RAPID: Retrieval-Augmented Parallel Inference Drafting for Text-Based Video Event Retrieval

Long S. T. Nguyen[1,2], Huy G. Nguyen[1,2], Bao G. Khuu[1,2], Huy A. T. Luu[1,2], Huy Q. Le[1,2], Tuan T. Nguyen[1,2], and Tho T. Quan[1,2(✉)]

[1] URA Research Group, Faculty of Computer Science and Engineering, Ho Chi Minh City University of Technology (HCMUT), Ho Chi Minh City, Vietnam
qttho@hcmut.edu.vn

[2] Vietnam National University Ho Chi Minh City, Ho Chi Minh City, Vietnam

Abstract. Retrieving events from videos using text queries has become increasingly challenging due to the rapid growth of multimedia content. Existing methods for text-based video event retrieval often focus heavily on object-level descriptions, overlooking the crucial role of contextual information. This limitation is especially apparent when queries lack sufficient context, such as missing location details or ambiguous background elements. To address these challenges, we propose a novel system called RAPID (Retrieval-Augmented Parallel Inference Drafting), which leverages advancements in Large Language Models (LLMs) and prompt-based learning to semantically correct and enrich user queries with relevant contextual information. These enriched queries are then processed through parallel retrieval, followed by an evaluation step to select the most relevant results based on their alignment with the original query. Through extensive experiments on our custom-developed dataset, we demonstrate that RAPID significantly outperforms traditional retrieval methods, particularly for contextually incomplete queries. Our system was validated for both speed and accuracy through participation in the Ho Chi Minh City AI Challenge 2024, where it successfully retrieved events from over 300 h of video. Further evaluation comparing RAPID with the baseline proposed by the competition organizers demonstrated its superior effectiveness, highlighting the strength and robustness of our approach.

Keywords: Text-Based Video Event Retrieval · Contextual Query Enrichment · Parallel Inference · Multimedia and Multimodal Retrieval

1 Introduction

The rapid advancement of *Artificial Intelligence* (AI), particularly in the field of Computer Vision, has driven significant breakthroughs in video event retrieval. Initially, research in this area focused on applications in Intelligent Transport

L. S. T. Nguyen and H. G. Nguyen—The two authors contributed equally to this work.

W. Buntine et al. (Eds.): SOICT 2024, CCIS 2353, pp. 181–194, 2025.
https://doi.org/10.1007/978-981-96-4291-5_15

Systems [14], but it has since expanded to include surveillance, security, and multimedia content analysis [27]. With the exponential growth of video data across online platforms, the ability to efficiently retrieve specific events from vast video datasets has emerged as a critical research challenge [22]. Developing more efficient and accurate retrieval methods is essential to managing the overwhelming volume of media content generated daily and improving retrieval accuracy [7].

To address the growing demand for efficient video retrieval solutions, several international competitions, such as the *Lifelog Search Challenge* (LSC) [8] and the *Video Browser Showdown* (VBS) [21], have been organized to foster the development of fast and effective video search methods. Similarly, in Vietnam, the *Ho Chi Minh City AI Challenge 2024*[1] was established to encourage innovation in video event retrieval. In this competition, participants were tasked with retrieving events from over 300 h of 1,500 news videos using a series of sequential text-based queries or short visual clips depicting the events. Notably, these queries often began with broad, context-poor descriptions, and many video segments lacked clear contextual cues, such as background details, making accurate retrieval a significant challenge.

Various approaches to video event retrieval have been explored over the years, leveraging a diverse range of input modalities. These include image frames [23], object tags [2], event labels [16], and even audio data [12]. More advanced methods use high-level sketches to visually represent events [32]. Recent progress in Natural Language Processing, especially with the development of multimodal models like *Contrastive Language-Image Pre-training* (CLIP) [19], has significantly advanced text-based video retrieval. These models map text queries and video frames into a shared embedding space using contrastive learning, enabling more intuitive and effective retrieval [6,17]. However, the effectiveness of text-based retrieval systems heavily depends on the completeness of the query. Queries that lack sufficient context often lead to reduced retrieval accuracy [5], underscoring the need for enriched contextual information [11]. Moreover, certain events are complex and difficult to describe in detail, further complicating the retrieval process.

To address these challenges, we propose a novel method called *Retrieval-Augmented Parallel Inference Drafting* (RAPID), illustrated in Fig. 1. RAPID is designed to enhance text-based video event retrieval, particularly in scenarios where queries lack sufficient context or where frames are challenging to describe. Unlike traditional methods that rely solely on single, unrefined queries, RAPID generates multiple *augmented queries* by enriching the original with additional contextual details, such as location or event-specific information. These *drafts* are created using *Large Language Models* (LLMs) combined with prompt-based learning techniques [4]. By performing *parallel inference*, RAPID retrieves the most relevant keyframes, which are subsequently re-evaluated for alignment with the original query. Additionally, we developed a user-friendly interface to support practical, real-world application of the system.

[1] http://aichallenge.hochiminhcity.gov.vn/.

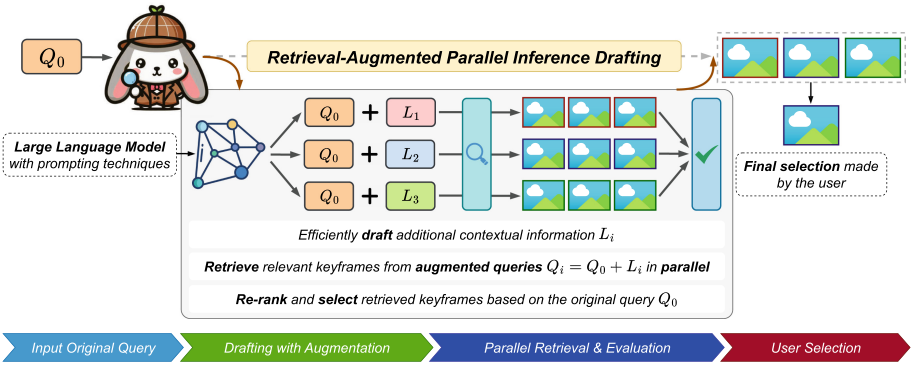

Fig. 1. Overview of the RAPID framework.

We evaluated RAPID using a dataset derived from 300 h of news videos provided by the challenge. Our experimental results demonstrate that enriching queries with contextual information, such as location, significantly improves retrieval accuracy and efficiency. Furthermore, RAPID outperformed the baseline retrieval pipeline provided by the competition organizers, showcasing its ability to handle real-world video retrieval challenges, especially when dealing with incomplete or context-poor queries.

In summary, our key contributions are as follows.

- We developed the RAPID system, which leverages parallel processing of multiple augmented queries to significantly improve video event retrieval performance.
- We demonstrated the effectiveness of RAPID in handling incomplete or context-poor queries, and compared its performance against the baseline provided by the competition, showing notable improvements.
- We successfully applied RAPID in a real-world scenario, achieving excellent results at the Ho Chi Minh City AI Challenge 2024, highlighting its accuracy and speed.

2 Related Work

2.1 Text-Based Video Event Retrieval Methods

Various video event retrieval methods have evolved from traditional text-based queries that match against metadata such as titles and tags [2,10,16]. These methods often fail when metadata is incomplete or when users need to locate complex events that are difficult to describe using simple tags or titles. To address these limitations, visual-based retrieval techniques, such as content-based image retrieval [23] and query-by-high-level sketches [20,32], were introduced to better analyze the visual content of videos. With the notable success of multimodal models like CLIP [19], baseline systems that retrieve both text and image

frames from videos in a shared vector space have become increasingly effective [17,18]. These advancements have enabled more intuitive and accurate retrieval processes, especially when handling large-scale datasets. However, despite the improved performance of modern retrieval systems, challenges remain, particularly in aligning different modalities. This process requires substantial computational resources and poses scalability challenges for real-time applications [26]. Moreover, these methods heavily rely on the accuracy and completeness of text queries as the primary alignment mechanism. When event descriptions are vague or imprecise, system effectiveness is reduced. Research has shown that including contextual components in queries, such as location information, can significantly enhance model performance, as highlighted in [5,11].

2.2 Augmented Query Strategies for Video Event Retrieval

The rapid evolution of LLMs, combined with the power of prompt-based learning across various tasks [4], has driven significant interest in developing more sophisticated systems for information retrieval [1]. By improving the semantic embedding of queries, LLMs can enhance retrieval systems by providing richer, more context-aware representations of the user's search intent. This is particularly useful when the original query lacks detailed context or contains ambiguities, as the augmented query versions generated by LLMs can fill in the missing information, thereby improving the overall precision and recall in video retrieval systems [9]. LLMs can also suggest query completions or augmentations, as demonstrated in fashion recommendation systems, where user history is leveraged to improve query suggestions [3]. Similarly, research [1] has shown that LLMs can enhance the semantic content of queries in retrieval systems, making them more context-aware and better aligned with the user's intent. Specifically, in the domain of video event retrieval, studies [30,33] have demonstrated the effectiveness of rewriting queries logically to automatically incorporate additional context, thereby significantly improving retrieval performance. However, despite the benefits, enriching queries with LLM-generated augmentations also carries risks, such as introducing irrelevant or excessive information, which can decrease efficiency and accuracy. Moreover, some events within video datasets are inherently complex, making it difficult to accurately describe their surrounding context through visual inspection alone. This further complicates the creation of detailed and context-rich queries necessary for effective video event retrieval.

3 RAPID: A Novel Approach for Text-Based Video Event Retrieval

3.1 Motivation

Previous studies have demonstrated the potential of augmented retrieval methods for video event retrieval using text-based queries. Building on the concept of *EmotionPrompt*, which highlights how enriching prompts with additional elements, such as positive emotions, can improve LLM performance [15], we

adapt this approach by focusing on context-specific elements, particularly location information, to enhance query relevance. Additionally, inspired by Google's recent *Speculative RAG* research [28], which explores parallel query processing across multiple drafts, we propose *RAPID*, a novel method for text-based event retrieval. RAPID augments the original query by generating several variations enriched with contextual details, such as location information, and processes these queries in parallel to improve both retrieval accuracy and speed. The formal methodology behind RAPID is described in Subsect. 3.2.

3.2 Methodology

The notation used throughout this paper is summarized in Table 1.

Table 1. Notation

Symbol	Description
$\mathbb{T}$	Text space
$\mathbb{I}$	Image space
$\mathcal{M}_{m \times n}$	Matrix with m rows and n columns
$\mathcal{L} : \mathbb{T} \to \mathbb{T}$	Function that transforms the original query into augmented versions by adding varying contextual details, particularly location information, leveraging a Large Language Model
$\mathcal{E} : \mathbb{T} \cup \mathbb{I} \to \mathbb{R}^d$	Multi-modal embedding function that maps both text and images to a shared d-dimensional vector space via a Multi-modal Embedding Model
$\text{Idx}_k : \mathcal{M}_{1 \times n} \to \mathcal{M}_{1 \times k}$	Function that retrieves the indices of the top-k highest values from a vector $\mathcal{M}$ with n values

Task Definition. We define the task of text-based event retrieval, particularly in the context of the Ho Chi Minh City AI Challenge, as identifying a frame $F_c \in \mathcal{F}$, where $\mathcal{F}$ denotes the set of frames relevant to the described event, from a video corpus $\mathcal{V} = \{V_1, V_2, .., V_p\}$ totaling approximately 300 h of footage. Given a text query Q_0, the objective is to retrieve a frame F_c such that $F_c \in \mathcal{F}$ and best matches the content of Q_0. For visual queries, represented by short video clips V_0, we employ a strategy of converting these videos into descriptive text queries Q_0. As a result, both textual and visual queries are handled under a unified text-based retrieval framework.

Data Preprocessing. Each video $V_i \in \mathcal{V} \subset \mathbb{I}$ consists of image frames, where a video recorded at x *frames per second* (fps) produces x frames for every second of playback. To reduce computational complexity, we extract *keyframes*, which are representative frames containing the key visual information for each segment [31]. For this, we employ *TransNet V2* [24], an effective deep learning architecture optimized for fast shot transition detection. For each video V_i, we generate a set of scenes $\mathcal{S}_i = \{S_{i_1}, S_{i_2}, .., S_{i_{s_i}}\}$, where each scene S_{i_j} consists of frames $\{S_{i_{j_1}}, S_{i_{j_2}}, .., S_{i_{j_e}}\} \subset \mathbb{I}$. From each scene j, we select three keyframes: the first, middle, and last frames, as described in Eq. 1.

$$S_{i_{j_{\text{selected}}}} = \{S_{i_{j_1}}, S_{i_{j_{\lfloor \frac{e-1}{2} \rfloor}}}, S_{i_{j_e}}\}. \tag{1}$$

These keyframes are aggregated into the set $\mathcal{F} \subset \mathbb{I}$, as described in Eq. 2.

$$\mathcal{F} = \bigcup_{i=1}^{p} \bigcup_{j=1}^{s_i} S_{i_{j_{\text{selected}}}}, \tag{2}$$

where p is the total number of videos, and s_i is the number of scenes in video i.

RAPID Inference. Let Q_0 represent the initial query, which may lack sufficient context.

The function $\mathcal{L}(Q_0)$ generates a set of *augmented queries*, denoted as $\mathcal{Q} = \{Q_1, Q_2, .., Q_n\}$. Each augmented query Q_i is then embedded as $\mathbf{Q} = \mathcal{E}(\mathcal{Q}) = \{\mathbf{q}_1, \mathbf{q}_2, .., \mathbf{q}_n\}$.

The keyframes obtained from preprocessing, $\mathcal{F} = \{F_1, F_2, .., F_m\}$, where $m = \sum_{i=1}^{p} s_i \times 3$, are also embedded, yielding $\mathbf{F} = \mathcal{E}(\mathcal{F}) = \{\mathbf{f}_1, \mathbf{f_2}, .., \mathbf{f}_m\}$.

Both the augmented *drafts* $\mathcal{Q}$ and the set of keyframes $\mathcal{F}$ are embedded in the shared vector space $\mathbb{R}^d$. To enable *parallel retrieval*, we compute the cosine similarity between the query embeddings $\mathbf{q}_i$ and the keyframe embeddings $\mathbf{f}_j$, as shown in Eq. 3.

$$\mathcal{P} = \text{sim}(\mathbf{Q}, \mathbf{F}) = \frac{\mathbf{Q} \cdot \mathbf{F}^\top}{\|\mathbf{Q}\|\|\mathbf{F}\|} \in \mathcal{M}_{n \times m}. \tag{3}$$

We then retrieve the top-k most similar frames for each draft Q_i, as shown in Eq. 4.

$$\mathcal{C}_i = \text{Idx}_k(\mathcal{P}_{i,:}), \quad \forall i = 1, 2, .., n. \tag{4}$$

This produces a matrix $\mathcal{C} \in \mathcal{M}_{n \times k}$, where each row contains the indices of the top-k frames for each augmented query. These selected frames are flattened into a set $\mathcal{K} = \{F_1, F_2, .., F_w\}$, where $w = n \times k$.

Finally, these frames are re-evaluated based on their similarity to the original query Q_0, as shown in Eq. 5.

$$\mathcal{C}_{\text{final}} = \text{Idx}_K(\text{sim}(\mathcal{E}(Q_0), \mathcal{E}(\mathcal{K}))) \in \mathcal{M}_{1 \times K}, \tag{5}$$

where $\mathcal{C}_{\text{final}}$ represents the indices of the top-K frames that best match the original query Q_0.

The final set of selected frames is presented to the user for evaluation, allowing them to verify whether the retrieved frames not only match the context of the query but also align with the specific information they are searching for.

Multi-modal Embedding and Prompting Techniques. RAPID integrates multi-modal embeddings and advanced prompting strategies to optimize the alignment between text and visual data, enhancing retrieval performance.

A key property of the multi-modal embedding function $\mathcal{E}(\cdot)$, implemented using a Multi-modal Embedding Model, is that for any image-text pair (I_i, T_i), where T_i accurately describes I_i, the cosine similarity between their embeddings, $\cos\theta(\mathcal{E}(I_i), \mathcal{E}(T_i))$, approaches 1. This property ensures that the more semantically aligned the text query is with the image, the closer their embeddings are in the shared vector space $\mathbb{R}^d$, forming the backbone of the RAPID approach.

To enhance the drafting process, denoted as $\mathcal{L}(\cdot)$, RAPID utilizes a Large Language Model combined with few-shot prompting [4], specifically employing the *Chain-of-Thought* (CoT) reasoning technique [29]. The prompt provided to the LLM for augmenting the query q_0 is formally described in Eq. 6.

$$\begin{aligned} \text{prompt}(q_0) &= \sum_{i=1}^{n} \langle q_i, \text{CoT}(q_i), \{q_{\text{augmented}_i^j}\}_j \rangle + q_0, \\ \text{prompt}(q_0) &\rightarrow \{q_{\text{augmented}_0^j}\}_j, \end{aligned} \tag{6}$$

where $q_1, q_2, .., q_n$ represent example queries, and each q_i is associated with a chain of reasoning $\text{CoT}(q_i)$, which includes logical reasoning steps to guide the LLM, such as identifying key content and verifying contextual information. This reasoning process generates a set of augmented queries $\{q_{\text{augmented}_i^j}\}_j$. Once these examples are processed, the LLM receives the target query q_0 and applies the learned patterns to produce a set of augmented queries $\{q_{\text{augmented}_0^j}\}_j$.

3.3 User Inference Design and Interaction

To enable the deployment of RAPID at the Ho Chi Minh City AI Challenge 2024, we developed an intuitive *User Interface* (UI) tailored to RAPID's operational workflow, as illustrated in Fig. 2. The process begins when the user inputs the initial query Q_0 in the `Original Query` field and presses the `Augment` button. This triggers the system to generate N augmented queries $\{Q_1, Q_2, .., Q_N\}$. The user then selects n suitable augmented queries and initiates the inference process by pressing the `Search` button, following the pipeline described in Subsect. 3.2.

After RAPID inference, we obtain a ranked list of the top K keyframes, denoted as $\mathcal{L}_K = \{F_1, F_2, .., F_K\}$. Displaying these keyframes sequentially could result in a cluttered view, as multiple keyframes may originate from the same video V_i. To improve clarity and organization, we group the keyframes in $\mathcal{L}_K$ by their source videos, forming clusters $\{C_{V_1}, C_{V_2}, .., C_{V_m}\}$, where each $C_{V_i} \subset \mathcal{L}_K$ contains all keyframes from video $V_i \in \mathcal{V}$. Within each cluster, the keyframes are ordered according to their appearance in $\mathcal{L}_K$, preserving their relative ranking. These clusters are then displayed on the UI in the order of the first appearance

Fig. 2. An illustration of RAPID's UI for the textual query Q_0: *A monk is writing*, where $n = 2$ augmented queries are selected from $N = 6$ generated drafts, and the parameter $K = 600$ specifies the number of final keyframes. The relevant result, highlighted in green, is displayed among the top-ranked keyframes. (Color figure online)

of their keyframes in $\mathcal{L}_K$, ensuring that clusters with higher-ranked keyframes appear first. This structured approach provides an intuitive display, allowing users to view related frames grouped by video while maintaining the overall relevance order.

Additionally, we incorporated a filtering feature using an *Optical Character Recognition* (OCR) model, $\mathcal{O}(\cdot)$. This feature enables the system to apply a filter $\mathcal{O}(\mathcal{L}) = \mathcal{L}'_K$, where $\mathcal{L}'_K$ represents the subset of keyframes containing the desired text. This capability is particularly valuable when searching for event frames that include visible text, thereby enhancing the system's effectiveness in retrieving context-specific keyframes.

Another key feature, depicted in Fig. 3, allows users to review a sequence of π keyframes before and after a selected frame F_c, as defined in Eq. 7.

$$\{F_{c-\pi}, .., F_{c-1}, F_c, F_{c+1}, .., F_{c+\pi}\}. \tag{7}$$

In this context, assume F_c is the target frame to be identified, which corresponds to the description Q_0 provided by the organizers. However, visually similar frames such as $F'_c \in V_i$ and $F_c \in V_j$, where $i \neq j$ and $F'_c \approx F_c$, can cause

confusion between events. To address this issue, the organizers provide additional queries Q'_0 that include detailed information about the event, specifically describing the frames surrounding F_c. This additional context helps users accurately identify the correct frame and avoid confusion with visually similar frames like F'_c. This feature is therefore essential for distinguishing between events.

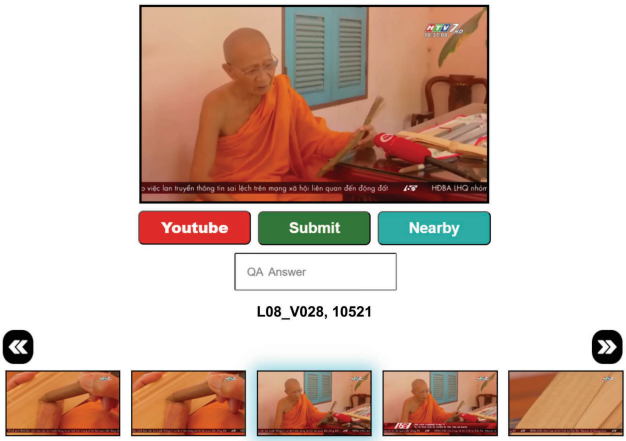

Fig. 3. The user can review frames adjacent to the selected keyframe to check for accuracy before pressing `Submit` and can view the complete video containing it by pressing the `Youtube` button.

4 Experimentation

We conducted two sets of experiments to evaluate the effectiveness of the RAPID system: one comparing retrieval performance between location-augmented queries and non-augmented queries, and the other assessing the overall performance of RAPID against the baseline system provided by the Ho Chi Minh City AI Challenge 2024, referred to as the *AIC'24 Baseline*.

4.1 Experimental Setup

Dataset. We utilized a dataset consisting of 300 h of news video footage provided by the challenge organizers. From this dataset, keyframes were extracted, and we selected keyframes corresponding to specific events, categorized into two types: *Type 1*, with clearly identifiable locations, and *Type 2*, where the location is less discernible. For each set of event keyframes, descriptive queries were generated by focusing on prominent visual elements such as character actions and visible objects. The final dataset consists of 500 records, each containing a `query` (a textual description of the frame) and the corresponding `keyframes` (range of

frame IDs). Of these, 84% are classified as Type 1, while the remaining 16% fall under Type 2.

Evaluation Metrics. To evaluate the system's performance, we employed well-established information retrieval metrics, including *Mean Reciprocal Rank* (MRR), *Precision@k* (P@k), and *Recall@k* (R@k) [25]. MRR measures how early the correct frame is retrieved, Precision@k evaluates the proportion of relevant results in the top-k retrieved frames, and Recall@k assesses the system's ability to retrieve all relevant frames within the top-k. These metrics are highly suitable for evaluating text-based video event retrieval systems, as they emphasize both the accuracy and coverage of retrieved results, which are critical for identifying specific events in large video datasets.

Implementation Details. For the RAPID baseline system, as outlined in Subsect. 3.2, we employed pre-trained models from the Sentence Transformers framework[2] for the Multi-modal Embedding Model. Specifically, we experimented sequentially with two models, namely clip-ViT-L-14 and clip-ViT-B-32, with embedding dimensions of 768 and 512, respectively. To perform fast vector similarity searches, we utilized the FAISS-GPU vector database [13]. For the Large Language Model in the query augmentation phase, we employed the state-of-the-art GPT-4o API from OpenAI[3].

4.2 Evaluation Results

Impact of Location-Augmented Queries. We compared two types of queries: *naive queries*, which focus on describing the subjects and their prominent actions in the event frames as provided in our dataset, and *location-augmented queries*, where the naive queries are enhanced through the RAPID drafting process by incorporating additional location-based context.

The results in Table 2 show a significant improvement in performance across all metrics when using location-augmented queries compared to naive queries, across all types of target frames. Notably, for event frames with difficult-to-interpret contexts, RAPID's augmented queries resulted in a 36%-46% increase in MRR across both models, demonstrating the effectiveness of location augmentation.

Additionally, this experiment demonstrates that the `CLIP-ViT-L/14` model consistently outperforms the `CLIP-ViT-B/32` model. The larger embedding dimension of 768 in the `L/14`, compared to 512 in the `B/32`, allows it to capture more detailed information, particularly when the query is enhanced with location-based context. Therefore, when participating in the AI Challenge, we selected the `CLIP-ViT-L/14` as the multi-modal embedding model for RAPID.

[2] https://sbert.net/.
[3] https://platform.openai.com/docs/models/gpt-4o.

Table 2. Performance Comparison Between Naive and Location-Augmented Queries Across Different Types of Target Frames

Model	Frame Type	Query Type	P@1	P@5	P@10	P@20	R@10	R@20	MRR
CLIP-ViT-B/32	Type 1	Naive	0.181	0.131	0.099	0.667	0.118	0.155	0.269
		Location-augmented	**0.190**	**0.137**	**0.102**	**0.073**	**0.136**	**0.192**	**0.299**
	Type 2	Naive	0.079	0.105	0.075	0.051	0.114	0.154	0.204
		Location-augmented	**0.184**	**0.113**	**0.088**	**0.062**	**0.136**	**0.195**	**0.298**
	All	Naive	0.166	0.127	0.095	0.064	0.118	0.155	0.260
		Location-augmented	**0.192**	**0.134**	**0.100**	**0.071**	**0.136**	**0.193**	**0.300**
CLIP-ViT-L/14	Type 1	Naive	0.220	0.200	0.153	0.100	0.177	0.231	0.334
		Location-augmented	**0.271**	**0.223**	**0.169**	**0.112**	**0.214**	**0.277**	**0.391**
	Type 2	Naive	0.158	0.150	0.101	0.070	0.148	0.209	0.271
		Location-augmented	**0.237**	**0.176**	**0.130**	**0.088**	**0.209**	**0.274**	**0.369**
	All	Naive	0.193	0.144	0.095	0.052	0.173	0.228	0.324
		Location-augmented	**0.216**	**0.163**	**0.108**	**0.056**	**0.213**	**0.276**	**0.388**

Comparison to AIC'24 Baseline. We compared the best version of the RAPID system against the baseline provided by the competition organizers, which was also used by several participating teams, using the dataset we developed.

Table 3. MRR Comparison Between AIC'24 Baseline and RAPID Across Different Target Frame Types

Method	Type 1	Type 2	All
AIC'24 Baseline	0.382	0.362	0.379
RAPID	**0.392**	**0.368**	**0.388**

The results in Table 3 indicate that RAPID shows a slight improvement over the baseline provided by the competition organizers. While the baseline is robust, RAPID demonstrates marginal gains in handling both contextually clear and less discernible target frames.

5 Discussion, Limitations, and Future Works

In this paper, we introduced RAPID, a novel approach designed to enhance the precision of text-based video event retrieval by leveraging LLMs to provide relevant background details and location-based context. This method effectively bridges the semantic gap between users' textual input and the complex visual content of video scenes. Our experiments demonstrated that RAPID achieves

competitive results, improving the system's ability to capture nuanced contextual information and deliver more accurate retrieval outcomes. By utilizing parallel processing, RAPID generates multiple relevant background queries simultaneously, enriching contextual understanding across diverse video scenarios and improving system robustness. Furthermore, augmenting textual input queries significantly optimizes performance by fully utilizing the multi-modal capabilities of the CLIP-based model and the structured representation of video frames, leading to substantial improvements in both retrieval accuracy and contextual relevance.

Despite the promising results, several limitations must be acknowledged. One significant challenge arises when the model encounters scenes with unexpected or uncommon content that LLMs have not been extensively trained on, which may lead to incorrect predictions or suggestions, adversely affecting the accuracy of retrieval outcomes. Additionally, environmental conditions in the keyframes, such as varying light intensity or inadequate background representations, can interfere with scene clarity, leading to irrelevant or misleading contextual information. This reduces the precision of the retrieval process and highlights the need for more robust model generalization across diverse scenarios.

In the future, we plan to incorporate additional inputs, such as filtering with image tags or even audio, to enhance retrieval accuracy in more complex scenarios. Additionally, fine-tuning the multi-modal embedding model offers promising potential for further improving system performance.

References

1. Ai, Q., et al.: Information retrieval meets large language models: a strategic report from Chinese IR community. AI Open **4**, pp. 80–90 (2023). ISSN: 2666-6510. https://doi.org/10.1016/j.aiopen.2023.08.001. https://www.sciencedirect.com/science/article/pii/S2666651023000049
2. Amir, A., et al.: A multi-modal system for the retrieval of semantic video events. Comput. Vis. Image Underst. **96**(2), 216–236 (2004). Special Issue on Event Detection in Video. ISSN: 1077-3142. https://doi.org/10.1016/j.cviu.2004.02.006. https://www.sciencedirect.com/science/article/pii/S107731420400075X
3. Barbany, O., et al.: Leveraging large language models for multimodal search. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1201–1210. IEEE Computer Society, Los Alamitos (2024). https://doi.org/10.1109/CVPRW63382.2024.00127. https://doi.ieeecomputersociety.org/10.1109/CVPRW63382.2024.00127
4. Brown, T., et al.: Language models are few-shot learners. In: Larochelle, H., et al. (eds.) Advances in Neural Information Processing Systems, vol. 33, pp. 1877–1901. Curran Associates, Inc. (2020)
5. Dong, J., et al.: Dual encoding for video retrieval by text. IEEE Trans. Pattern Anal. Mach. Intell. **44**(8), 4065–4080 (2022). https://doi.org/10.1109/TPAMI.2021.3059295
6. Fang, X., et al.: Multi-modal cross-domain alignment network for video moment retrieval. IEEE Trans. Multimedia **25**, 7517–7532 (2023). https://doi.org/10.1109/TMM.2022.3222965

7. Gabeur, V., Sun, C., Alahari, K., Schmid, C.: Multi-modal transformer for video retrieval. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12349, pp. 214–229. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58548-8_13
8. Gurrin, C., et al.: Introduction to the seventh annual lifelog search challenge, LSC'24. In: Proceedings of the 2024 International Conference on Multimedia Retrieval, ICMR 2024, Phuket, Thailand, pp. 1334–1335. Association for Computing Machinery (2024). ISBN: 9798400706196. https://doi.org/10.1145/3652583.3658891
9. Harte, J., et al.: Leveraging large language models for sequential recommendation. In: Proceedings of the 17th ACM Conference on Recommender Systems, RecSys 2023, Singapore, pp. 1096–1102. Association for Computing Machinery (2023). ISBN: 9798400702419. https://doi.org/10.1145/3604915.3610639
10. Hu, W., et al.: A survey on visual content-based video indexing and retrieval. IEEE Trans. Syst. Man Cybern. Part C (Appl. Rev.) **41**(6), 797–819 (2011). https://doi.org/10.1109/TSMCC.2011.2109710
11. Ji, W., et al.: Weakly supervised video moment retrieval via location-irrelevant proposal learning. In: Companion Proceedings of the ACM Web Conference 2024, WWW 2024, Singapore, pp. 1595–1603. Association for Computing Machinery (2024). ISBN: 9798400701726. https://doi.org/10.1145/3589335.3651942
12. Jin, Q., et al.: Event-based video retrieval using audio. In: Interspeech (2012). https://api.semanticscholar.org/CorpusID:18006304
13. Johnson, J., Douze, M., Jégou, H.: Billion-scale similarity search with GPUs. IEEE Trans. Big Data **7**(3), 535–547 (2021). https://doi.org/10.1109/TBDATA.2019.2921572
14. Jung, Y.K., Lee, K.W., Ho, Y.S.: Content-based event retrieval using semantic scene interpretation for automated traffic surveillance. IEEE Trans. Intell. Transp. Syst. **2**(3), 151–163 (2001). https://doi.org/10.1109/6979.954548
15. Li, C., et al.: Large Language Models Understand and Can be Enhanced by Emotional Stimuli (2023). https://api.semanticscholar.org/CorpusID:260126019
16. Mazloom, M., Li, X., Snoek, C.G.: Few-example video event retrieval using tag propagation. In: Proceedings of International Conference on Multimedia Retrieval, ICMR 2014, Glasgow, United Kingdom, pp. 459–462. Association for Computing Machinery (2014). https://doi.org/10.1145/2578726.2578793
17. Nguyen, T. P., et al.: Traffic video event retrieval via text query using vehicle appearance and motion attributes. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 4160–4167 (2021). https://doi.org/10.1109/CVPRW53098.2021.00470.
18. Le-Quynh, M. D., et al.: Enhancing video retrieval with robust CLIP-based multimodal system. In: Proceedings of the 12th International Symposium on Information and Communication Technology, SOICT 2023, Ho Chi Minh, Vietnam, pp. 972–979. Association for Computing Machinery (2023). ISBN: 9798400708916. https://doi.org/10.1145/3628797.3629011
19. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning, vol. 139, pp. 8748– 8763. Proceedings of Machine Learning Research, PMLR (2021). https://proceedings.mlr.press/v139/radford21a.html
20. Sauter, L., et al.: Exploring effective interactive text-based video search in vitrivr. In: Dang-Nguyen, D.T., et al. (eds.) MultiMedia Modeling, pp. 646–651. Springer, Cham (2023). ISBN: 978-3-031-27077-2

21. Schoeffmann, K.: Video browser showdown 2012-2019: a review. In: 2019 International Conference on Content-Based Multimedia Indexing (CBMI), pp. 1–4 (2019). https://doi.org/10.1109/CBMI.2019.8877397
22. Singh, B., et al.: Selecting relevant web trained concepts for automated event retrieval. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 4561–4569 (2015). https://doi.org/10.1109/ICCV.2015.518
23. Smeulders, A., et al.: Content-based image retrieval at the end of the early years. IEEE Trans. Pattern Anal. Mach. Intell. **22**(12), 1349–1380 (2000). https://doi.org/10.1109/34.895972
24. Souek, T., Loko, J.: Transnet V2: an effective deep network architecture for fast shot transition detection (2020). https://api.semanticscholar.org/CorpusID:221095572
25. Tamm, Y.M., Damdinov, R., Vasilev, A.: Quality metrics in recommender systems: do we calculate metrics consistently? In: Proceedings of the 15th ACM Conference on Recommender Systems, Rec-Sys 2021, Amsterdam, Netherlands, pp. 708–713. Association for Computing Machinery (2021). ISBN: 9781450384582. https://doi.org/10.1145/3460231.3478848
26. Tian, K., et al.: Towards efficient and effective text-to-video retrieval with coarse-to-fine visual representation learning. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI) (2024)
27. Wali, A., et al.: A new system for event detection from video surveillance sequences. In: Blanc-Talon, J., et al. (eds.) Advanced Concepts for Intelligent Vision Systems, pp. 110–120. Springer, Heidelberg (2010). ISBN: 978-3-642-17691-3
28. Wang, Z., et al.: Speculative RAG: enhancing retrieval augmented generation through drafting (2024). https://api.semanticscholar.org/CorpusID:271097348
29. Wei, J., et al.: Chain-of-thought prompting elicits reasoning in large language models. Adv. Neural. Inf. Process. Syst. **35**, 24824–24837 (2022)
30. Wu, J., et al.: LLM-based query paraphrasing for video search (2024). https://api.semanticscholar.org/CorpusID:271245154
31. Yang, Z., et al.: Video event restoration based on keyframes for video anomaly detection. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 14592–14601 (2023). https://doi.org/10.1109/CVPR52729.2023.01402
32. Zhang, Y., et al.: High-level representation sketch for video event retrieval. Sci. China Inf. Sci. **59** (2016). https://api.semanticscholar.org/CorpusID:18659831
33. Zhu, H., et al.: Enhancing interactive image retrieval with query rewriting using large language models and vision language models. In: Proceedings of the 2024 International Conference on Multimedia Retrieval, ICMR 2024. Phuket, Thailand, pp. 978–987. Association for Computing Machinery (2024). ISBN: 9798400706196. https://doi.org/10.1145/3652583.3658032