细粒度文本丢失的问题 (Xiao 等, 2026)。当前研究主要通过分析帧与问答间的语义因果关系来选择帧 (Fang 等, 2026), 引入自适应关键帧采样来最大化信息量 (Tang 等, 2025), 这些方法证明了关键帧选择对视频理解性能的显著影响 (Xu 等, 2024)。但这些方法多关注全局视觉相似度, 仍忽视了视频中隐含的字符级线索。

其次, 多模态模型在推理过程中难以有效聚焦于关键文本区域。真实场景中的视频常伴有运动模糊、遮挡、复杂背景干扰以及光照变化 (Xie 等, 2025)。现有的视频问答模型 (如图 1 右侧红框所示) 往往更倾向于关注视觉显著区域 (如人脸、大幅物体), 而忽略了背景中面积较小但语义密度高的文本区域。这种注意力分散导致模型在处理文字密集型任务时产生幻觉或推理错误。针对视觉干扰导致文本聚焦难的问题, 部分研究尝试通过边界框硬标注策略标出位置, 但往往引入额外的计算开销且破坏图像自然语义。为了克服这种硬性截取的局限性, 后续部分研究转向通过引入额外的文本识别模块或空间注意力来增强理解 (Fang 等, 2023)。然而, 如何显式地引导模型在复杂的视觉空间中排除背景噪声, 实现精准的文本区域对齐, 仍是一个尚未解决的难题。



图 1 场景文字视频问答任务中的主要局限性

Fig. 1 Major limitations in scene text video question answering task

为应对上述两个挑战, 本文提出一种文本语义协同筛选与区域感知的场景文字视频问答方法。该方法引入“答案驱动”的逆向索引策略, 通过候选答

案引导精准选帧, 并设计文字区域感知模块增强视觉聚焦。

为了提升模型对细粒度文本帧的捕获能力, 本
© 中国图象图形学报版权所有

文提出一种基于文本-语义协同的关键帧选择方法。该方法首先利用多模态大语言模型在粗粒度输入下生成候选答案,作为指导检索的“语义锚点”。通过计算候选答案与场景文本之间的 Levenshtein 编辑距离,精准选择关键帧。这种字符级的匹配度量比基于全局视觉相似度的帧选择(Su 等, 2023)更能关注到细粒度文本。同时,引入对比语言-图像预训练(contrastive language - image pre-training, CLIP)(Radford 等, 2021)视觉语义补充机制,在文本稀疏时利用全局语义兜底,实现了显式文本匹配与隐式视觉关联的平衡。

同时,针对背景噪声干扰推理的问题,本文设计了文字区域感知方法。该方法利用光学字符识别坐标构建空间掩膜(Wang 等, 2023a),并执行基于亮度的软性衰减策略。通过保持文本区域原始亮度并适度降低背景亮度,在视觉上构建“聚光灯”效应,强制模型聚焦于高价值文本语义载体。此外,提出基于模型置信度的双路径推理框架,通过比较全局基线答案与局部精炼答案的后验概率,动态优选最终输出,进一步增强了推理的鲁棒性。

本文的主要贡献总结如下:

1)提出一种基于文本-语义协同的关键帧选择方法,利用候选答案与帧内文本的编辑距离精准定位关键帧,将精炼推理阶段的输入帧数压缩至3帧的同时有效避免了细粒度文本信息的丢失。

2)设计了基于亮度衰减的文字区域感知方法和置信度融合机制,通过视觉引导抑制背景噪声,结合多路径推理实现了答案的动态自适应优选。

3)本文在 M4-ViteVQA (multi-category multi-frame multi-resolution multi-modal benchmark for video text visual question answering)(Zhao 等, 2022)、RoadTextVQA (road text video question answering)(Tom 等, 2023)和 NewsVideoQA (news video question answering)(Jahagirdar 等, 2023)三个标准数据集上进行了广泛实验,定量比较验证了本文方法的有效性。本文方法在三个数据集的准确率(accuracy, Acc)上分别达到 65.55%、72.03% 和 75.54%,凸显了在处理复杂场景文字视频问答任务上的显著优势。

1 本文方法

1.1 总体框架

本文提出的文本语义协同筛选与区域感知的场景文字视频问答方法总体框架如图 2(a)所示。该方法主要包括两个基本模块:1)基于文本-语义协同的关键帧选择模块,旨在利用候选答案作为语义锚点,精准筛选包含关键文本线索的高价值帧,以解决视频中细粒度文本稀疏容易被忽略的问题;2)文字区域高亮与置信度融合模块,通过亮度衰减策略抑制无关背景噪声以显式增强文本区域特征,并结合全局与局部推理的置信度差异动态优选最终答案。

1.2 基于文本-语义协同的关键帧选择模块

在包含密集文本的视频场景中,单纯的视觉相似度往往难以捕捉细微的文字线索。因此,本模块采用“答案驱动”策略,通过文本层面的精准锚定与视觉层面的语义补充,构建高质量的输入帧集合。具体结构如图 2(b)所示。

1.2.1 候选答案生成

不同于传统的基于视觉特征的帧选择,本文提出一种由答案驱动的反向推理策略。给定视频 $V = \{v_1, v_2, \dots, v_T\}$ (T 表示视频总帧数)和问题 q ,我们首先利用多模态大语言模型 M_θ 在稀疏采样帧 V_{sparse} 上的推理能力,获取答案的先验分布。我们选取概率最高的 K 个生成序列作为候选答案集 $A = \{a_1, \dots, a_K\}$,其中 a_k 服从: $a_k \sim p_{M_\theta}(a | V_{\text{sparse}}, q)$ 这些候选答案充当了潜在的“语义锚点”,用于指导后续细粒度帧检索和置信度融合。

1.2.2 场景文本感知与细粒度匹配度量

视频中的场景文本往往承载着回答问题的关键语义线索。为了从非结构化的视频流中提取这些线索,我们首先对视频采样帧序列 $V = \{v_i\}_{i=1}^T$ 进行场景文本检测与识别。我们使用光学字符识别引擎提取场景文本,对于第 t 帧 v_t ,提取到的文本实例集合定义为 T_t :

$$T_t = \{s_{t,1}, s_{t,2}, \dots, s_{t,m}\} \quad (1)$$

式中 $s_{t,m}$ 表示第 t 帧中识别出的第 m 个文本字符串。

为了量化候选答案集合 $A = \{a_1, \dots, a_K\}$ 与帧内文本的语义一致性,我们引入 Levenshtein 编辑距离作为核心度量指标。相比于基于向量的余弦相似

度,编辑距离能更精确地捕捉字符级的细微差异(如拼写错误或光学字符识别噪声)。对于候选答案 a_k (长度 $|a_k|$)和帧内文本 $s_{t,m}$ (长度 $|s_{t,m}|$),其编辑距离 $\text{lev}(a_k, s_{t,m})$ 通过动态规划递归计算:

$$\text{lev}_{i,j}(a_k, s_{t,m}) = \min \begin{cases} \text{lev}_{i-1,j}(a_k, s_{t,m}) + 1, \\ \text{lev}_{i,j-1}(a_k, s_{t,m}) + 1, \\ \text{lev}_{i-1,j-1}(a_k, s_{t,m}) + \mathbb{I}(a_k[i] \neq s_{t,m}[j]) \end{cases} \quad (2)$$

式中 $\text{lev}_{i,j}(\cdot)$ 表示候选答案 a_k 的前缀子串 $a_k[1:i]$ 与场景文本 $s_{t,m}$ 的前缀子串 $s_{t,m}[1:j]$ 之间的最小转换代价。公式中的三项分别对应三种基础编辑操作:1)删除:对应 $\text{lev}_{i-1,j} + 1$,即将 a_k 的第 i 个字符视为冗余并移除;2)插入:对应 $\text{lev}_{i,j-1} + 1$,即在 $s_{t,m}$ 中补全缺失的字符以匹配 a_k ;3)替换:对应 $\text{lev}_{i-1,j-1} + \mathbb{I}(\cdot)$,式中 $\mathbb{I}(\cdot)$ 为指示函数,当字符 $a_k[i]$ 与 $s_{t,m}[j]$ 不相等时,函数值为1(引入替换代价),否则为0(零代价匹配)。通过这种状态转移机制,该度量方法能够寻找两个字符串之间的最优对齐路径,从而在允许一定程度的光学字符识别误差(如将数字'0'误识为字母'O')的前提下,鲁棒地评估答案与场景文本的语义关联度。

基于上述度量,我们定义第 t 帧相对于第 k 个候选答案的最小语义距离 $D(a_k, v_t)$ 为该答案与帧内所有文本实例距离的最小值:

$$D(a_k, v_t) = \min_{s \in T_t} \text{lev}(a_k, s) \quad (3)$$

式中 s 表示 T_t 中的场景文本。

为了构建文本相关的关键帧集合 F_{text} ,我们采用“答案驱动”的逆向索引策略。对于每个候选答案 a_k ,我们在时间域上搜索使其语义距离最小的最佳匹配帧 v_t^k :

$$v_t^k = \arg \min_{v_t \in V} D(a_k, v_t) \quad (4)$$

最终的文本支持帧集合由所有候选答案的最佳匹配帧去重构成:

$$F_{\text{text}} = \bigcup_{k=1}^K \{v_t^k\} \quad (5)$$

该集合包含了视频中与预测答案文本重合度最高的关键视觉片段,确保了模型关注到细粒度的文本。

1.2.3 视觉语义互补与动态帧补全

尽管文本匹配策略在文本密集型场景下表现优

异,但在部分视觉主导或文本稀疏的场景中,仅依赖 F_{text} 可能导致输入信息不足。为保证推理的鲁棒性,我们设定最小关键帧数量阈值 τ (本文设为3),并引入 CLIP 模型构建视觉语义互补机制。设当前已选文本支持帧集合为 $F_{\text{curr}} = F_{\text{text}} \circ$ 若 $|F_{\text{curr}}| < \tau$,说明仅凭文本无法提供充足的上下文,此时我们需要从剩余帧集合 $V_{\text{rem}} = V \setminus F_{\text{text}}$ 中补充视觉相关帧。我们利用 CLIP 的双塔架构计算剩余帧 $v_t \in V_{\text{rem}}$ 与问题文本 q 的全局语义相似度 $S_{\text{clip}}(v_t, q)$:

$$S_{\text{clip}}(v_t, q) = \frac{\varepsilon_I(v_t) \cdot \varepsilon_T(q)}{\|\varepsilon_I(v_t)\|_2 \|\varepsilon_T(q)\|_2} \quad (6)$$

式中 $\varepsilon_I(\cdot)$ 和 $\varepsilon_T(\cdot)$ 分别表示 CLIP 的图像编码器和文本编码器, $\|\cdot\|_2$ 表示 L_2 范数。基于相似度排序,我们选取 Top-K 个帧构成补充集合 F_{supp} ,其中 $K = \tau - |F_{\text{curr}}|$:

$$F_{\text{supp}} = \text{Top-K}_{v_t \in V_{\text{rem}}}(S_{\text{clip}}(v_t, q), K) \quad (7)$$

最终的关键帧集合 F_{final} 由文本支持帧与视觉补充帧动态融合而成:

$$F_{\text{final}} = F_{\text{text}} \cup F_{\text{supp}} \quad (8)$$

通过这种文本-语义协同机制,算法能够在“显式文本匹配”与“隐式视觉关联”之间自动平衡,既能在文本密集场景下精准定位,又能在文本稀疏场景下利用全局语义兜底,有效解决了单一模态选帧的信息偏差问题。

1.3 文字区域感知模块

1.3.1 文字区域高亮处理

为了增强多模态大语言模型对文本语义信息的关注度,并抑制无关背景引入的视觉噪声,我们在选定的关键帧集合 F_{final} 上提出了一种显式的视觉注意力引导策略。该策略旨在构建一种视觉上的“聚光灯”效应,强制模型聚焦于高对比度的文本区域。具体结构如图2(c)所示。

具体而言,对于任意关键帧 $v_t \in F_{\text{final}}$,我们利用光学字符识别引擎返回的文本边界框集合 $B_{v_t} = \{b_1, b_2, \dots, b_m\}$ 来构建二值空间掩膜 $M \in \{0, 1\}^{H \times W}$, H 表示掩膜高度, W 表示掩膜宽度。掩膜中像素 (x, y) 的值由其空间位置决定:

$$M(x, y) = \mathbb{I}\left((x, y) \in \bigcup_{b_m \in B_{v_t}} \Omega(b_m)\right) \quad (9)$$

式中 $\Omega(b_m)$ 表示边界框 b_m 所覆盖的像素区域, $\mathbb{I}(\cdot)$

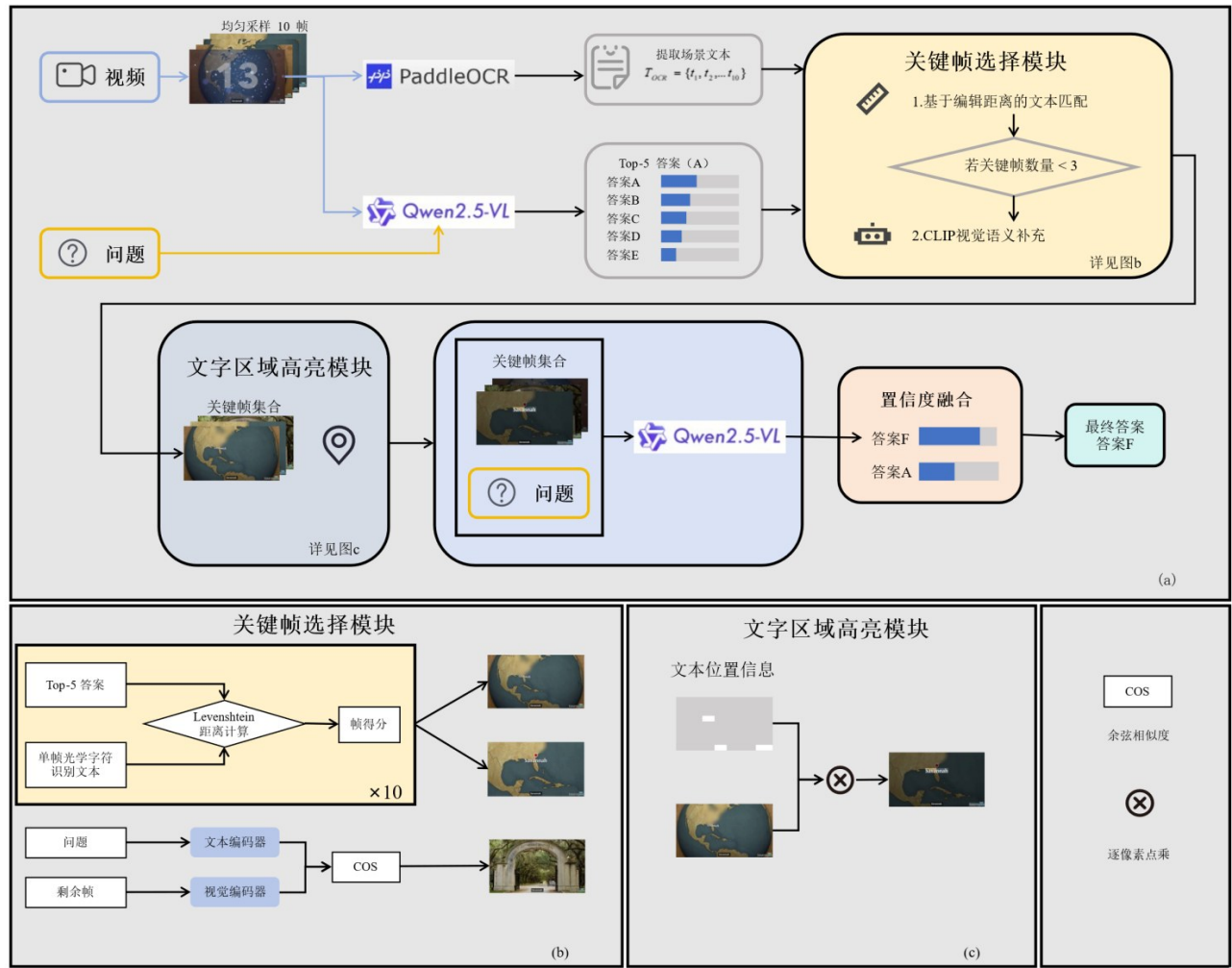


图2 文本语义协同筛选与区域感知的场景文字视频问答方法总体框架图

Fig. 2 The overall framework of text semantic collaborative filtering and region-aware scene text video question answering

为指示函数。即文本区域内的 $M(x, y) = 1$, 背景区域为0。

基于该掩膜,我们对原始帧 v_i 执行像素级加权,生成增强后的特征帧 $\hat{v}_i$:

$$\hat{v}_i = v_i \odot M + \alpha \cdot v_i \odot (1 - M) \quad (10)$$

式中 $\odot$ 表示逐元素乘积, $\mathbf{1}$ 为全1矩阵。参数 $\alpha \in [0, 1]$ 为背景亮度衰减系数,用于控制背景信息的保留程度。本文,我们设置 $\alpha = 0.8$,即保持文本区域原始亮度不变,同时将非文本背景亮度衰减至原图像的80%。

这种非线性的亮度变换 $T(v_i) \rightarrow \hat{v}_i$ 有效地最大化了文本区域与背景环境的信噪比,使得模型在重推理阶段能够排除视觉干扰,显式地聚焦于关键文本语义载体。

1.3.2 基于模型置信度的双路径推理

为了在全局时序上下文与局部细粒度文本支持帧集合之间实现自适应权衡,我们提出了一种基于模型置信度的双路径推理框架。该框架通过在不同信息粒度下并行生成候选答案,并依据模型的生成置信度进行选择,从而提升推理的鲁棒性与准确性。

为提高整体推理效率,我们引入计算复用策略,避免对已有结果的重复计算。具体而言,双路径机制包含以下两个互补的推理路径:1)基线路径:直接复用第1.2.1节中基于均匀采样视频帧集合 V_{sparse} 所生成的Top-1候选答案 A_{base} 。该路径不引入额外计算开销,主要用于提供基于全局上下文的稳健参考。2)精炼路径:将经注意力调制后的关键帧 $\hat{v}_i$ 输入模型,执行细粒度推理以获得精炼答案 A_{ref} ,该路径更侧重于局部细粒度文本的建模。

对于任意生成序列 $y \in \{A_{\text{base}}, A_{\text{ref}}\}$, 我们利用模型在自回归生成过程中输出的后验概率来量化其置信度。设 z_t 表示模型在第 t 个生成步的 Logits 向量, W 表示模型词表, k 为词表中的 token 索引, 则当前 token y_t 的生成概率 $p(y_t | y_{<t}, q, V)$ 为:

$$p(y_t | y_{<t}, q, V) = \frac{\exp(z_t[y_t])}{\sum_{k=1}^{|W|} \exp(z_t[k])} \quad (11)$$

式中 $y_{<t}$ 表示第 t 步之前已经生成的 token 序列。

由于序列生成过程中逐步概率相乘可能引入长度偏置, 我们采用平均标记置信度作为序列级的置信度得分 $C(y)$:

$$C(y) = \frac{1}{L} \sum_{t=1}^L p(y_t | y_{<t}, V) \quad (12)$$

式中 L 表示生成序列的长度。

最终答案 A^* 的选择遵循“最大置信度原则”, 即通过比较两个独立推理过程的后验概率来确定最优输出:

$$A^* = \begin{cases} A_{\text{ref}}, & C(A_{\text{ref}}) > C(A_{\text{base}}) \\ A_{\text{base}}, & \text{otherwise} \end{cases} \quad (13)$$

当 $C(A_{\text{ref}})$ 较高时, 表明关键帧中包含明确且与问题高度相关的文本线索, 模型“确信”其局部推理结果, 因此采纳精炼答案 A_{ref} 。当 $C(A_{\text{ref}})$ 较低时, 通常意味着关键帧选择有误或文本信息不足, 此时系统稳健地回退到基于全局视野的基线答案 A_{base} 。

2 实验结果与分析

2.1 数据集与评估指标

2.1.1 数据集

本文方法在三个公开的场景文本视频问答数据集: M4-ViteVQA、RoadTextVQA 和 NewsVideoQA 进行了充分的实验。M4-ViteVQA 是 ByteDance 提出的场景文字视频问答基准, 包含 9 个类别的 7620 段视频剪辑, 共 25123 个问答对; RoadTextVQA 是驾驶场景下的视频问答数据集, 共收集了 3222 段驾驶视频和 10500 个基于道路标志文字的问题; NewsVideoQA 则包含来自全球多个新闻频道的 3000 余段新闻视频和 8600 多个问答对, 要求模型结合视频中的文本和视觉信息回答问题。上述数据集均包含大量需要识别和推理视频中场景文本才能回答的问题, 是验

证我们方法的理想测试集。

2.1.2 评估指标

本文采用准确率和答案归一化编辑相似度 (answer normalized Levenshtein similarity, ANLS) 作为评价指标。准确率用于衡量模型推理的精确性, 定义为经文本归一化 (包括小写化及去除标点符号) 后, 预测答案与人工标注真值完全匹配的样本比例。答案归一化编辑相似度基于 Levenshtein 编辑距离量化预测与真值之间的语义接近度, 并引入 0.5 的截断阈值, 当相似度低于该阈值时得分为零, 从而更合理地评价长文本生成的质量。

2.1.3 实施细节

本文方法基于 PyTorch 框架实现, 实验在配备 NVIDIA RTX PRO 6000 GPU 的服务器上完成。为了平衡推理效率与显存占用, 多模态大语言模型推理全程采用半精度浮点数格式进行运算。本文方法由多模态大语言模型、场景文本提取模块和视觉语义补充模块组成, 其具体配置如下: 1) 多模态大语言模型: 采用 Qwen2.5-VL-7B-Instruct (Bai 等, 2025) 作为核心推理模型。在推理阶段, 我们设置最大生成长度为 128; 2) 场景文本提取: 调用飞桨光学字符识别系统 (paddle optical character recognition, PaddleOCR) (Du 等, 2020) 工具库进行端到端的文本检测与识别, 置信度阈值设为默认值, 以保证高召回率; 3) 视觉语义补充: 采用预训练的 CLIP 模型计算问题和帧图像之间的余弦相似度。所有输入图像均缩放到 224×224 大小, 问题文本经过标准的分词处理。

2.2 定量分析

为了定量比较, 将本文的方法与一些经典方法和最近的先进方法进行了比较, 包括 M4C (multi-modal multi-copy mesh) (Hu 等, 2020)、JustAsk (Just ask: Learning to answer questions from millions of narrated videos) (Yang 等, 2021)、All-in-one-B (All in one: Exploring unified video-language pre-training) (Wang 等, 2023b)、GIT (GIT: A generative image-to-text transformer for vision and language) (Wang 等, 2022a)、SINGULARITY (Revealing single frame bias for video-and-language learning) (Lei 等, 2023)、Video-LLaVA* (Video-LLaVA: Learning united visual representation by alignment before projection) (Lin 等, 2024)、VideoLLaMA2* (VideoLLaMA 2: Advancing

spatial-temporal modeling and audio understanding in video-LLMs) (Cheng 等, 2024)、NVILA* (NVILA: Efficient frontier visual language models) (Liu 等, 2025)、T5-ViteVQA (Towards video text visual question answering: Benchmark and baseline) (Zhao 等, 2022)、TEA (Track the answer: Extending TextVQA from image to video with spatio-temporal clues) (Zhang 等, 2025a)、GAT (Gather and trace: Rethinking video TextVQA from an instance-oriented perspective) (Zhang 等, 2025b)。

由表1结果可见,本文提出的文本语义协同筛选与区域感知的场景文字视频问答方法提升了场景文字视频问答任务的性能。总的来说,在三个以场景文本为核心的公开数据集上,本文方法均优于现有的先进方法及均匀采样基线。

在 M4-ViteVQA 数据集上,本文方法取得了最优结果,准确率达到 65.55%,ANLS 达到 73.23%。得益于本文提出的基于文本-语义协同的关键帧选择策略,模型能够通过计算候选答案与光学字符识别引擎提取文本的 Levenshtein 编辑距离,精准定位

包含细粒度文本线索的关键帧;同时配合文字区域感知模块对无关背景噪声的有效抑制,与采用均匀采样的基线模型(Baseline)相比,本文模型在准确率和 ANLS 上分别提升了 3.04% 和 2.27%。

在 RoadTextVQA 数据集上,得益于关键帧选择策略对稀疏场景文本的有效保留,本文模型准确率达到了 72.03%。与基线相比,通过引入视觉语义补充机制解决文本较少时的帧选择问题,本文方法使准确率提升了 0.76%。而在 NewsVideoQA 数据集上,本文模型同样展示了出色的鲁棒性,准确率达到 75.54%。相较于基线模型,准确率提升了 1.08%。这些结果表明,本文提出的场景文本与视觉语义的答案驱动视频问答方法有效地解决场景文字视频问答中计算开销与信息丢失的矛盾。

在 RoadTextVQA 数据集上的性能提升幅度相对有限,为进一步验证性能提升是否具有统计意义,本文采用 McNemar 检验对预测结果进行统计显著性分析, p 值(p -value)为 0.0433,小于 0.05,表明性能提升达到统计显著水平

表 1 各方法在 M4-ViteVQA、RoadTextVQA 和 NewsVideoQA 的评估结果
Table 1 Evaluation results of various methods in M4-ViteVQA, RoadTextVQA, and NewsVideoQA.

方法类别	方法	发表来源	M4-ViteVQA Task1 Split1		RoadTextVQA		NewsVideoQA	
			Acc.(%) ↑	ANLS.(%) ↑	Acc.(%) ↑	ANLS.(%) ↑	Acc.(%) ↑	ANLS.(%) ↑
场景文字图片问答	M4C	CVPR20	18.66	24.20	28.92	32.27	-	-
	JustAsk	ICCV21	10.81	15.40	-	-	-	-
	All-in-one-B	CVPR23	11.47	15.30	-	-	-	-
	GIT	TMLR22	-	-	29.58	35.23	-	-
	SINGULARITY	ACL23	-	-	24.62	30.79	-	-
视频语言预训练	T5-ViteVQA	NeurIPS22	23.17	30.10	-	-	-	-
	TEA-Large	AAAI25	37.49	46.38	48.14	54.85	-	-
	GAT-Large	MM25	38.01	47.53	50.23	58.12	-	-
	Video-LLaVA	EMNLP24	15.82	17.77	30.82	40.92	-	-
场景文字视频问答	VideoLLaMA2	arXiv24	20.04	21.73	25.11	36.53	-	-
	NVILA	CVPR25	37.89	47.67	49.98	57.22	-	-
	本文-Baseline	-	62.51	70.96	71.27	76.92	74.46	82.88
	本文	-	65.55	73.23	72.03	77.24	75.54	83.42

注:加粗字体表示各列最优结果,“↑”表示数值越高越好,“-”表示没有可用的结果。

2.3 消融实验

为了验证在该模型中添加基于文本-语义协同的关键帧选择模块和文字区域感知模块对场景文字视频问答任务的有效性,同时验证部分模型设置对性能的影响,本文在 M4-ViteVQA 验证集上进行了消融研究。

2.3.1 模块有效性分析

实验结果如表 2 所示,详细展示了不同策略下的性能表现。第一行表示仅使用均匀采样策略且不包含任何附加模块的基线模型方案。当仅采用关键帧选择模块或文字区域感知模块时,模型的性能均显著上升,准确率分别从基线的 62.51% 提升至 64.23% 和 64.08%,验证了这两个模块在捕捉关键文本信息和增强文本感知的关键作用。

表 2 在 M4-ViteVQA 验证集上验证不同模块的有效性
Table 2 Evaluate the effectiveness of different modules on the M4-ViteVQA validation set.

模块		评估指标	
关键帧选择	文字区域感知	Acc.(%) ↑	ANLS.(%) ↑
-	-	62.51	70.96
✓	-	64.23	72.02
-	✓	64.08	71.87
✓	✓	65.55	73.23

注:“↑”表示数值越高越好,“-”表示不包含对应模块,“✓”表示包含对应模块,加粗字体为每列最优值。

与基线模型相比,最后一行完整方案通过同时引入基于文本-语义协同的关键帧选择模块和文字区域感知模块,不仅实现了对细粒度文本线索的精准定位,还有效抑制了无关背景噪声的干扰,从而取得了最佳性能。本文提出的完整模型方案在准确率和 ANLS 指标上均表现出显著优势,分别达到了 65.55% 和 73.23%,充分证明了本文提出的两个模块的有效性。

2.3.2 关键帧中答案文本覆盖率的消融分析

为了进一步探究关键帧筛选模块对关键文本信息的保留能力,本文统计不同筛选策略下所选帧中包含答案文本的帧比例,并统计最终筛选帧数量,结

果如表 3 所示。

第一行展示了基线模型的表现,采用简单的均匀采样策略虽然保留了较多的视频帧,但包含答案文本的帧比例较低。这表明均匀采样引入大量与问题无关的视觉信息,导致关键文本线索的密度较低。

第二行采用基于语义嵌入余弦相似度的文本匹配策略。具体而言,将候选答案与场景文本分别编码为语义向量,并利用余弦相似度衡量二者的语义相关性,据此筛选与候选答案最相关的帧。该实验用于与本文采用的 Levenshtein 编辑距离进行直接对比,从而分析不同文本匹配度量对关键帧筛选效果的影响。

第三行验证了第 1.2.2 节提出的场景文本感知与细粒度匹配度量的精确性。通过利用候选答案作为“语义锚点”并计算 Levenshtein 编辑距离,模型筛选出的帧中包含答案文本的比例显著提升。在 RoadTextVQA 和 NewsVideoQA 数据集上,该比例分别从基线的 12.78% 和 32.89% 跃升至 28.39% 和 56.18%,同时总帧数大幅下降。相比基于语义嵌入余弦相似度的匹配策略,Levenshtein 编辑距离能够更加直接地刻画候选答案与场景文本之间的字符级一致性,适用于场景文字中存在短文本、局部字符匹配以及光学字符识别结果存在一定偏差的情况。

第四行展示了第 1.2.3 节引入视觉语义互补机制后的最终结果。尽管引入 CLIP 补充视觉相关帧后,包含答案文本的帧比例较第三行略有下降,但仍远高于均匀采样基线。这一变化符合预期,因为 CLIP 补充机制旨在召回那些虽不含直接文本但具备重要视觉上下文的帧,从而在“显式文本匹配”与“隐式视觉关联”之间实现平衡。最终方案在保持较低总帧数的同时,确保了输入信息的高密度与完整性。

2.3.3 基于 CLIP 视觉相似度的帧选择策略分析

为了验证仅依赖视觉-语义相似度进行关键帧选择的局限性,我们在 M4-ViteVQA 上进行了消融实验。具体而言,我们直接利用 CLIP 模型计算问题文本与所有视频帧的余弦相似度,并选取相似度最高的 Top - K 帧($K = 1 \sim 6$)输入模型进行推理。实验结果如图 3 所示。

结果表明,基于 CLIP 的帧选择策略在场景文本问答任务中表现不佳。即使在选择帧数较多($K =$

表 3 关键帧选择策略下包含答案文本的帧比例与总帧数对比分析

Table 3 Comparative analysis of the proportion of frames containing answer text and total frame counts under key frame selection strategies.

方法	M4-ViteVQA Task1Split1		RoadTextVQA		NewsVideoQA	
	比例(%)	总帧数	比例(%)	总帧数	比例(%)	总帧数
基线-均匀采样 10 帧	41.14	19710	12.78	10505	32.89	9200
语义嵌入余弦相似度匹配	45.08	4075	21.3	2019	50.34	1599
场景文本感知与细粒度匹配	51.36	3616	28.39	1937	56.18	1360
视觉语义互补与动态帧补全	48.7	6010	23.38	3182	51.92	2766

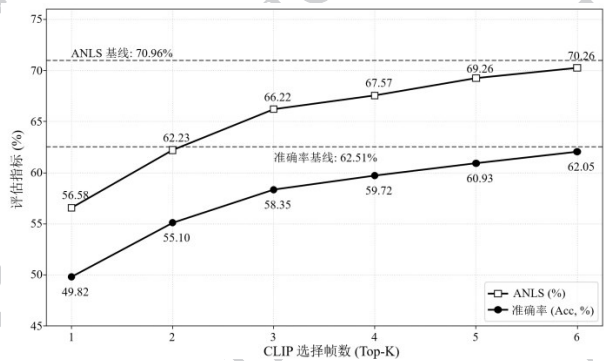


图 3 在 M4-ViteVQA 验证集上评估基于 CLIP 视觉相似度的帧选择策略

Fig. 3 Evaluate the frame selection strategy based on CLIP visual similarity on the M4-ViteVQA validation set.

6)的情况下,其准确率(62.05%)和 ANLS(70.26%)仍低于均匀采样的基线准确率(62.51%)。随着选取帧数的减少,性能呈现急剧下降趋势,当仅选择 1 帧时,准确率跌至 49.82%。这充分说明 CLIP 模型虽然在通用的视觉对象匹配上表现出色,但缺乏对细粒度场景文本的感知能力。对于需要精确识别文字(如路牌、字幕)的问题,CLIP 往往倾向于检索包含相关物体但文字模糊或无关的帧,导致关键信息丢失。因此,单纯依赖 CLIP 进行帧筛选是不可靠的,引入显式的场景文本匹配至关重要。

2.3.4 输入帧数量与自适应选择策略的对比分析

为了探究输入帧数量对推理准确性的影响,我们对比了固定选取 Top - K 帧(K = 1 ~ 5)与本文提出的“自适应选择”策略的性能差异,结果如表 4 所示。

实验结果表明,输入帧数与模型性能之间并非简单的正相关关系。当仅选择 1 帧时,由于视觉信息缺失严重,模型准确率仅为 59.51%。随着帧数增加至 3 帧,模型能够获取更充分的场景文本线索,

表 4 不同帧数量选择策略在 M4-ViteVQA 验证集上的性能对比

Table 4 Performance comparison of different frame quantity selection strategies on M4-ViteVQA validation set.

选择 Top-K 帧	评估指标	
	Acc.(%) ↑	ANLS.(%) ↑
固定 5 帧	64.18	72.18
固定 4 帧	64.18	72.12
固定 3 帧	64.49	72.03
固定 2 帧	62.86	70.53
固定 1 帧	59.51	68.04
自适应帧数	64.23	72.02

注:“↑”表示数值越高越好,加粗字体为每列最优值。

准确率达到峰值 64.49%。然而,当进一步强制增加帧数至 4 帧或 5 帧时,准确率反而略有下降(64.18%)。这表明,强制引入匹配度较低的帧会引入视觉噪声,干扰多模态大语言模型的推理判断。

相比之下,本文采用的自适应选择策略在不设定固定帧数硬约束的情况下,取得了 64.23% 的准确率,与最优的固定 3 帧策略性能基本持平(差距仅 0.26%)。该策略通过对候选答案对应的最佳帧进行去重,能够根据视频内容的复杂度和答案的一致性动态调整输入帧数。这证明了自适应策略能够在保证关键信息覆盖的同时,有效避免冗余帧带来的噪声干扰,是一种更具泛化能力的帧选择方案。

2.3.5 文字区域高亮方式与背景抑制参数的消融

为了确定引导模型关注场景文本的最佳视觉提示方式,我们对比了不同形式的高亮策略(边界框 vs. 亮度衰减)及不同强度的背景抑制参数,结果如表 5 所示。

实验结果表明,传统的“硬提示”策略(绘制边界
© 中国图象图形学报版权所有

表5 M4-ViteVQA 验证集上不同文本高亮策略与参数的消融实验

Table 5 Ablation study of different text highlighting strategies and parameters on the M4-ViteVQA validation set.

处理方式	评估指标	
	Acc.(%) ↑	ANLS.(%) ↑
无处理	64.23	72.02
非文本区域图片亮度为0.8	64.49	72.14
非文本区域图片亮度为0.5	64.08	71.87
非文本区域图片亮度为0.3	63.72	71.38
红色边界框标出文本	63.01	71.10
绿色边界框标出文本	63.12	70.79

注：“↑”表示数值越高越好，加粗字体为每列最优值。

框)表现不佳。使用红色或绿色边界框框选文本区域后,模型准确率分别降至 63.01% 和 63.12%,低于“无处理”基线(64.23%)。这可能是因为显式的几何边框引入了额外的视觉干扰,破坏了图像的自然语义特征。

其次,基于亮度的“软提示”策略效果取决于背景信息的保留程度。当将非文本区域亮度调整为原图的 0.8 时,模型在保持环境上下文可见的同时有效地聚焦于文本,准确率达到最高的 64.49%。然而,随着背景亮度进一步降低(0.5 至 0.3),模型性能开始下滑。特别是当亮度降至 0.3 时,准确率跌至 63.72%。这表明视频中的非文本背景对多模态推理仍重要,过度的背景抑制会导致关键上下文丢失。因此,采用温和的亮度衰减(0.8)是平衡文本聚焦与全局感知的最佳方案。

2.3.6 候选答案数量对模型推理性能的影响

为了验证候选答案数量(K)在引导关键帧选择中的作用,我们测试了在自适应选帧策略下,利用不同数量的候选答案(K = 1 ~ 5)进行推理的性能表现,结果如表 6 所示。实验数据呈现出明显的单调上升趋势,证明了增加候选答案数量能够显著提升系统的容错率和检索能力。当仅依赖 Top - 1 候选答案(K = 1)时,模型准确率最低,仅为 59.51%。这是因为单一候选答案一旦出现幻觉或错误,后续的文本匹配环节将无法定位到正确的支持帧,导致推理链条断裂。随着 K 值从 2 增加到 5,准确率和 ANLS 指标稳步增长。当 K = 5 时,模型达到了最佳性能(准确率 64.23%,ANLS 72.02%)。这一结果

表明,通过引入更多的候选答案,我们有效地扩大了对视频内容的文本搜索空间,从而提高了找回潜在关键帧的召回率。结合本文的自适应去重机制,这种策略确保了模型能够从多个可能的视角聚合视觉证据,从而做出更稳健的回答。

表6 M4-ViteVQA 验证集上候选答案数量对模型性能的影响

Table 6 Impact of the number of candidate answers on model performance on the M4-ViteVQA validation set.

候选答案数量	评估指标	
	Acc.(%) ↑	ANLS.(%) ↑
1 个候选答案	59.51	68.04
2 个候选答案	61.69	69.95
3 个候选答案	63.06	71.07
4 个候选答案	63.32	71.22
5 个候选答案	64.23	72.02

注：“↑”表示数值越高越好，加粗字体为每列最优值。

2.4 计算效率分析

2.4.1 不同输入帧数配置下的性能对比

为了探究不同输入帧数配置对模型性能的影响,本文测试了减少第一阶段输入帧数的多种配置方案,结果如表 7 所示。

本文方法采用 10+3 帧配置(即第一阶段均匀采样 10 帧生成候选答案,第二阶段选择 3 帧精炼推理),虽然取得了最优性能,但相比基线的 10 帧输入确实增加了额外计算开销。当第一阶段输入帧数减少至 7 帧或 5 帧时(即 7+3 和 5+3 配置),模型性能虽有小幅下降,但仍优于均匀采样基线。

特别是 5+3 配置,仅需 8 帧总输入即实现了 64.89% 的准确率和 72.73% 的 ANLS,分别超出基线 2.38 和 1.77 个百分点,表明本文方法在降低计算开销的同时仍能保持显著的性能优势。

2.4.2 计算复杂度分析

为进一步量化本文方法的计算开销,表 8 从自注意力机制的角度分析了不同输入帧配置下的计算复杂度。由于多模态大语言模型中自注意力机制的计算复杂度与输入序列长度的平方成正比,本文以视觉 token 数(V)为主导项进行分析($V \gg Q$,其中 Q 为文本 token 数)。在 10+3 配置下,相对计算开销为

表7 不同输入帧数配置在 M4-ViteVQA 验证集上的性能对比

Table 7 Performance comparison of different input frame configurations on M4-ViteVQA validation set.

方法	输入帧数	评估指标	
		Acc.(%) ↑	ANLS.(%) ↑
Baseline	10	62.51	70.96
本文	10+3	65.55	73.23
本文	7+3 +	65.14	72.96
本文	5+3	64.89	72.73

注:“↑”表示数值越高越好,加粗字体为每列最优值。

109%,略高于基线;但7+3配置的主导项仅为58V²,相对开销降至58%;5+3配置更进一步降至34%。这说明通过合理减少第一阶段输入帧数,本文方法能够在保持性能优势的同时显著降低计算开销。

2.5 与强基线方法的对比

在同一基础模型(Qwen2. 5-VL-7B)和同一帧预算(10+3帧总输入)下进行公平对比,本文补充了三种代表性的选帧基线方法:(1) Qwen2. 5-VL推理选帧——直接利用 Qwen2. 5-VL 对 10 帧输入进行推理,选取模型注意力权重最高的 3 帧作为关键帧;(2) CLIP 视觉相似度选帧——利用 CLIP 计算问题文本与各帧的视觉相似度,选取相似度最高的 3 帧;(3) 场景文本检索选帧——提取各帧的场景文本,通过文本嵌入模型计算与问题的语义相似度,选取最相关的 3 帧。

如表9所示,上述三种基线方法的性能均低于

本文方法。其中, Qwen2. 5-VL 直接选帧取得了 63. 67% 的准确率, CLIP 选帧为 63. 62%, 场景文本选帧仅为 62. 56%。相比之下,本文方法通过文本-语义协同机制将 Levenshtein 编辑距离匹配与 CLIP 视觉语义互补相结合,取得了 65. 55% 的最优准确率,分别超出上述基线 1. 88、1. 93 和 2. 99 个百分点,验证了本文提出选帧策略的有效性。

2.6 可视化分析

图4直观展示了本方法与基线模型在四个视频场景下的定性对比结果。图中完整呈现了原始视频序列、基线模型的预测结果以及本方法的预测答案。特别地,图中被标记为绿色五角星的帧,代表了经由本文提出的基于文本-语义协同的关键帧选择模块筛选出的关键帧。

首先,可视化结果表明本方法在精准捕获细粒度文本线索方面具有显著优势。在涉及场景文本理解的问答任务中,关键语义信息往往只在视频的特定时刻短暂出现,极易被忽略。例如,在左上角的第一人称射击游戏场景(问题:“Who has 22 hp?”)中,血条数值“22 hp”仅在少数几帧中清晰可见。基线模型由于采用均匀采样策略,只有少数帧包含该关键信息,多数帧引入了无关信息,导致回答错误(“karrigan”);相比之下,本方法通过计算光学字符识别引擎文本与候选答案的 Levenshtein 编辑距离,成功定位并锁定了包含“22 hp”和玩家用户名的关键帧(如绿色星号所示),从而推理出正确答案“hunter”。类似地,在右下角的足球比赛场景(问题:“What is written on the orange clothes?”)中,尽管守门员球衣上的文字“carlsberg”受到运动模糊和镜

表8 不同输入帧配置下的计算复杂度分析

Table 8 Computational complexity analysis under different input frame configurations.

方法	输入帧数	输入 token 数	复杂度	主导项 (V ≫ Q)	相对计算开销 ↓
Baseline	10	Q + 10V	$O((Q + 10V)^2) = O(Q^2 + 20QV + 100V^2)$	100V ²	100%
本文	10+3	(Q + 10V) + (Q + 3V)	$O((Q + 10V)^2 + (Q + 3V)^2) = O(2Q^2 + 26QV + 109V^2)$	109V ²	109%
本文	7+3+ +	(Q + 7V) + (Q + 3V)	$O((Q + 7V)^2 + (Q + 3V)^2) = O(2Q^2 + 20QV + 58V^2)$	58V ²	58%
本文	5+3	(Q + 5V) + (Q + 3V)	$O((Q + 5V)^2 + (Q + 3V)^2) = O(2Q^2 + 16QV + 34V^2)$	34V ²	34%

注:“↓”表示数值越低越好

表9 不同选帧策略在 M4-ViteVQA 验证集上的性能对比
Table 9 Performance comparison of different frame selection strategies on M4-ViteVQA validation set.

方法	评估指标	
	Acc.(%) ↑	ANLS.(%) ↑
Qwen2.5-VL推理选帧	63.67	71.61
CLIP视觉相似度选帧	63.62	71.38
场景文本检索选帧	62.56	69.84
本文		
+	65.55	73.23
+		

注：“↑”表示数值越高越好，加粗字体为每列最优值。

头切换的干扰，本方法筛选出的关键帧仍能准确捕捉到包含清晰文字特征的特写视角，有效避免了基

线模型因关键视觉信息缺失而产生的幻觉 (“reina”)。

本方法通过筛选高价值帧，有效抑制了视觉冗余与噪声对推理过程的干扰。通过对比观察可见，本方法筛选出的关键帧通常是视频序列中信息密度最高、与问题语义相关性最强的时刻。这种策略有效地过滤了视频中大量存在的过渡帧、模糊帧等视觉冗余，使得多模态大语言模型在推理阶段能够将注意力完全聚焦于核心视觉证据。与之形成鲜明对比的是，基线模型虽然具备强大的通用多模态理解能力，但受限于固定采样或均匀采样策略，往往因输入了包含误导性信息或缺乏关键线索的帧，最终导致推理失败。这进一步印证了本文提出的关键帧选择策略在复杂动态场景下的鲁棒性与有效性。



图4 本方法与基线模型的可视化定性对比(红色叉号表示基线模型的错误预测,绿色对勾表示本方法的正确预测)
Fig. 4 Visual qualitative comparison between our method and the baseline (red crosses denote incorrect predictions by the baseline, green checks denote correct predictions by ours)

3 结论

针对现有场景文字视频问答任务中计算开销大与关键细粒度文本信息易丢失的矛盾,本文提出了一种文本语义协同筛选与区域感知的场景文字视频问答方法。具体而言,本文提出基于文本-语义协同的关键帧选择策略,利用多模态大语言模型生成的候选答案作为“语义锚点”,通过计算 Levenshtein 编辑距离精准定位包含关键文本线索的帧,并辅以 CLIP 视觉语义补充机制,在大幅降低输入帧数的同时有效解决了稀疏文本信息的丢失问题。此外,设

计了文字区域感知模块,通过基于掩膜的亮度衰减策略构建视觉“聚光灯”效应,有效抑制了无关背景噪声,强制模型聚焦于高价值文本区域,并结合置信度融合策略实现了答案的动态优选。在 M4-ViteVQA、RoadTextVQA 和 NewsVideoQA 三个公开数据集上的实验结果表明,本文方法在准确率和 ANLS 指标上均优于现有的先进方法,验证了所提框架在处理场景文字视频问答中细粒度文本理解任务上的有效性与鲁棒性。

然而,当前方法的性能在一定程度上依赖于初始候选答案的质量以及光学字符识别引擎的检测精度。当视频环境极度复杂导致光学字符识别失效,

或初始候选答案出现严重幻觉时, 关键帧检索的准确性可能会受到影响。未来将考虑引入端到端的文本特征学习模块, 探索更高效的特征交互方式, 以进一步提升模型在极端场景下的泛化能力。

参考文献 (References)

- Bai J, Bai S, Yang S, Wang S, Tan S, Wang P, et al. 2023. Qwen-VL: A frontier large vision-language model with versatile abilities [EB/OL]. [2026-01-25].
<https://arxiv.org/abs/2308.12966>
- Bai S, Chen K, Liu X, Wang J, Ge W, Song S, et al. 2025. Qwen2.5-VL technical report [EB/OL]. [2026-01-25].
<https://arxiv.org/abs/2502.13923>
- Cheng Z, Leng S, Zhang H, Xin Y, Li X, Chen G, et al. 2024. VideoL-LaMA 2: Advancing spatial-temporal modeling and audio understanding in video-LLMs [EB/OL]. [2026-01-25].
<https://arxiv.org/abs/2406.07476>
- Dai W, Li J, Li D, Tiong A M H, Zhao J, Wang W, et al. 2023. InstructBLIP: Towards general-purpose vision-language models with instruction tuning // Proceedings of the 37th Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 49250-49267 [DOI: 10.52202/075280-2142]
- Du Y, Li C, Guo R, Yin X, Liu W, Zhou J, et al. 2020. PP-OCR: A practical ultra lightweight OCR system [EB/OL]. [2026-01-25].
<https://arxiv.org/abs/2009.09941>
- Fang C Y, Jiang W H, Fang Y M, Peng Y X and Liu Y. 2025. Separate, locate, and align: Determine context relation of scene text from multiple perspectives in TextVQA. IEEE Transactions on Circuits and Systems for Video Technology, 35(11): 11172-11185 [DOI: 10.1109/TCSVT.2025.3577188]
- Fang C Y, Li J N, Li L, Ma C and Hu D Y. 2023. Separate and locate: Rethink the text in text-based visual question answering // Proceedings of the 31st ACM international conference on multimedia. Ottawa, Canada: Association for Computing Machinery: 4378-4388 [DOI: 10.1145/3581783.3611753]
- Fang C Y, Zhu C, Jiang W H, Fang Y M and Yan J B. 2026. Dual adaptive redundancy elimination for training-free video question answering. Journal of Software, 37(5): 1950-1963 (方承炀, 朱畅, 姜文晖, 方玉明, 鄢杰斌. 2026. 面向免训练视频问答的双重自适应冗余消除. 软件学报, 37(5): 1950-1963) [DOI: 10.13328/j.cnki.jos.007544]
- Gao P, Han J, Zhang R, Lin Z, Geng S, Zhou A, et al. 2023. LLaMA-Adapter V2: Parameter-efficient visual instruction model [EB/OL]. [2026-01-25].
<https://arxiv.org/abs/2304.15010>
- Hu R H, Singh A, Darrell T and Rohrbach M. 2020. Iterative answer prediction with pointer-augmented multimodal transformers for TextVQA // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 9989-9999 [DOI: 10.1109/CVPR42600.2020.01001]
- Jahagirdar S, Mathew M, Karatzas D and Jawahar C V. 2023. Watching the news: Towards VideoQA models that can read // Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 4430-4439 [DOI: 10.1109/WACV56688.2023.00442]
- Lei J, Berg T L and Bansal M. 2023. Revealing single frame bias for video-and-language learning // Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Toronto, Canada: Association for Computational Linguistics: 487-507 [DOI: 10.18653/V1/2023.ACL-LONG.29]
- Li B, Zhang Y, Chen L, Wang J, Pu F, Cahyono J A, et al. 2025a. Otter: A multi-modal model with in-context instruction tuning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 47(9): 7543-7557 [DOI: 10.1109/TPAMI.2025.3571946]
- Li J, Li D, Savarese S and Hoi S C H. 2023. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models // Proceedings of the 40th International Conference on Machine Learning. Honolulu, USA: PMLR: 19730-19742
- Li K, Wang Y, He Y, Li Y, Wang Y, Liu Y, et al. 2024. MVBench: A comprehensive multi-modal video understanding benchmark // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 22195-22206 [DOI: 10.1109/CVPR52733.2024.02095]
- Li K C, He Y, Wang Y, Li Y, Wang W, Luo P, et al. 2025b. VideoChat: chat-centric video understanding. Science China Information Sciences, 68(10): 200102 [DOI: 10.1007/s11432-024-4321-9]
- Lin B, Ye Y, Zhu B, Cui J, Ning M, Jin P, et al. 2024. VideoLLaVA: Learning united visual representation by alignment before projection // Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Miami, USA: Association for Computational Linguistics: 5971-5984 [DOI: 10.18653/V1/2024.EMNLP-MAIN.342]
- Liu Z J, Zhu L G, Shi B F, Zhang Z Y, Lou Y M, Yang S, et al. 2025. NVILA: Efficient frontier visual language models // Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 4122-4134 [DOI: 10.1109/CVPR52734.2025.00390]
- Maaz M, Rasheed H A, Khan S and Khan F. 2024. Video-ChatGPT: Towards detailed video understanding via large vision and language models // Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Bangkok, Thailand: Association for Computational Linguistics: 12585-12602 [DOI: 10.18653/V1/2024.ACL-LONG.679]
- Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. 2021. Learning transferable visual models from natural language supervision // Proceedings of the 38th International Conference on

- Machine Learning. Virtual Event: PMLR: 8748-8763
- Su Y, Lan T, Li H, Xu J, Wang Y and Cai D. 2023. PandaGPT: One model to instruction-follow them all//Proceedings of the 1st Workshop on Taming Large Language Models: Controllability in the era of Interactive Assistants. Prague, Czech Republic: Association for Computational Linguistics: 11-23
- Tang X, Qiu J H, Xie L X, Tian Y J, Jiao J B and Ye Q X. 2025. Adaptive keyframe sampling for long video understanding//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 29118-29128 [DOI: 10.1109/CVPR52734.2025.02711]
- Tom G, Mathew M, Garcia-Bordils S, Karatzas D and Jawahar C V. 2023. Reading between the lanes: Text VideoQA on the road//Proceedings of the 17th International Conference on Document Analysis and Recognition (Part VI). San José, USA: Springer: 137-154 [DOI: 10.1007/978-3-031-41731-3_9]
- Tong Z, Song Y B, Wang J and Wang L M. 2022. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training//Proceedings of the 36th Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 10078-10093 [DOI: 10.52202/068431-0732]
- Wang F, Shi F Y, Zhao J, Zhang X S and Wang X F. 2023a. Answer mask-fused visual question answering model. Journal of Image and Graphics, 28(11): 3562-3574 (王峰, 石方宇, 赵佳, 张雪松, 王雪枫. 2023. 融合答案掩码的视觉问答模型. 中国图象图形学报, 28(11): 3562-3574) [DOI: 10.11834/jig.211137]
- Wang J P, Ge Y X, Yan R, Ge Y Y, Lin K Q, Tsutsui S, et al. 2023b. All in one: Exploring unified video-language pre-training//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 6598-6608 [DOI: 10.1109/CVPR52729.2023.00638]
- Wang J, Yang Z, Hu X, Li L, Lin K, Gan Z, et al. 2022a. GIT: A generative image-to-text transformer for vision and language. Transactions on Machine Learning Research
- Wang Y, Li K, Li Y, He Y, Huang B, Zhao Z, et al. 2022b. InternVideo: General video foundation models via generative and discriminative learning[EB/OL].[2026-1-25].
https://arxiv.org/abs/2212.03191
- Xiao J F, Jiang W H, Fang Y M, Fang C Y and Zhao X W. 2026. Video captioning with trajectory-based spatiotemporal perceiving and adaptive semantic focusing. Journal of Image and Graphics, 31(2): 0556-0572 (肖景富, 姜文晖, 方玉明, 方承扬, 赵小伟. 2026. 融合轨迹时空感知与自适应语义聚焦的视频描述方法. 中国图象图形学报, 31(2): 0556-0572) [DOI: 10.11834/jig.250266]
- Xie J L, Zhang R F and Li G B. 2025. Video question answering with large language models: a survey. Journal of Image and Graphics, 30(12): 3760-3781 (谢君琳, 张锐斐, 李冠彬. 2025. 大语言模型下的视频问答方法综述. 中国图象图形学报, 30(12): 3760-3781) [DOI: 10.11834/jig.240535]
- Xu L, Zhao Y, Zhou D, Lin Z, Ng S K and Feng J. 2024. PLLaVA: Parameter-free LLaVA extension from images to videos for video dense captioning[EB/OL].[2026-01-25].
https://arxiv.org/abs/2404.16994
- Yan S, Zhu T, Wang Z, Cao Y, Zhang M, Ghosh S, et al. 2022. VideoCoCa: Video-text modeling with zero-shot transfer from contrastive captioners[EB/OL].[2026-01-25].
https://arxiv.org/abs/2212.04979
- Yang A, Miech A, Sivic J, Laptev I and Schmid C. 2021. Just ask: Learning to answer questions from millions of narrated videos//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 1666-1677 [DOI: 10.1109/ICCV48922.2021.00171]
- Ye Q H, Xu H Y, Xu G H, Ye J B, Yan M, Zhou Y, et al. 2023. mPLUG-Owl: Modularization empowers large language models with multimodality[EB/OL].[2026-01-25].
https://arxiv.org/abs/2304.14178
- Zhang Y, Zeng G Y, Shen H W, Wu D Q, Zhou Y and Ma C. 2025a. Track the answer: Extending TextVQA from image to video with spatio-temporal clues//Proceedings of the 39th AAAI Conference on Artificial Intelligence. Philadelphia, USA: AAAI Press: 10275-10283 [DOI: 10.1609/AAAI.V39I10.33115]
- Zhang Y, Zeng G Y, Wu D Q, Shen H W, Li B B, Zhou Y, et al. 2025b. Gather and trace: Rethinking video TextVQA from an instance-oriented perspective//Proceedings of the 33rd ACM International Conference on Multimedia. Dublin, Ireland: ACM: 876-885 [DOI: 10.1145/3746027.3754718]
- Zhao M Y, Li B J, Wang J, Li W Q, Zhou W J, Zhang L, et al. 2022. Towards video text visual question answering: Benchmark and baseline//Proceedings of the 36th Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 35549-35562 [DOI: 10.5555/3600270.3602846]

作者简介

方承扬,男,讲师,主要研究方向为计算机视觉。E-mail: fangchengyang@jxufe.edu.cn

章楠,男,硕士研究生,研究方向为计算机视觉。E-mail: 2202426584@stu.jxufe.edu.cn

姜文晖,通信作者,男,副教授,主要研究方向为计算机视觉、多模态信号处理。E-mail: wenhui@jxufe.edu.cn

刘学林,男,讲师,主要研究方向为视觉质量评价和计算机视觉。E-mail: xuelinliu-bill@foxmail.com

方玉明,男,教授,主要研究方向为计算机视觉、多媒体信号处理、视觉质量评估。E-mail: fa0001ng@e.ntu.edu.sg