This CVPR Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Toward Automation in Text-based Video Retrieval with LLM Assistance

Khanh-An C. Quan*, Qui Ngoc Nguyen, Duc-Tuan Luu
University of Information Technology, VNU-HCM, Ho Chi Minh City, Vietnam
Vietnam National University, Ho Chi Minh City, Vietnam
{anqck,quinn,tuanld}@uit.edu.vn

Abstract

Video retrieval is a challenging task that requires capturing both semantic relevance and temporal consistency to return accurate results. Recent advancements in Vision-Language Models (VLMs) have significantly improved retrieval performance, but challenges remain in handling complex queries and ensuring robust ranking. To address these issues, we explore the integration of Large Language Models (LLMs) into multiple modules within the retrieval process to enhance video retrieval using textual queries. Specifically, we propose four key modules: Temporal-assisted retrieval, Query refinement, Results reranking, and Multimodal combinations. These modules leverage LLMs to improve temporal understanding, refine input queries, rerank retrieval results, and integrate multimodal cues, ultimately enhancing the relevance and accuracy of retrieved video segments. Our study focuses on a fully automated retrieval system, where queries are processed without human intervention. Comprehensive experiments on the textual KIS dataset from the Video Browser Showdown (VBS) competition demonstrate that LLM-assisted retrieval significantly improves retrieval performance. The proposed framework outperforms conventional approaches by effectively handling complex search queries and improving retrieval accuracy, highlighting the potential of LLMs in automated video search.

1. Introduction

Video retrieval [32] is a critical and challenging task in multimodal information processing, particularly as the volume of video content on online platforms continues to grow exponentially. Users may search for specific video segments for various purposes, including research, education, surveillance, or entertainment. Unlike traditional text retrieval, video retrieval requires systems not only to understand textual queries accurately but also to align them with relevant visual features. A common scenario is *Known-Item Search (KIS)*, in which users attempt to retrieve a video or segment they have previously seen but can no longer recall in detail. These queries are often vague, involving partial memories of scenes, characters, or general context, and cannot be expressed through precise keywords. This creates a significant challenge, as the system must interpret the user's intent and identify semantically relevant content within a large-scale video repository. Among the various approaches to video retrieval, keyframe-based methods with textual query input have demonstrated superior performance in both retrieval speed and result accuracy [24]. However, natural language queries often pose challenges due to their complexity and narrative structure. Such queries typically contain multiple elements—such as characters, actions, and scenes—described in a continuous flow, making it difficult for systems to extract and match relevant details. In addition, many queries imply a temporal sequence of events, which traditional systems often fail to model effectively.

Recent advances in artificial intelligence, particularly with the rise of Large Language Models (LLMs) [4], have led to a paradigm shift in the representation of vision-language cross-domain tasks, enabling unprecedented levels of automation and cognitive assistance across a wide range of applications [14, 30]. The inherent ability of LLMs to process and generate natural language, combined with their advanced reasoning and inferential capabilities, has paved the way for the development of sophisticated models powered by LLM [5, 8, 11]. These models are increasingly being deployed across various domains, including complex data analysis, knowledge representation, and intuitive information retrieval, to optimize and streamline cognitive and operational workflows. Among various challenges [6, 24], numerous advanced video search frameworks have been introduced incorporating LLMs [19, 23].

To address the limitations of video retrieval methods using textual input, we propose an LLM-assisted retrieval framework designed to improve video retrieval performance. By incorporating LLMs into the retrieval stages, our approach improves accuracy, temporal understanding,

*Corresponding author.

and ranking, leading to more effective and semantically coherent video retrieval. Experimental results on the V3C1 dataset [18] demonstrate that our method outperforms existing approaches, underscoring the impact of LLM-assisted retrieval in video search tasks. The proposed retrieval process is fully automated, with queries provided in advance to participants, and does not require direct human involvement to resolve. Our contributions can be outlined in two primary points, detailed as follows:

- We propose an enhanced retrieval framework that integrates Large Language Models (LLMs) for query refinement, temporal search, results reranking, and multi-model fusion. Extensive ablation studies show that LLM-assisted modules significantly improve retrieval accuracy. Additionally, we extend our framework with a question-answering module that extracts contextualized questions from queries and generates responses based on retrieved video content.
- Second, we conduct an in-depth analysis of state-of-the-art Vision-Language Models (CLIP, SigLIP, EVA-CLIP) and different keyframe extraction methods for video retrieval. Our results show that SigLIP and EVA-CLIP outperform CLIP across all retrieval metrics on the V3C1 dataset, while keyframes extracted using FFmpeg [3] yield superior performance compared to the default keyframes extracted by Cineast [17] in the V3C1 dataset.

2. Related Works

2.1. Video Retrieval

The ubiquity of video data in daily activities, such as news reports, cinematic entertainment, or automated surveillance, has led to a huge demand for sophisticated video retrieval systems. These systems must exhibit scalability and precision to efficiently manage and extract pertinent information from the vast amount of visual data collection. Specifically, in the context of video retrieval, users may perform the search process to locate predefined or known items using their own natural language description. The attempt to retrieve a specific, familiar contextual visual content is the primary task of video segment retrieval, namely the *Known-Item Search* or *KIS* task. Additionally, the *Adhoc Video Search* or *AVS* is also an important task where the goal is to find as many keyframes as possible that are semantically relevant to the textual query input. Multiple annual competitions, e.g. Lifelog Search Challenge [6], Video Browser Showdown [24], or NTCIR-Lifelog [31] are organized to promote advanced search frameworks with real-time benchmarks and achieve competitive results. Previous approaches [2, 12, 21] demonstrated that integrating CLIP-based semantic knowledge [7, 16, 29] into the search process can improve the accuracy and relevance of the retrieval results. Therefore, in recent years, various CLIP variants have emerged to enhance video retrieval tasks. In 2023, Zhai et al. introduced SigCLIP [29], which replaces the traditional loss function with a sigmoid-based loss, decoupling batch size from the loss function. SigLIP improves performance even with limited computational resources while maintaining high accuracy. Building on these advancements, Sun et al. proposed EVA-CLIP [22] to optimize training efficiency without significantly increasing computational costs. The model leverages pre-trained EVA, the LAMB optimizer, and flash attention, ensuring stable training and improved performance on zero-shot tasks.

2.2. Integration of LLM in Multimodal Models

Advancements in LLMs have led to the creation of AI models that can integrate multimodal inputs for richer understanding and reasonable decisions for human-computer interaction. Tools such as Visual ChatGPT [26] bridge the gap between text and visuals, enabling image generation and editing during chats. MM-REACT [27] takes it a step further, integrating vision experts with ChatGPT for multimodal understanding and action. HuggingGPT [20] leverages various AI models from HuggingFace, tackling complex tasks across different modalities and domains. TaskMatrix.AI [8] expands the scope even further, connecting foundation models with millions of APIs for diverse digital and physical tasks. InternGPT [10] emphasizes clear communication by combining pointing actions with natural language instructions. VisProg [5] and the followed application [11] introduced the concept of visual programming in which modular pseudocodes are generated based on user instructions. The corresponding programs are then executed to achieve the target requirement. With the in-context learning attribute of LLMs, the models learn specific patterns from a limited set of task examples and uses this understanding to interpret new instances. Our work builds upon these approaches by integrating LLMs into multiple stages of the retrieval process, including query refinement, temporal-assisted search, and reranking, demonstrating significant improvements in retrieval accuracy.

3. Approaches

3.1. Base Retrieval Model

The base video retrieval pipeline is illustrated in Figure 1. Our approach follows the standard keyframe-based video retrieval framework used in previous systems [24]. The pipeline consists of two main stages: Indexing Stage and Searching Stage.

3.1.1. Indexing Stage

In the indexing stage, keyframes are extracted from a large-scale video dataset to capture the core content of each video. These keyframes act as visual anchors for retrieval, ensuring that only the most relevant frames are stored. Once ex-

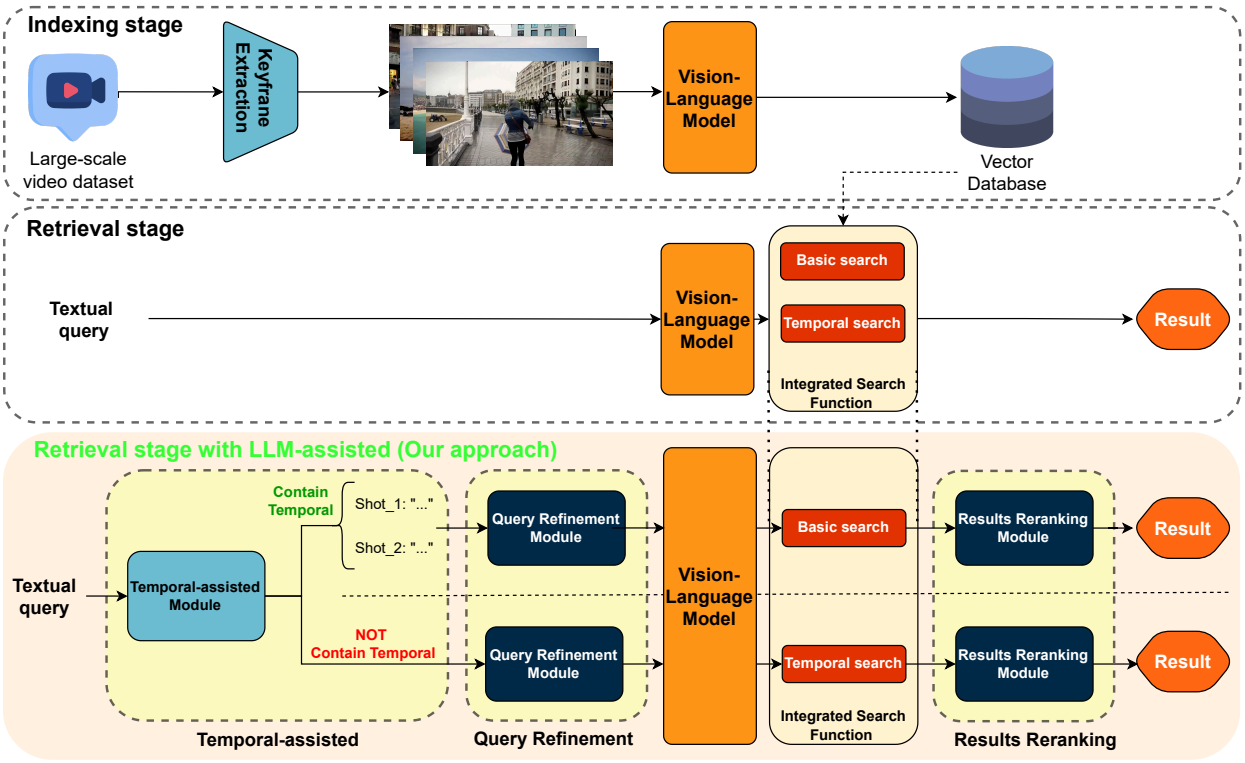


Figure 1. The overall pipeline of the video retrieval system using textual description input. The system consists of two main stages: the Base Retrieval Model and the proposed LLM-assisted Retrieval Stage, which enhances retrieval accuracy through query refinement, temporal-assisted search, results reranking, and multimodal fusion.

tracted, the keyframes are processed by a Vision-Language Model (VLM), which converts each image into a high-dimensional feature vector. These feature representations are then indexed in a vector database, enabling efficient similarity-based retrieval. The quality of keyframe extraction and the choice of VLM significantly impact the accuracy of the entire retrieval system. Selecting the right keyframe selection strategy and model ensures that the indexed representations accurately capture the semantics of the video content, leading to more precise search results.

Keyframes Extraction We tested two types of keyframes extracted from videos: 1. keyframes extracted by Cineast [17] provided in the dataset and 2. keyframes extracted by FFmpeg [3]. For keyframes extracted by FFmpeg, we defined the segment's start time as the time of the extracted keyframe and the end time of the segment as the time of the next extracted keyframe.

Joint Embedding Models Many prior works have utilized CLIP as a joint embedding model for textual and visual data, achieving remarkable results in video retrieval competitions, such as the Video Browser Showdown (VBS). However, recent advancements have introduced enhanced variations of CLIP that offer superior performance, particularly in Text-to-Image retrieval tasks. In our approach, we integrate two state-of-the-art CLIP variations, SigLIP [28] and EVA-CLIP [22], both of which demonstrate stronger performance in zero-shot image-text retrieval, significantly surpassing the original CLIP, especially on datasets like MS-COCO [9]. By leveraging these advanced joint embedding models, our system enhances retrieval accuracy and robustness, ensuring improved alignment between textual queries and visual content.

3.1.2. Retrieval Stage

In the retrieval stage, the input textual query is processed using the same Vision-Language Model (VLM) as in the indexing stage to generate its feature representation. This ensures consistency between the textual and visual embeddings, allowing for accurate similarity-based retrieval. We employ two types of search strategies in our retrieval system: Basic Search and Temporal Search. By integrating these search strategies, our system effectively retrieves keyframes that align with both static and temporally connected queries, enhancing the overall retrieval performance.

Basic Search This type of search takes a textual query

as input and returns a list of relevant video segments that match the query. The retrieval process works by comparing the similarity between the textual query representation—extracted using a Vision-Language Model (VLM)—and the keyframe representations stored in the vector database during the indexing stage. Pairs of vectors with the highest similarity scores are considered the most relevant results, enabling the retrieval of video frames that best align with the linguistic description. This approach enhances both efficiency and accuracy in video retrieval by leveraging the power of VLMs to understand and bridge the gap between visual and textual content.

Temporal Search While input queries often contain temporal information, our video retrieval system also incorporates temporal search to enhance retrieval accuracy. Specifically, our temporal search feature takes as input two chronological descriptions corresponding to two scenes appearing in the video. For this type of search, we first retrieve keyframes using the first input query, generating a list of the most relevant keyframes along with their corresponding similarity scores. Then, for each keyframe in this list, we compute the similarity scores of the second input query with all its successive keyframes in the next segments, selecting the one with the highest similarity score as the most relevant successive keyframe. To further refine the ranking, we combine the similarity score from the initial retrieval with the highest similarity score from the successive keyframe search. This ensures that the final ranking reflects both the relevance of the initial keyframe to the first query and the temporal continuity between keyframes with respect to the second query. As a result, the retrieved keyframe pairs maintain a coherent temporal alignment, making them highly effective for video retrieval tasks that require both semantic relevance and temporal consistency.

3.2. Retrieval Stage LLM-Assisted

In this approach, we integrate Large Language Models (LLMs) into multiple modules within the base retrieval process to enhance video retrieval using textual queries. Specifically, we propose four key modules: Temporal-assisted retrieval, query refinement, results reranking, and multimodel combinations - all designed to improve retrieval performance with LLM support. The LLM-assisted retrieval stage is illustrated in Figure 1-bottom. The retrieval process begins when a user submits a textual query. From there, the system leverages LLMs across the proposed modules to enhance temporal understanding, refine the query, re-rank results, and incorporate multi-modal cues, ultimately improving the relevance and accuracy of retrieved video segments.

For the Temporal-Assisted and Query Refinement modules, we use ChatGPT-4o [13], while InternVL 2.5 [1] is adopted for the Results Re-ranking module. Detailed prompts are provided in the Supplementary Material.

3.2.1. Temporal-Assisted Module

While textual queries often describe multiple scenes within a video, a basic search approach that considers only a single scene from the query may introduce noise into the retrieval results. To address this, the temporal-assisted module is designed to first determine whether the given query contains temporal information. By identifying queries that describe sequential events or multiple scenes, this module enables more accurate retrieval by incorporating temporal context into the search process. Based on this classification, the retrieval system can decide whether to perform a basic search (using a single input description) or a temporal search (handling multiple input descriptions). If the query contains temporal information, this module further helps decompose it into multi-shot descriptions, ensuring that each segment is processed separately for more accurate video retrieval.

3.2.2. Query Refinement Module

The query refinement module enhances video retrieval by optimizing textual queries for better search relevance. It restructures the input description into a concise, keyword-rich, and highly searchable format, ensuring alignment with common search patterns. By removing unnecessary details while preserving core meaning, this module improves retrieval accuracy and ranking, making video searches more efficient and effective.

3.2.3. Results Reranking Module

The results re-ranking module leverages vision LLMs to refine retrieval accuracy by reassessing the similarity between the textual query and the retrieved images. After the initial search retrieves candidate keyframes, this module re-evaluates their relevance using a Vision LLMs-powered similarity assessment. This ensures that the top-ranked results are the most semantically and visually aligned with the query. This process helps filter out less relevant frames and improves the overall quality of video retrieval.

Specifically, given an input textual query and its retrieved keyframes, a Vision LLM evaluates their relevance and assigns an "LLM score" as a refined similarity measure. The retrieved results are then re-ranked based on this LLM score. If multiple results share the same LLM score, they are further ranked using the similarity score between the textual query and keyframe representations extracted by the Vision LLM. For temporal searches, the module computes LLM similarity scores between retrieved keyframe pairs and the multi-shot descriptions generated by the query refinement module.

3.2.4. Multi-model Combinations

The multi-model combinations module enhances retrieval diversity by integrating results from multiple Vision-Language Models (VLMs). Instead of relying on a single

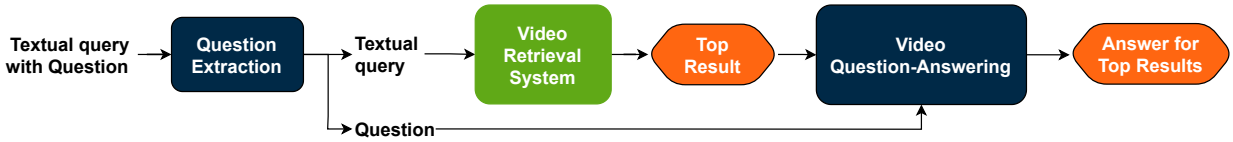


Figure 2. The video question-answering pipeline based on the proposed LLM-assisted video retrieval system. From the input containing both a textual description and a question, we introduce a Question Extraction Module to separate them. The extracted description is used to retrieve the top relevant videos, and a Video Question-Answering Module answers the question based on the top relevant result.

model, this approach combines the top retrieval results from different VLMs, ensuring broader coverage and improved retrieval robustness.

Specifically, the most relevant keyframes retrieved by different models are concatenated into a unified result set. To ensure optimal ranking, the Results Reranking Module is then applied to evaluate the relevance of these retrieved frames against the textual query. This process helps leverage the strengths of multiple models, leading to more accurate and diverse video retrieval outcomes.

3.3. Question-Answering with LLM-Assisted

The Question-Answering with LLM-Assisted pipeline is illustrated in Figure 2. For the question-answering task, we utilize the same Retrieval Stage LLM-Assisted pipeline described in Section 3.2. However, unlike standard retrieval, the retrieved keyframes are grouped by video, meaning each video is represented only by the keyframe with the highest relevance score. To generate answers based on video content, we introduce two additional modules: Question Extraction and Video Question-Answering.

3.3.1. Question Extraction

In textual queries, the question is always located in the last sentence. However, relying solely on the question itself may result in a loss of contextual information, as previous sentences often describe the background or setting of the query. This module is designed to extract the question while preserving its relevant contextual information, ensuring that the response remains accurate and meaningful.

3.3.2. Video Question-Answering

To generate answers based on video content, we employ the InternVideo 2.5 model [25], which has demonstrated state-of-the-art performance across various short and long video benchmarks. This model takes the retrieved video as input along with the extracted question from the previous step and generates an answer based on the visual content. By leveraging advanced video-language modeling, this approach enhances the accuracy and relevance of question-answering in video retrieval.

4. Experiment

4.1. Datasets and Evaluation Metrics

Validation Dataset For validation, we use the textual KIS dataset from the Video Browser Showdown (VBS) competition from 2019 to 2024[1]. To ensure relevance, we filter out all queries where the answer is contained within the V3C1 dataset [18], resulting in a final validation set of 46 items. Each item in the dataset consists of a textual description that uniquely specifies a target time segment in a video. For our retrieval pipeline, we use the last hint provided in the dataset as the input textual query, as it contains the most detailed description of the target scene. For an answer to be considered accurate, it must: Correctly identify the video containing the target segment and ensure that the retrieved interval fully overlaps with the ground-truth segment.

Evaluation Metrics We employed Hit@K as the evaluation metric, where a score of "1" is assigned if at least one correct item appears within the top K retrieved entries; otherwise, a score of "0" is recorded. To determine the correctness of an answer, the target video must be correctly identified and ensured that the predicted interval is entirely contained within the ground-truth segment. Intervals that are fully enclosed within the target, even if shorter in duration, are considered correct. However, segments that only partially overlap with the target are regarded as incorrect. This strict evaluation criterion ensures precise temporal localization in video retrieval.

4.2. Experiment setup

Vision-Language Models & Similarity Search For the Vision-Language Models (VLMs), we evaluate CLIP [15], SigLIP [28], and EVA-CLIP [22] using their respective pretrained weights: clip_vit_large_patch14_336, SigLIP-large-384, and EVA-CLIP-large-336. The retrieval system employs Milvus as the vector database, with cosine similarity used as the similarity metric to measure the relevance between textual and visual representations.

LLM-Assisted Modules We integrate ChatGPT-4o [13] as the Large Language Model (LLM) in the Temporal-Assisted Retrieval, Query Refinement, and Question Extraction modules. To ensure deterministic outputs, we con-

[1] https://github.com/lucaro/VBS-Archive

Table 1. Retrieval results with different vision language models and keyframe extractions on the V3C1 dataset.

Models	Type of Keyframes	#Keyframes	Validation		
			H@1	H@5	H@10
CLIP [15]	Default keyframes (Cineast)	1,082,659	0.1087	0.2391	0.2826
SigLIP [28]			**0.1739**	**0.3478**	**0.3913**
EVA-CLIP [22]			0.1739	0.3043	0.3695
SigLIP [28]	Custom keyframes (FFmpeg)	1,552,550	0.2391	0.5217	0.6304
EVA-CLIP [22]			**0.4130**	**0.6304**	**0.6957**

figure it with a fixed seed of 0, a temperature of 0, and top_p set to $1e-21$. Despite these settings, we observed slight variations in the generated outputs across different calls. To account for this variability, we report the average performance over five runs in Table 2, including the standard deviation. For the Results Re-ranking and Multi-Model Combinations modules, we use InternVL2_5-8B [1] with 4-bit quantization for evaluating retrieval relevance. For the Video Question-Answering module, we use InternVideo2.5 [25]. The InternVL2_5-8B and InternVideo2.5 models are set with do_sample = false to maintain deterministic outputs. All experiments run on a single NVIDIA RTX 4090 GPU with 24GB of VRAM, ensuring efficient processing and evaluation of retrieval quality. In Multi-Model Combinations, we merge the top-10 retrieval results from each SigLIP and EVA-CLIP model.

4.3. Results

4.3.1. Keyframes Extraction and Features Extraction

The retrieval results for different feature extraction models and keyframe extraction methods are presented in Table 1. When using the default keyframes provided in the V3C dataset, both SigLIP and EVA-CLIP demonstrated significantly higher performance than CLIP across all evaluation metrics. SigLIP, in particular, achieved the highest H@5 and H@10 scores, outperforming the CLIP baseline by 45.45% and 38.46%, respectively. With keyframes densely extracted using FFmpeg, the retrieval performance of both SigLIP and EVA-CLIP improved substantially compared to the default keyframes extracted by Cineast. Specifically, SigLIP achieved a 61.11% improvement over its results with default keyframes, while EVA-CLIP exhibited an even greater improvement of 88.28%. In this configuration, EVA-CLIP outperformed SigLIP across all retrieval metrics, making it the most effective model in this setup. However, FFmpeg keyframe extraction produced 43.40% more keyframes than Cineast, increasing the number of indexed frames in the retrieval system. Despite this, the superior retrieval performance achieved with FFmpeg-extracted keyframes justified its use as the keyframe extraction method for our retrieval pipeline.

4.3.2. LLM-Assisted Modules

The ablation study of the proposed LLM-assisted modules in our retrieval process is illustrated in Table 2. The application of the Temporal-Assisted module led to a significant performance boost for both SigLIP and EVA-CLIP compared to not using it. Specifically, for SigLIP, the retrieval performance improved by 103.65% in H@1, 22.49% in H@5, and 19.31% in H@10. Similarly, for EVA-CLIP, the Temporal-Assisted module resulted in increases of 32.64% in H@1, 5.52% in H@5, and 2.49% in H@10.

Further improvements were observed when incorporating the Query Refinement module. With EVA-CLIP, applying this module led to performance gains across all retrieval metrics. However, for SigLIP, while H@1 and H@5 improved, a slight decrease was noted in H@10. These results indicate that while the Query Refinement module generally enhances retrieval accuracy, its impact may vary depending on the feature extraction model used.

When applying the reranking module, the EVA-CLIP model shows a notable improvement in retrieval accuracy, particularly in H@1, which increases by 10.07% compared to not using reranking. This demonstrates the effectiveness of Vision LLM-based assessment in refining top-ranked retrieval results. On the other hand, the SigLIP model exhibits an improvement in H@5, but a slight decrease in H@1, suggesting that the reranking module may prioritize broader relevance over exact top-1 matches in this case. Additionally, reranking across multiple retrieved elements has a negligible overall impact, indicating that the module is most effective when refining the highest-ranked results.

4.3.3. Multi-Model Combinations

Table 3 presents the results of the Multi-Model Combinations approach. Specifically, we integrate both SigLIP and EVA-CLIP with LLM-assisted retrieval. Compared to using each model separately, the combined approach demonstrates an improvement in H@5, highlighting the diversity in retrieval results between the two models. This suggests that SigLIP and EVA-CLIP capture complementary aspects of the query, and their combination enhances the overall retrieval performance. The results validate the effectiveness of leveraging multiple VLMs to improve the robustness and accuracy of video retrieval.

4.3.4. Constrain Relaxation

In this experiment, rather than using strict evaluation constraints—where an answer is considered correct only if it correctly identifies the video and the specified interval is entirely contained within the target—we explore a more flexible evaluation approach. We assess our system under three relaxed scenarios to better understand its effectiveness. In the Inside scenario, an answer is considered correct if the predicted interval is fully contained within the target segment. The Overlap scenario allows partial overlap between

Table 2. Ablation study of the proposed LLM-assisted modules in our retrieval process. The best results for each model are highlighted in **bold**, while the overall best results across both models are highlighted in **red**.

Models	Temporal assisted	Query refinement	Results Reranking	Validate		
				H@1	H@5	H@10
SigLIP [28]				0.2391 ± 0.0000	0.5217 ± 0.0000	0.6304 ± 0.0000
	✓			0.4870 ± 0.0194	0.6391 ± 0.0119	0.7522 ± 0.0119
	✓	✓		0.5130 ± 0.0248	0.6435 ± 0.0119	0.7043 ± 0.0119
	✓	✓	✓(Top 10)	0.4913 ± 0.0248	0.6609 ± 0.0119	0.7043 ± 0.0119
	✓	✓	✓(Top 20)	**0.4957 ± 0.0283**	**0.6739 ± 0.0154**	**0.7174 ± 0.0154**
EVA-CLIP [22]				0.413 ± 0.0000	0.6304 ± 0.0000	0.6957 ± 0.0000
	✓			0.5478 ± 0.0182	0.6652 ± 0.0119	0.7130 ± 0.0097
	✓	✓		0.5609 ± 0.0182	**0.7391 ± 0.0000**	**0.7783 ± 0.0097**
	✓	✓	✓(Top 10)	**0.6174 ± 0.0119**	0.7348 ± 0.0097	**0.7783 ± 0.0097**
	✓	✓	✓(Top 20)	0.6000 ± 0.0119	0.7174 ± 0.0000	**0.7783 ± 0.0097**

Table 3. Results of Multi-Model Combinations. The LLM-assisted approach includes Temporal-Assisted retrieval, Query Refinement, and Results Reranking (Top 10 results considered).

Model	Validate		
	H@1	H@5	H@10
#1: SigLIP + LLMs	0.4913 ± 0.0248	0.6609 ± 0.0119	0.7043 ± 0.0119
#2: EVA + LLMs	**0.6174 ± 0.0119**	0.7348 ± 0.0097	**0.7783 ± 0.0097**
Model #1 + Model #2	0.5957 ± 0.0119	**0.7478 ± 0.0119**	0.7522 ± 0.0194

Table 4. Results of constraint relaxation on our Multi-Model Combinations approach, showing the impact of different evaluation criteria on retrieval performance.

Constraints	Validate		
	H@1	H@5	H@10
Within	0.5957 ± 0.0119	0.7478 ± 0.0119	0.7522 ± 0.0194
Overlap	**0.7391 ± 0.0000**	**0.8043 ± 0.0000**	**0.8087 ± 0.0097**
Mean	0.6957 ± 0.0000	**0.8043 ± 0.0000**	**0.8087 ± 0.0097**

the predicted and target segments, considering them correct as long as there is some intersection. In the Mean scenario, instead of using the original start and end times of a segment to define its temporal boundaries, we compute the mean of the start and end times and assign both the new start and end to this average value, resulting in a fixed-duration segment centered within the original interval. These relaxed evaluation criteria provide a broader perspective on the retrieval system's performance, capturing its ability to retrieve semantically relevant content even when the exact time segment prediction is not strictly precise.

Table 4 illustrates the results when simplifying the evaluation constraints. When considering the Overlap condition instead of the strict Within constraint, the retrieval performance improves significantly across all metrics, with a notable increase in H@1 by 24.07%, reaching 73.81% accuracy. This indicates that the proposed retrieval system provides results that are closely aligned with the intended search targets, even if the exact time intervals do not fully match. Similarly, when evaluating results based on the Mean condition—where the start and end times of the retrieved segment are set equal to the averaged value—the performance improves compared to the strict within the condition. The results obtained using the Mean condition are only slightly lower than those using the Overlap condition but still significantly higher than Within constraints.

4.4. Qualitative Results

The qualitative results of the video retrieval system using EVA-CLIP, Temporal-assisted, and Query Refinement modules are illustrated in Figure 3. Compared to using the default input query, the Temporal-assisted module effectively extracts temporal information, breaking down the query into descriptions of two distinct scenes. This allows the retrieval system to conduct a temporal-aware search, leading to higher accuracy in retrieving relevant keyframes.

Furthermore, the Query Refinement module further enhances retrieval performance by optimizing the input query. It refines the textual descriptions to be more search-friendly while eliminating redundant information, ensuring a more precise match with the indexed keyframes. The combination of these modules results in a more effective video retrieval pipeline, demonstrating the benefits of integrating LLM-assisted enhancements into the search process.

5. Conclusion

In this paper, we present Toward Automation in Video Retrieval with LLM-Assist, a novel automated video retrieval framework that leverages Large Language Models (LLMs) to handle complex search queries and enhance retrieval accuracy. Our system integrates LLMs for query refinement, temporal-assisted search, results reranking, and multi-model fusion, significantly improving performance compared to traditional methods. Experimental results demonstrate that incorporating LLM-assisted modules boosts retrieval effectiveness, with the Temporal-Assisted module increasing H@1 by 103.65% for SigLIP and 32.64% for EVA-CLIP. Additionally, applying the reranking module further improved H@1 performance for EVA-CLIP by 10.07%, showcasing the effectiveness of

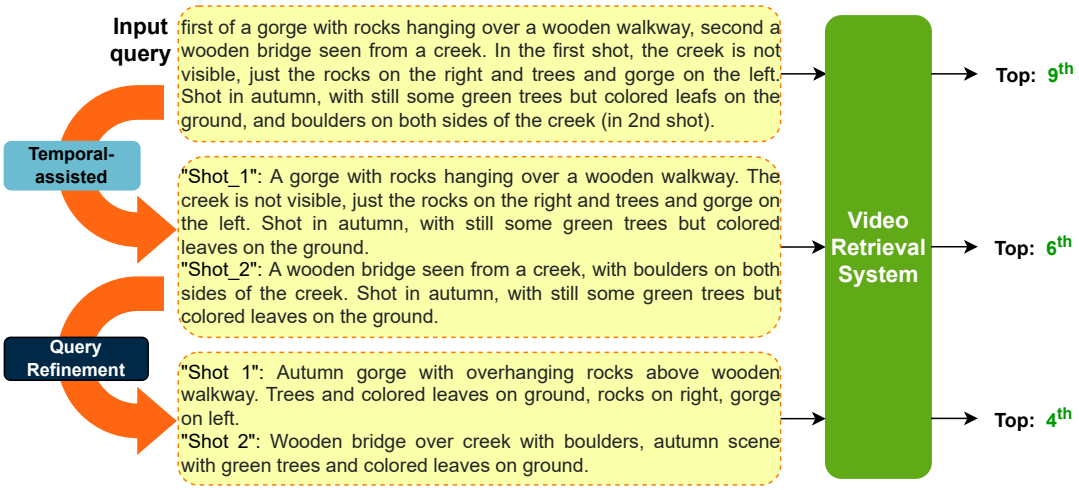


Figure 3. Qualitative results of the video retrieval system with EVA-CLIP, Temporal-assisted, and Query Refinement modules.

LLM-based ranking adjustments.

Our study also highlights the advantages of combining different Vision-Language Models (VLMs), where the fusion of SigLIP and EVA-CLIP improves retrieval diversity and enhances results, particularly at higher recall levels. Furthermore, relaxing evaluation constraints reveals that a substantial number of retrieval results are near the target interval, emphasizing the robustness of our approach in handling ambiguous or complex queries.

Further progress on refining LLM-based query decomposition, adaptive temporal search strategies, and optimizing multimodal fusion techniques could enhance the overall accuracy and efficiency of the system. Future work may also explore real-time retrieval optimization and user feedback-driven refinement to improve search relevance in large-scale video datasets. Expanding the framework to support interactive query reformulation and integrating advanced scene understanding models could further push the boundaries of automated video retrieval.

Acknowledgments.

This research is funded by University of Information Technology - Vietnam National University Ho Chi Minh City under grant number D1-2025-06

References

[1] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 4, 6

[2] Yu-Tong Cheng, Jiaxin Wu, Zhixin Ma, Jiangshan He, Xiao-Yong Wei, and Chong-Wah Ngo. Interactive video search with multi-modal llm video captioning. In *International Conference on Multimedia Modeling*, pages 302–309. Springer, 2025. 2

[3] FFmpeg Developers. ffmpeg tool. `http://ffmpeg.org/`, 2016. [Software]. 2, 3

[4] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 1

[5] Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14953–14962, 2023. 1, 2

[6] Cathal Gurrin, Liting Zhou, Graham Healy, Werner Bailer, Duc-Tien Dang Nguyen, Steve Hodges, Björn Thór Jónsson, Jakub Lokoč, Luca Rossetto, Minh-Triet Tran, et al. Introduction to the seventh annual lifelog search challenge, lsc'24. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pages 1334–1335, 2024. 1, 2

[7] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 2

[8] Yaobo Liang, Chenfei Wu, Ting Song, Wenshan Wu, Yan Xia, Yu Liu, Yang Ou, Shuai Lu, Lei Ji, Shaoguang Mao, et al. Taskmatrix. ai: Completing tasks by connecting foundation models with millions of apis. *Intelligent Computing*, 3:0063, 2024. 1, 2

[9] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In

Computer Vision – ECCV 2014, pages 740–755, Cham, 2014. Springer International Publishing. 3

[10] Zhaoyang Liu, Yinan He, Wenhai Wang, Weiyun Wang, Yi Wang, Shoufa Chen, Qinglong Zhang, Zeqiang Lai, Yang Yang, Qingyun Li, et al. Interngpt: Solving vision-centric tasks by interacting with chatgpt beyond language. *arXiv preprint arXiv:2305.05662*, 2023. 2

[11] Duc-Tuan Luu, Khanh-Linh Bui-Le, Duy-Ngoc Nguyen, Vinh-Tiep Nguyen, and Minh-Triet Tran. Visa: Video interactive search with advanced visual programming. In *2024 International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*, pages 1–6. IEEE, 2024. 1, 2

[12] Duc-Tuan Luu, Khanh-An C Quan, Duy-Ngoc Nguyen, Khanh-Linh Bui-Le, Nhat-Sang Doan, Minh-Duc Le-Ngo, Vinh-Tiep Nguyen, and Minh-Triet Tran. Imsearch 2.0: Toward user-centric and efficient interactive multimedia retrieval system. In *International Conference on Multimedia Modeling*, pages 294–301. Springer, 2025. 2

[13] OpenAI. Chatgpt-4o. https://openai.com/index/chatgpt, 2024. 4, 5

[14] Chenbin Pan, Burhaneddin Yaman, Tommaso Nesti, Abhirup Mallik, Alessandro G Allievi, Senem Velipasalar, and Liu Ren. Vlp: Vision language planning for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14760–14769, 2024. 1

[15] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, pages 8748–8763. PMLR, 2021. 5, 6

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2

[17] Luca Rossetto, Ivan Giangreco, and Heiko Schuldt. Cineast: A multi-feature sketch-based video retrieval engine. In *2014 IEEE International Symposium on Multimedia*, pages 18–23, 2014. 2, 3

[18] Luca Rossetto, Heiko Schuldt, George Awad, and Asad A Butt. V3c–a research video collection. In *MultiMedia Modeling: 25th International Conference, MMM 2019, Thessaloniki, Greece, January 8–11, 2019, Proceedings, Part I 25*, pages 349–360. Springer, 2019. 2, 5

[19] Luca Rossetto, Athina Kyriakou, Svenja Lange, Florian Ruosch, Ruijie Wang, Kathrin Wardatzky, and Abraham Bernstein. Lifegraph 4-lifelog retrieval using multimodal knowledge graphs and vision-language models. In *Proceedings of the 7th Annual ACM Workshop on the Lifelog Search Challenge*, pages 88–92, 2024. 1

[20] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information Processing Systems*, 36:38154–38180, 2023. 2

[21] Michael Stroh, Vojtěch Kloda, Benjamin Verner, Zuzana Vopálková, Raphael Buchmüller, Bastian Jäckl, Jakub Hajko, and Jakub Lokoč. Prak tool v3: Enhancing video item search using localized text and texture queries. In *International Conference on Multimedia Modeling*, pages 326–333. Springer, 2025. 2

[22] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023. 2, 3, 5, 6, 7

[23] Ly Duyen Tran, Thanh-Binh Nguyen, Cathal Gurrin, and Liting Zhou. Myeachtrax: Lifelog question answering on mobile. In *Proceedings of the 7th Annual ACM Workshop on the Lifelog Search Challenge*, pages 93–98, 2024. 1

[24] Lucia Vadicamo, Rahel Arnold, Werner Bailer, Fabio Carrara, Cathal Gurrin, Nico Hezel, Xinghan Li, Jakub Lokoc, Sebastian Lubos, Zhixin Ma, et al. Evaluating performance and trends in interactive video retrieval: Insights from the 12th vbs competition. *IEEE Access*, 2024. 1, 2

[25] Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyu Zeng, Chenting Wang, Changlian Ma, Haian Huang, Jianfei Gao, Min Dou, Kai Chen, Wenhai Wang, Yu Qiao, Yali Wang, and Limin Wang. Internvideo2.5: Empowering video mllms with long and rich context modeling. *arXiv preprint arXiv:2501.12386*, 2025. 5, 6

[26] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023. 2

[27] Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*, 2023. 2

[28] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023. 3, 5, 6, 7

[29] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. 2

[30] Yang Zhong and Bhiman Kumar Baghel. Multimodal understanding of memes with fair explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2007–2017, 2024. 1

[31] Liting Zhou, Cathal Gurrin, Duc-Tien Dang-Nguyen, Graham Healy, Chenyang Lyu, Tianbo Ji, Longyue Wang, Joho Hideo, Ly-Duyen Tran, and Naushad Alam. Overview of the ntcir-17 lifelog-5 task. In *Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies. https://doi. org/10.20736/0002001329*, 2023. 2

[32] Cunjuan Zhu, Qi Jia, Wei Chen, Yanming Guo, and Yu Liu. Deep learning for video-text retrieval: a review. *International Journal of Multimedia Information Retrieval*, 12(1):3, 2023. 1