 DOI: https://doi.org/10.12345/bdai.v7i2.37658

Temporal Alignment and Retrieval-Augmented Method for Long Video Clip Localization

Yikai Wang Qiwei Shen

Beijing University of Posts and Telecommunications, Beijing, 100876, China

Abstract

Long video clip localization requires models to simultaneously achieve long-temporal semantic matching and fine-grained temporal alignment. However, existing large video-language models (LVLMs) suffer from issues such as context redundancy, semantic drift, and "evidence–time" mismatch. Expanding visual tokens easily increases computational overhead and introduces noise, while semantic similarity-based video retrieval methods are prone to temporal drift. To address these problems, this paper proposes the TAEC-RAG framework, which avoids blind expansion of visual context: it extracts multi-source timestamped evidence to build a library, converts redundant information into controllable evidence units through fragmentation and compression, suppresses temporal drift via temporal consistency constraints, and feeds compact evidence together with queries into LVLMs for enhanced reasoning. Experimental results on long video benchmarks validate that TAEC-RAG stably improves generative localization performance across different query granularities, with particularly significant gains in event-level localization.

Keywords

Long Video Moment Retrieval; Retrieval-Augmented Generation; Temporal Alignment; Evidence Compression

长视频片段定位的时间对齐与检索增强方法

王羿凯 沈奇威

北京邮电大学，中国 · 北京 100876

摘 要

长视频片段定位需模型兼顾长时序语义匹配与精细时间对齐，但现有大型视频 - 语言模型（LVLMs）存在上下文冗余、语义漂移及“证据 — 时间” 错配等问题。扩大视觉 token 易增计算开销与噪声，语义相似度检索法则易出现 temporal drift。为此，本文提出 TAEC-RAG 框架，无需盲目扩展视觉上下文：提取多源带时间戳证据构建库，经片段化压缩转化为可控证据单元，通过时间一致性约束抑制 temporal drift，将紧凑证据与查询输入 LVLM 实现增强推理。实验验证，该方法在长视频基准不同查询粒度下稳定提升定位性能，事件级定位增益尤为显著。

关键词

长视频片段定位；检索增强生成；时间对齐；证据压缩

1 介绍

近年来，大规模语言模型与多模态学习的快速发展推动了大型视频 – 语言模型的兴起，使模型能够在视频与文本之间进行更强的语义理解与推理能力 [1]。在此基础上，长视频片段定位作为连接视频理解与下游应用（如视频问答、事件检索与内容导航）的关键任务，逐渐受到关注。该任务要求模型在完整、连续的视频时间轴上，根据查询语义精确预测对应的时间区间，对模型的时序建模与时间对齐能力提出了更高要求。然而，尽管现有 LVLMs 在短视频或片段级输入上已取得较好效果，其在长视频片段定位场景中仍面临显著瓶颈。

针对上述问题，现有改进思路主要有两类：一是通过扩展上下文长度增强长程建模，但计算开销大且冗余信息可能加剧噪声干扰；二是采用检索增强生成（RAG）范式补全证据，但现有方法多将视频“文本化”，削弱了检索与原始时间轴的对齐关系，并容易出现“时间漂移”（temporal drift）——检索到语义相关但时间错位的证据，导致定位偏差。

为此，本文提出 TAEC-RAG（ Temporal-Aligned Evidence Compression Retrieval-Augmented Generation），一种面向长视频片段定位任务的时间对齐证据压缩式检索增强框架。TAEC-RAG 的核心目标是在不盲目扩大视觉上下文规模的前提下，将长视频中分散且冗余的多源信息压缩为可检索、可对齐且长度可控的证据单元，并通过显式的时间一致性约束确保检索证据与视频时间轴同步，从而提升长视频片段定位的鲁棒性与稳定性。具体而言，TAEC-RAG 首先从视频

【作者简介】王羿凯（2001–），男，中国山西晋中人，硕士，从事多模态处理研究。

中提取多源辅助文本证据（自动语音识别 ASR、场景文字识别 OCR、目标/动作检测 DET），构建带时间戳的证据库；随后，针对查询在语义空间进行候选证据召回，并引入时间对齐约束对检索结果进行重排序与筛选，以抑制长视频检索中的 temporal drift；最终，将检索得到的紧凑证据与查询共同输入 LVLM，形成以证据为驱动的上下文增强生成过程，直接输出目标片段对应的时间区间。

本文主要贡献如下：1）提出 TAEC-RAG 框架，在有限上下文预算下提升生成式模型的时间定位稳定性；2）构建带时间戳的多源证据表示，通过时间一致性约束缓解时间漂移；3）在长视频片段定位基准上验证了该方法在不同查询粒度下均能稳定提升定位性能。

2 相关方法

2.1 面向视频理解的大语言模型

近年来，大语言模型（Large Language Models, LLMs）的快速发展推动了通用视频 – 语言系统的持续演进。早期工作如 Video-ChatGPT 通过对视频帧进行独立编码，并借助时空池化实现视频级特征融合；VideoChat 在外观特征的基础上引入在线生成的文本描述，以构建更丰富的片段语义表示。为缩小图像与视频路径之间的表征差距，Video-LLaVA 进一步设计共享投影层，将图像与视频特征对齐至统一的语言潜空间，其后续工作 LLaVA-NeXT-Video 则在 LLaVA-NeXT 主干模型基础上进行针对性的视频微调，以提升视频理解能力。尽管上述方法在跨模态对齐与语义建模方面取得了显著进展，但在长视频场景中仍存在明显局限：面对信息密度高、语义层次复杂且依赖细粒度时序关系的视频内容，现有方法在捕捉长程时间依赖与细微语义演化方面仍显不足。

2.2 基于生成范式的长视频片段定位方法

针对长视频片段定位任务，生成式方法（generation-based）提供了一种不同于传统检索范式的解决思路。该类方法通常将定位问题建模为时间区间生成任务：在推理阶段，将视频以均匀下采样帧序列或原始 mp4 形式输入模型，并显式提供视频总时长及每帧对应的时间戳信息，通过指令提示模型直接生成包含答案的一段或多段时间区间。为保证评测公平性，生成式方法通常与检索式方法采用一致的基于 IoU 的评估指标，并遵循各模型官方推荐的输入帧数与时间指令格式，同时通过不同帧数设置的消融实验分析上下文长度对定位性能的影响。在基线模型选择上，既包括专为时间片段定位设计的方法（如 TimeChat、LITA），也涵盖主流开源多模态大模型（如 LLaVA-Video、InternVL3、Qwen2.5-VL）以及面向长视频理解的模型（如 VideoLLaMA3、Eagle2.5、VideoChat-Flash），并可进一步引入 GPT-4o、Gemini-2.5 Pro 等闭源模型作为能力上界参考，从而形成对生成式长视频定位方法的系统性对比分析。

3 方法

本文提出一种无需训练（training-free）的生成式长视频片段定位流程，由三阶段组成：

1）时间对齐证据抽取与压缩：从长视频中提取多源证据，并压缩为可控长度、带时间戳的证据片段集合；2）时间窗检索与一致性约束：针对查询召回语义相关证据，并通过时间窗聚合获得候选时间范围，抑制长视频中的 temporal drift 与位置偏置；3）双层生成推理：冻结 VLM 先对证据进行结构化归纳（Level-1），再生成最终区间预测（Level-2），并输出可解释的证据引用。

3.1 时间对齐证据抽取与压缩

从视频中提取多源证据（如 ASR/OCR/DET/CAP），统一表示为带时间戳的四元组：

$$e_k=(m_k,x_k,[t_k^s,t_k^e],c_k)$$

其中m_k为证据模态类型，x_k为文本化内容，$[t_k^s,t_k^e]$为覆盖时间范围，c_k为置信度。所有原子证据构成集合：

$$\varepsilon = \{e_k\}_{k=1}^{|\varepsilon|}$$

为了便于统一检索，我们将证据文本映射到向量空间：

$$\mathrm{z_k}=f_{emb}(x_k),$$

$$z_k=f_{emb}(Q),$$

其中$f_{emb}(\,)$是冻结的文本嵌入器

长视频中证据碎片化明显，我们将相邻且语义相近的原子证据合并为证据片段 s:

$$z_k=f_{emb}(Q),$$

合并判据可用“时间邻近 + 语义相似”形式化：

$$\Delta_t(e_i,e_j) = max(0,t_j^s-t_i^e)$$

$$sim(e_i,e_j) = cos(z_i,z_j)$$

当$\Delta_t(e_i,e_j)$较小且 $sim(e_i,e_j)$较大时，将二者归于同一片段。

片段文本通过压缩函数获取：

$$c_{s_u} = g(\varepsilon_u)$$

其中 $g(\,)$可由轻量摘要进行完成，最终得到证据片段库：

并将片段向量表示为 $z_{s_u}=f_{emb}(c_{s_u})$

3.2 时间窗检索与一致性约束

基于语义相似度召回候选证据片段集合：

$$S_k = TopK_{s\in S}cos(Z_q,Z_s)$$

考虑到之前仅用语义相似易出现语义漂移，我们引入了时间一致性项，构造联合打分：

$$Score(s) = \alpha\cdot cos(Z_q,Z_s)+(1-\alpha)\varphi(s_u)$$

其中$\varphi(s_u)$为基于时间分布的连贯选项。通过时间窗聚合获得候选区间作为生成阶段的时间锚点：

$$[t_s^{(0)},t_e^{(0)}] = arg\ \max_{\omega} \sum_{s_u \in \omega} Score(s_u)$$

3.3 双层生成推理

我们采用冻结 *VLM* F_θ（θ固定）进行两部生成，以提升输出结构化与可解释性。首先将查询 Q、视频时长 T、粗候选窗口 ω 及证据集合 E，组织为提示 P_1, 生成结构化证据归纳结果：

$$C=F_\theta\big(P_1(Q,T,\omega,C,E)\big)$$

其中 C 包含关键事实、时间边界线索及对应证据引用。随后，在 C 基础上构造提示 P_2，生成最终区间预测：

$$I= F_\theta\big(P_2(Q,T,\omega,C,E)\big)$$

为抑制生成式模型在长视频中的时间幻觉，我们对输出施加时间约束：预测区间需满足合法性，并与至少一条证据在时间上重叠。要求 $IoU([t_s,t_e],[T_s^s,T_s^e])>0$。此外，通过窗口一致性约束鼓励预测区间位于粗候选窗口附近。最终，对高度重叠的区间进行合并，输出区间集合及其证据引用，形成可解释的定位结果。

4 实验

4.1 数据集

MomentSeeker 是一个专为长视频片段定位任务构建的基准数据集，旨在系统性评估模型在长时序视频中的精细时间定位与语义对齐能力。数据集中的视频通常具有较长时长，涵盖多个连续或交叠的事件过程，使得目标片段往往仅占据视频整体的一小部分，从而显著增加了定位难度。MomentSeeker 支持多种查询形式，包括文本查询以及图像或视频条件查询，要求模型在完整视频时间轴上输出一个或多个与查询语义高度匹配的时间区间。该数据集按照任务语义粒度进一步划分为 Global-level、Event-level 和 Object-level 等不同类型，用以分别考察模型对整体语境、具体事件以及细粒度对象或状态变化的理解能力。

4.2 实现细节

我们将 TAEC-RAG 作为生成式定位推理前置模块，统一采用“多源证据抽取—证据压缩—时间一致性检索—双层生成”的 training-free 流程。对于含音轨的视频，首先从 mp4 中提取音频并使用现成 Whisper 得到带时间戳的分段转写，再对持续时间过短或信息量较低的片段执行相邻合并以减少碎片化；对于所有视频均可用的 OCR 通道，我们按每两秒采样关键帧并运行 OCR，其中仅保留文本与时间戳用于后续处理。引入时间一致性项抑制长视频中的 temporal drift，其中 采用候选证据中心时间的密度峰值一致性。打分函数中的其中用于权衡语义相关性与时间连贯性：当证据噪声较大或视频更长时适当降低以强调时间一致性，而当 ASR/CAP 等证据质量较高时可提高以增强语义召回，我们最终使用进行实验，随后在时间轴上对高分证据进行窗口聚合得到粗候选窗口，选取其邻域内证据构成输入冻结 VLM 执行双层生成，视频以均匀下采样帧序列或原始 mp4 提供给模型，Frames 表示输入帧数 (N)，并显式提供总时长 (T) 与每帧时间戳，不同模型的 (N) 取其官方推荐设置（如表中 96/100/768），以保证对比的公平性，实验基线涵盖 TimeChat、LITA、InternVL3、Qwen2.5-VL 等生成式视频多模态模型，评测统一采用基于 temporal IoU 的 R@1 与 mAP@5，并按 Global/Event/Object 及 Overall 进行汇总。

4.3 实验结果

Method	Size	#Frames	Global-level		Event-level		Object-level		Overall	
			R@1	mAP@5	R@1	mAP@5	R@1	mAP@5	R@1	mAP@5
InternVL3	8B	96	6.4	5.8	13.8	15.0	7.1	7.1	10.3	10.7
InternVL3 + TAEC-RAG	8B	96	**6.7**	5.7	**15.8**	**17.1**	7.0	**7.3**	**11.3**	**11.6**
Qwen2.5VL	7B	768	9.8	8.8	22.5	22.9	9.1	9.0	15.7	15.7
Qwen2.5VL + TAEC-RAG	7B	768	9.6	**9.0**	**25.4**	**25.9**	**9.4**	8.9	**16.6**	**16.3**
TimeChat	7B	96	5.3	5.3	16.0	16.0	6.6	6.6	13.1	13.1
TimeChat + TAEC-RAG	7B	96	**5.5**	5.2	**18.2**	**18.1**	6.5	**6.8**	**14.0**	**13.8**
LITA	13B	100	11.8	11.8	18.3	18.3	9.3	9.3	15.8	15.8
LITA + TAEC-RAG	13B	100	**12.0**	11.7	**20.9**	**20.8**	**9.6**	9.2	**16.8**	**16.5**

Table 1 MomentSeeker 数据集在 IoU=0.1 条件下的实验结果，其中 +TAEC-RAG 表示基于该方法的性能

4.4 结果分析

从表 1 实验结果显示，TAEC-RAG 在多种生成式基线模型上实现稳定一致的性能提升，验证了其时间对齐证据压缩与检索增强策略在长视频片段定位任务中的有效性。Event-level 任务中提升最为显著，如 InternVL3（8B）的 R@1 从 13.8 升至 15.8、mAP@5 从 15.0 升至 17.1，Qwen2.5VL 等模型也呈现相同趋势，这表明其时间一致性约束和证据窗口聚合机制能有效缓解 temporal drift 问题，凸显时间对齐证据对细粒度时序推理的重要性。Global-level 与 Object-level 任务中，TAEC-RAG 虽使部分模型个别指标轻微下降，但多数模型在核心指标上仍获提升，尤其 Object-level 的 R@1 增益稳定，这是通过抑制模型 “过度泛化” 换取更可靠时间对齐能力的结果。Overall 指标上，所有评测模型均有明确提升，如 InternVL3 的 Overall R@1 从 10.3 升至 11.3，Qwen2.5VL 从 15.7 升至 16.6，证明无需参数微调，仅通过推理阶段的证据组织与时间对齐策略，即可显著增强模型整体表现。综上，TAEC-RAG 的提升源于时间一致性显式建模，其 training-free 特性在改善长视频场景下模型稳定性与可解释性的同时，尤其适用于对时间定位要求较高的复杂查询场景。

参考文献

[1] Liu Z, Dong Y, Liu Z, et al. Oryx mllm: On-demand spatial-temporal understanding at arbitrary resolution[J]. arXiv preprint arXiv:2409.12961, 2024.

[2] Zhang P, Zhang K, Li B, et al. Long context transfer from language to vision[J]. arXiv preprint arXiv:2406.16852, 2024.

[3] Liu H, Li C, Li Y, et al. Llavanext: Improved reasoning, ocr, and world knowledge[EB/OL].(2024-1)