NewsInsight: A Comprehensive Video Event Retrieval System with Spatial Insights and Query Assistance

Gia-Huy Vuong*
Van-Son Ho*
Tien-Thanh Nguyen-Dang*
Xuan-Dang Thai*
University of Science, VNU-HCM
Vietnam National University
Ho Chi Minh City, Vietnam
vghuy@selab.hcmus.edu.vn
20120021@student.hcmus.edu.vn
19120036@student.hcmus.edu.vn
txdang@selab.hcmus.edu.vn

Van-Tu Ninh
Minh-Khoi Pham
Tu-Khiem Le
Graham Healy
Dublin City University
Ireland
tu.ninhvan@adaptcentre.ie
minhkhoi.pham@adaptcentre.ie
tukhiem.le4@mail.dcu.ie
graham.healy@dcu.ie

Minh-Triet Tran†
University of Science, VNU-HCM
Vietnam National University
Ho Chi Minh City, Vietnam
tmtriet@fit.hcmus.edu.vn

ABSTRACT

Video event retrieval is the task of finding videos that are relevant to a given query. It is a challenging problem because videos are typically much larger than images, and they can contain a variety of different objects and scenes. However, there are a number of different approaches to video retrieval, and the field is rapidly evolving. Some of the most promising research directions include the use of deep learning and multimodal features. In this paper, we introduce NewsInsight – a comprehensive video event retrieval system developed for participating AI Challenge 2023. The system under investigation leverages the Bootstrapping Language-Image Pre-training (BLIP) model for zero-shot image-text retrieval, demonstrating superior recall scores on the Flickr30K dataset compared to the Contrastive Language–Image Pretraining (CLIP) model. In addition, it employs an Elastic Search filtering mechanism to discard irrelevant images. Beyond semantic search mechanisms, the system supports visual similarity search by calculating the inner product distance between vectors in the video frames corpus and the query image. The system also incorporates an explicit relevance feedback function, AI-based query description rewriting, and visual-example-generating features, enhancing the precision of the query description and aiding end-users in formulating a more accurate depiction of the targeted image for retrieval.

*These authors contributed equally to this research.
†Corresponding author

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
SOICT 2023, December 07–08, 2023, Ho Chi Minh, Vietnam
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0891-6/23/12…$15.00
https://doi.org/10.1145/3628797.3628805

CCS CONCEPTS

• **Information systems** → **Multimedia databases**; **Users and interactive retrieval**; *Search interfaces*; • **Human-centered computing** → **Interactive systems and tools**.

KEYWORDS

Lifelog, Interactive Retrieval, Spatial Insights, AI-based Assistance, Video Event retrieval

ACM Reference Format:
Gia-Huy Vuong, Van-Son Ho, Tien-Thanh Nguyen-Dang, Xuan-Dang Thai, Van-Tu Ninh, Minh-Khoi Pham, Tu-Khiem Le, Graham Healy, and Minh-Triet Tran. 2023. NewsInsight: A Comprehensive Video Event Retrieval System with Spatial Insights and Query Assistance. In *The 12th International Symposium on Information and Communication Technology (SOICT 2023), December 07–08, 2023, Ho Chi Minh, Vietnam.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3628797.3628805

1 INTRODUCTION

The increasing ubiquity of video data in our daily lives, ranging from traditional entertainment and broadcasting to camera systems used in manufacturing, medicine, smart cities, transportation, and wearable devices, has created the need for efficient and effective systems for storing, organizing, and managing this data. These systems must also allow users to quickly and easily retrieve the videos that they need from large collections.

In this context, we present NewsInsight, a comprehensive video event retrieval system designed for the AI Challenge 2023[3]. Our system stands out due to its utilization of the Bootstrapping Language-Image Pre-training (BLIP) model for zero-shot image-text retrieval, which has demonstrated superior recall scores on the Flickr30K dataset compared to the Contrastive Language–Image Pretraining (CLIP) model. Furthermore, NewsInsightincorporates an Elastic Search filtering mechanism to eliminate irrelevant images, thereby enhancing the precision of search results. It also supports visual similarity search, calculating the inner product distance between vectors in the video frames corpus and the query image.

The AI Challenge 2023 is a competition that focuses on the problem of video event retrieval. The format of the challenge is expected to be similar to international competitions such as the

Lifelog Search Challenge [4] (LSC) and Video Browser Showdown [10] (VBS). In the 2023 edition of the challenge, participants are expected to develop solutions to retrieve specific events from video data. This involves using AI technologies to analyze video event and identify specific events based on given queries. The challenge is not only a competition but also a learning platform. It encourages participants to enhance their entrepreneurial skills while receiving guidance and feedback from industry experts. Participants can apply the skills they're learning in the challenge and use AI services to develop real-world solutions. The challenge involves two types of queries: Textual Known-Item Search (KIS-Textual) and Video Known-Item Search (KIS-Visual).

In KIS-Textual, the judges provide a natural language description of an event, and the teams must find the exact video segment of this event. The description may include multiple ideas and sentences. In the preliminary round, the description content is provided in full for the teams to search for the video according to the description. In the final round, the description content will be gradually provided during the query time.

In Video KIS-Visual, instead of a natural language description, the teams will watch a short video clip (no more than 20 seconds) extracted from an event in the given video data. The teams are not allowed to take pictures or record this video to put into their tool but must find a suitable way to express the search request, such as describing the video event in natural language or redrawing the scene they see. This type of query is only applied in the final round of the competition.

However, thanks to the rapid development of vision-language models (e.g. Contrastive Language-Image Pre-training – CLIP [15], VisualBERT [9]), we can now conduct a semantic search for a specific moment based on the abstract description of the moment with high precision without heavily relying on the query parsing process for exact-matching with the metadata using the DBMS.

In this study, we introduce a novel video event retrieval system, referred to as NewsInsight. This system inherits the semantic search mechanisms of several leading lifelog retrieval systems from LSC'22. However, the unique aspect of NewsInsight is its emphasis on utilizing spatial information. This spatial information can be derived from either the semantic search results or the entire lifelog data. The integration of spatial information provides valuable insights, enhancing the effectiveness of the video event retrieval process.

A key distinction between the semantic search mechanism of our system and those considered state-of-the-art in LSC'22 lies in our employment of the Bootstrapping Language-Image Pre-training (BLIP) model [8]. This model is utilized for zero-shot image-text retrieval and has been demonstrated to yield a higher recall score compared to the CLIP model when tested on the Flickr30K dataset [14].

Our system, NewsInsight, incorporates several advanced features to enhance the accuracy and relevance of video event retrieval. We employ filtering mechanisms using Elastic Search, utilizing both string query [1] and term query [2] methods to effectively remove irrelevant images from the search results. In addition, we have integrated a visual similarity search function into the NewsInsight system. This feature works by comparing the inner product distance between vectors in the video frames corpus and the vector of the query image. The combination of these features ensures a highly accurate and efficient retrieval process, significantly improving the user experience in video event retrieval.

Beyond the conventional functionalities of a video event retrieval system, our system, NewsInsight, introduces several supplementary features designed to enhance the user experience during the retrieval process. These include an AI-powered query description rewriting mechanism and a feature that generates visual examples based on the query description. These additional features are particularly beneficial for Query-by-Example or Visual Similarity Search functions, offering users a more streamlined and efficient retrieval process. Our system, NewsInsight, has demonstrated its efficacy in retrieving pertinent images corresponding to a given query. This has been substantiated by its ranking within the top five at the HCM AI Challenge 2023, thereby attesting to its robust performance.

The structure of this paper is organized as follows: Section 2 provides a brief review of existing retrieval systems for video event search. The overview and user interface of our system are presented in Section 3. Essential components and solutions of our system are discussed in detail in Section 4. We also demonstrate some experiments of our system in Section 5. The paper concludes with Section 6, where we summarize our findings and discuss future work.

2 RELATED WORK

The domain of video retrieval has witnessed substantial advancements in recent years, prompting the organization of related competitions aimed at developing interactive retrieval systems. These systems have been expected to be efficient in finding target images from a vast array of video event frames within a specified, time-limited query. Two of the most prominent challenges in this field are the Lifelog Search Challenge [4] and Video Browser Showdown [10]. These competitions have garnered significant interest, attracting numerous participants tasked with tackling the video retrieval challenge. As a result, a diverse range of techniques have been employed to address this problem, including to concept-based search, vision-language joint embedding, and semantic-driven systems.

Concept-based search represents a methodology that entails the annotation of images with visual markers or concepts to augment the retrieval process. These visual concepts typically encompass detected objects within the image, image captions, and text extracted from images utilizing Optical Character Recognition (OCR) models. These annotations are subsequently indexed in conjunction with other metadata and utilized to align the images with input keywords. The concept-based search approach can also integrate Natural Language Processing (NLP) to enhance the retrieval experience. Through NLP, the system is capable of analyzing the text associated with the visual content, including captions, tags, and annotations, and extracting pertinent concepts and keywords. The extracted concepts and keywords can then be employed to match the visual content with input queries, thereby facilitating a more precise and relevant retrieval experience. Systems such as

[1] https://www.elastic.co/guide/en/elasticsearch/reference/current/query-dsl-query-string-query.html
[2] https://www.elastic.co/guide/en/elasticsearch/reference/current/query-dsl-term-query.html

Memoria [17], LifeSeeker [12], vitrivr [5], FIRST [21], Myscéal [19] and LifeInsight [13] offered multiple search modalities based on concepts.

The advent of vision-language models, exemplified by the development of the CLIP [15] in 2021, has facilitated the concurrent comprehension of visual and textual data. By jointly processing and embedding visual and textual information, vision-language models engender a more holistic and nuanced understanding of multimedia content, thereby paving the way for accurate and contextually aware retrieval. This approach has shown impressive outcomes in visual retrieval tasks.

Traditional visual retrieval systems rely on the visual features of images to match queries. However, this approach can be limited, as it does not take into account the context and meaning of the images. Semantic search can address this limitation by incorporating the meaning of the images into the retrieval process. This can be done by using machine learning models to extract the semantic features of images. The use of semantic search in visual retrieval systems can significantly improve the retrieval experience. This is because semantic search can capture the context and meaning of the images, as well as the abstract description queries. This allows for more precise and relevant retrieval results. In addition, the use of semantic search can turn traditional visual retrieval systems into semantic-driven systems. This means that the retrieval process is driven by the meaning of the images, rather than just their visual features. This can lead to a deeper understanding of the context of the images, which can enrich the retrieval process. Numerous systems have embraced the use of vision-language pre-trained models, with a particular emphasis on the CLIP model [16], such as LifeSeeker 4.0 [11], E-Myscéal [20], Memento 2.0 [1], FIRST 3.0 [6], Voxento [2] and LifeInsight [13]. These systems significantly improved their performance from the previous version in zero-shot image-text retrieval. These systems have exhibited a significant enhancement in their performance from their previous iterations in the realm of zero-shot image-text retrieval. Therefore, we incorporate the state-of-the-art vision-language pre-training model, BLIP [8], to construct semantic-driven video event retrieval systems. These systems use multiple AI-based features to support the retrieval process, in conjunction with the semantic search function described in Section 4.1.

3 OVERVIEW OF NEWSINSIGHT

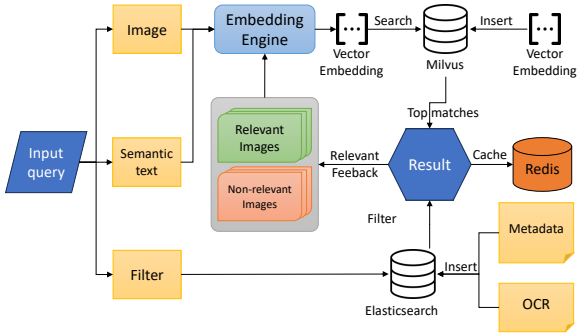
3.1 Overview

We present a system called NewsInsight, which is illustrated in Figure 1. The system can process input queries in the form of visual images, semantic texts or filters (OCR). The query semantic texts or visual images are encoded into vector embedding by the Embedding engine and then searched in the Milvus vector database to retrieve the most relevant images. The preprocessed information from OCR is then applied to filter the results through Elasticsearch. Finally, the system caches the results to gather relevant feedback requests from users.

3.2 User Interface

Our video event retrieval system, NewsInsight, offers a unique and intuitive user interface that distinguishes it from most existing

Figure 1: Overview of NewsInsight system.

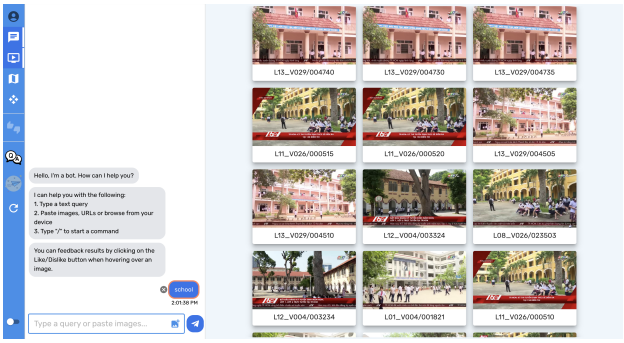


systems. Unlike conventional systems that typically have a search bar and an image result gallery, our system interface includes an image gallery that displays search results in various cluster view modes. Our system also has a versatile chatbox that stores query history and supports both free-text and image input. Users can also use commands for advanced actions. Finally, the vertical navigation panel has convenient buttons that provide easy access to essential features, such as hiding or opening the chatbox, providing feedback, and changing gallery view modes.

In addition, the image gallery is a main component of NewsInsight's user interface. It provides users with a visually appealing and intuitive way to explore and navigate through a vast collection of search results images. The gallery offers moment and video cluster view modes, which enable users to organize and explore images based on temporal or spatial relationships.

By default, the NewsInsight system shows the results of a query in the **moment mode**. The search results are sorted decreasingly based on their relevance to the query, from left to right and top to bottom. This means that the most relevant search results are displayed first. Users can click on individual images to access more detailed information and related images, facilitating a comprehensive understanding of the search results. For example, if a user searches for "school", the results will be shown as in Figure 2.

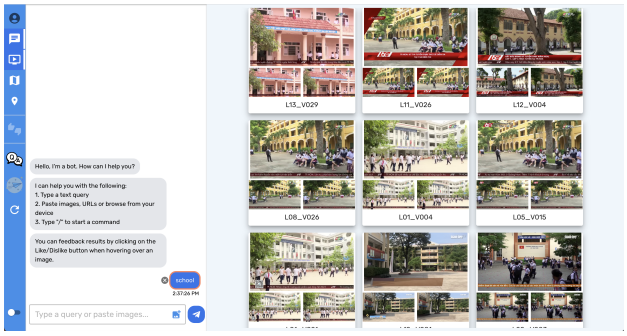
Figure 2: Moment mode.



On the other hand, users can also select the **video cluster mode** to view the results by video ID. This mode groups together images

that are part of the same video, making it easier to find the correct event frames. It also displays the top three images from the ranked list, and users can collapse or expand clusters to customize their browsing experience. For example, if a user searches for "school" and uses the video cluster mode instead of the moment mode, the NewsInsightUI will show as in Figure 3.

Figure 3: Video mode.



The Chatbox feature in NewsInsightis a innovative feature that distinguishes it from existing retrieval systems. It serves as a central hub for user interactions, providing an interactive and dynamic interface that enhances the user experience for visual data retrieval. The query input field in the Chatbox is the primary way for users to interact with the system. Users can enter their search queries in both text and image formats:

- For text queries, the system uses semantic search techniques to match the query with relevant visual content. For example, a user could enter the query "beach" to retrieve images of beaches.
- For image queries, the system retrieves visually similar images based on the provided image input. For example, a user could paste an image of a beach into the query input field to retrieve visually similar images.

This allows users to have more flexibility in how they interact with the system and to retrieve the images they are looking for more easily.

4 ENHANCEMENT FEATURES

4.1 Semantic Search

We use the Bootstrapping Language-Image Pre-training (BLIP) framework, a novel Vision-Language Pre-training (VLP) framework developed by the Salesforce team. BLIP achieves the highest recall score for zero-shot image-text retrieval task on the Flickr30 dataset [8]. The semantic search process using BLIP is similar to that of Contrastive Language-Image Pre-training (CLIP). Both models encode images and query sentences into the same vector space for similarity matching. This is done by transforming the context of the image (the relationship between visual and semantic concepts) and the context of the query sentence (the relationship between the key terms in the query) into vectors with the same dimension. The similarity computation is performed by Milvus [3] – a vector database built for scalable similarity search. The inner product is used to compare the distance between embeddings in the vector space.

4.2 Visual Similarity Search

Visual similarity search is performed in the same way as semantic search, as described in Section 4.1. However, instead of using the embedding of a query sentence, the embedding of the selected images is used. The user can choose to search for visually similar images using either one or multiple images. The final input embedding vector for visual similarity search is the centroid of the selected vectors. For example, suppose we have N images. We can feed these images into a system to encode them into embedding vectors. The average of these embedding vectors is then used as the vector representation of the N images.

$$\vec{q} = \frac{1}{N}\sum_{i=1}^{N}\vec{q}_i$$

where $\vec{q}_i$ is the embedding vector of the i-th selected image, while the vector representation of the total N selected images is denoted by $\vec{q}$.

4.3 Relevance Feedback

We also provide the ability for users to provide explicit **relevance feedback**, as shown in Figure 4. This allows users to re-rank the results or search for a new set of images that are visually similar to the relevant examples provided by the user, while being visually different from the non-relevant examples. To do this, we propose to use the Rocchio algorithm [18] to either initialize the query vector or modify the previous input query vector for searching or re-ranking. The Rocchio algorithm is a relevance feedback algorithm that is used to improve the performance of information retrieval systems. It works by taking the original query vector and moving it closer to the centroid of the relevant documents and farther away from the centroid of the non-relevant documents. The Rocchio algorithm is given by the following formula:

$$\vec{q} = \alpha\vec{q}_0 + \beta\frac{1}{|Q_r|}\sum_{\vec{q}_j\in Q_r}\vec{q}_j - \gamma\frac{1}{|Q_{nr}|}\sum_{\vec{q}_k\in Q_{nr}}\vec{q}_k$$

The embedding vector of the set of relevant examples Q_r is denoted by $\vec{q}_j$, while the embedding vector of the set of non-relevant examples Q_{nr} is denoted by $\vec{q}_k$. The original query vector is denoted by $\vec{q}_0$, and the modified query vector is denoted by $\vec{q}_m$.

4.4 Filtering Mechanism

The NewsInsight system relies heavily on semantic search, which allows users to input a description of a specific moment into the conversation box to retrieve relevant images. However, users can also input a text-based query to filter the ranked list obtained from the semantic search. This is done by employing the query string query mechanism of Elasticsearch, which allows users to perform searches with Boolean operators, wildcards, and fuzzy matching.

[3] https://milvus.io/

Figure 4: Rerank results from relevance feedback.

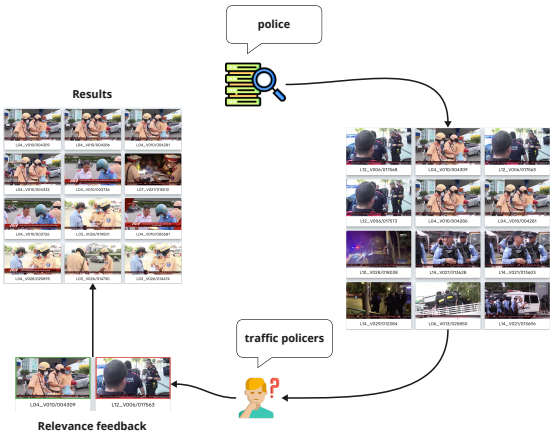
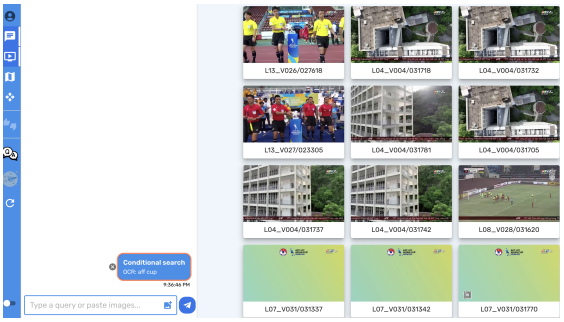


Figure 5: OCR Filtering.



For example, user can use this feature to filter frames which have "aff cup" as Fig 5.

4.5 Query Assistance

Different users may use different words or phrases to describe the same event, and these descriptions may be more or less efficient. To address this, we propose Query Assistance, an AI-based method that draws inspiration from crowd-sourcing principles to rewrite query descriptions. Query Assistance aims to enhance the search experience by providing users with alternative search query suggestions based on their initial input.

When a user provides an initial text description query, our system uses AI to generate a list of alternative query sentences that are relevant to the provided description. The user can then select the most relevant query sentence from the list to perform semantic search, which enables a more precise and efficient retrieval process. This allows users to explore different ways of expressing their search intent and potentially discover more relevant results.

As mentioned above, it can be difficult to describe a specific event or moment in a way that generates accurate results when searching for it. Our system addresses this issue by leveraging AI-based text generation and rewriting tools to provide multiple descriptions of an initial query. In practice, we use the *humarin/chatgpt_paraphraser_on_T5_base* model [4] [22], which is a generative pre-trained transformer model that has been fine-tuned for the task of text paraphrasing.

This approach is inspired by crowd-sourcing and allows users to select the best description for their needs. Figure 6 illustrates this process. Additionally, our system offers an AI-assisted crowd-sourcing mode that runs multiple retrieval processes with different description variants and re-ranks the list of candidate moments based on a voting scheme. In this scheme, the initial text is assigned a weight of 2, while any other descriptions are assigned a weight of 1. This approach ensures that users can find the most relevant and accurate results for their needs.

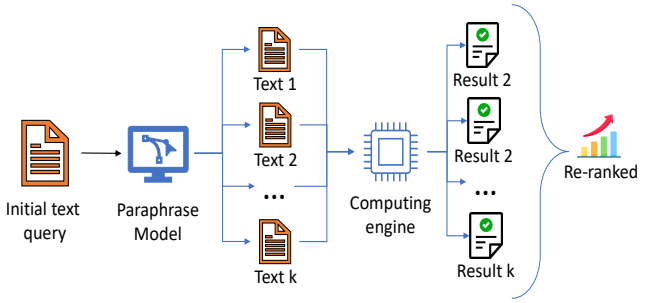


Figure 6: Re-rank results from multiple text-query variants.

4.6 Visual Examples Generating

In order to assist users in visualizing the images they want to seek, we introduce the **visual example generation** feature. Utilizing a state-of-the-art text-to-image model, the system analyzes user queries and generates images that closely align with the user's search criteria. This feature bridges the gap between a user's mental image and their search query, aiding in the visualization of targeted images. Our system's visual search capability allows users to select an image or a collection of images that catch their interest, and then request the system to find similar images. For instance, if a user is searching for "police", they can utilize this feature to generate an image associated with "police", as depicted in Fig. 7.

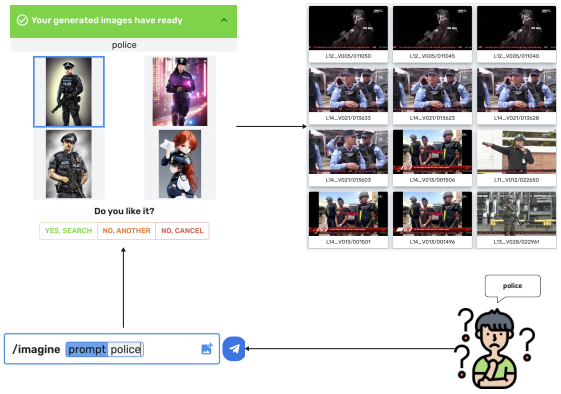
5 EXPERIMENTS AND RESULTS

5.1 Dataset

The dataset used in this paper was obtained from the AI Challenge 2023 competition. It primarily focuses on events and occurrences related to daily life in Vietnam, as covered by a comprehensive news program. This ensures a rich set of events with many different contextual scenarios. The dataset comprises 740 video clips with a total duration of 220 hours, while each video has a consistent frame rate of 25 frames per second (fps). The videos were meticulously gathered from the YouTube platform and published between December 2022 and August 2023.

[4] https://huggingface.co/humarin/chatgpt_paraphraser_on_T5_base

Figure 7: Visual Example Generating.



5.2 Data preprocessing

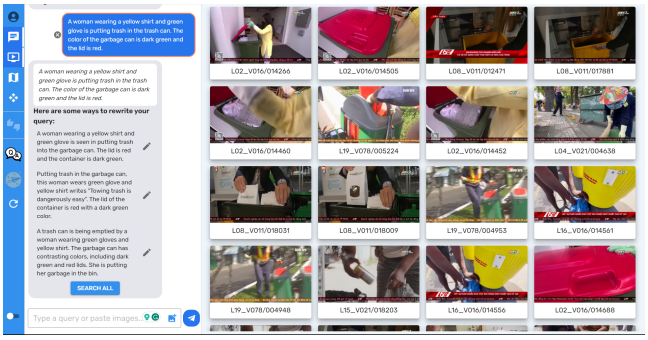
Given that the AI Challenge dataset contains 740 video clips, we need to extract frames from them and insert them into Milvus. However, extracting all frames from a video would be inefficient, so we used the PySceneDetect tool to obtain a scene list. For each scene in the scene list, we also extract K frames (we use $K = 3$ in this paper). Finally, we feed all frames extracted from all videos into the BLIP-2 model [7] to encode them to vector embeddings, which we then insert into Milvus for search.

5.3 Testing

Query 1: *A woman wearing a yellow shirt and green glove is putting trash in the trash can. The color of the garbage can is dark green and the lid is red.*

We employ a straightforward text search method, utilizing the image description as our search query. By searching this way, we identify the top-1 candidate image, shown in Figure 8, which is the image matches the result.

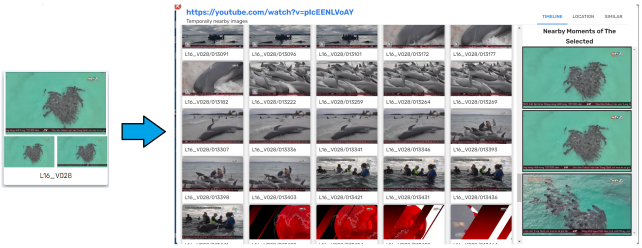
Figure 8: Use text search to find expected image.



Query 2: *A school of fish forms the shape of a heart, and then they swim in a line, tail to tail. The next scene depicts some people assisting these fish back into the sea.*

We determine that the given description contains two main elements: one is the school of fish swimming in the shape of a heart, and the other is the scene of local people assisting the fish in returning to the sea. Therefore, we can search using the query, "A school of fish forms a heart shape. Some people help these fish return to the sea," and employ cluster-by-video techniques, where our system will organize the videos based on the "average relevance of frames within the video," in descending order. This way, we can easily find the video, shown in Figure 9, which contains both scenes we want to search for.

Figure 9: Use cluster-by-video to display clusters of images.



Query 3: *The video is from a press conference of the WHO. The video shows close-up shots of the faces of the following individuals in order: Mr. Abdi Rahman Mahamud, Ms. Maria Van Kerkhove, Mr. Tedros Adhanom Ghebreyesus.*

Upon perceiving this query, we acknowledge the significance of the personal names contained within it, recognizing their potential to be readily searched via a filter employing optical character recognition (OCR) technology. As a result, we promptly prompt the system with the name "Abdi Rahman Mahamud" (shown in Fig 10), and within moments, NewsInsight delivers a set of relevant images, ensuring that the top-10 results are particularly pertinent. Furthermore, in order to instill confidence in the returned frames, NewsInsight provides the user with the ability to swiftly peruse related images, namely those frames that are deemed to be closely associated with the primary image.

Query 4: The video features a woman in a purple traditional dress standing and posing for a photo next to a white cyclo. This cyclo has the license plate number 79H-6072.

Similar to the third query, we can observe the information "79H-6072", which can be readily identified using OCR techniques. Consequently, we incorporate it into our prompt and receive a pertinent image from the top-ranked candidate (described in Fig 11).

Query 4: *The video consists of 4 consecutive images related to the digestive system. The second image lists common digestive diseases. The third image is of a person holding their stomach*

Upon reviewing the query, we opted to employ a visual search strategy. This led to the discovery of an image pertinent to the digestive system. The visual search feature facilitated the identification of the correct answer in the top-ranked result (see Fig 12). As a result, this approach can help users find relevant images more easily.

Query 5: *An image of a giant Buddha statue 34m high and weighing 250 tons in Hong Kong*

Figure 10: Apply filter with Optical Character Recognition(OCR) to find relevant frames.

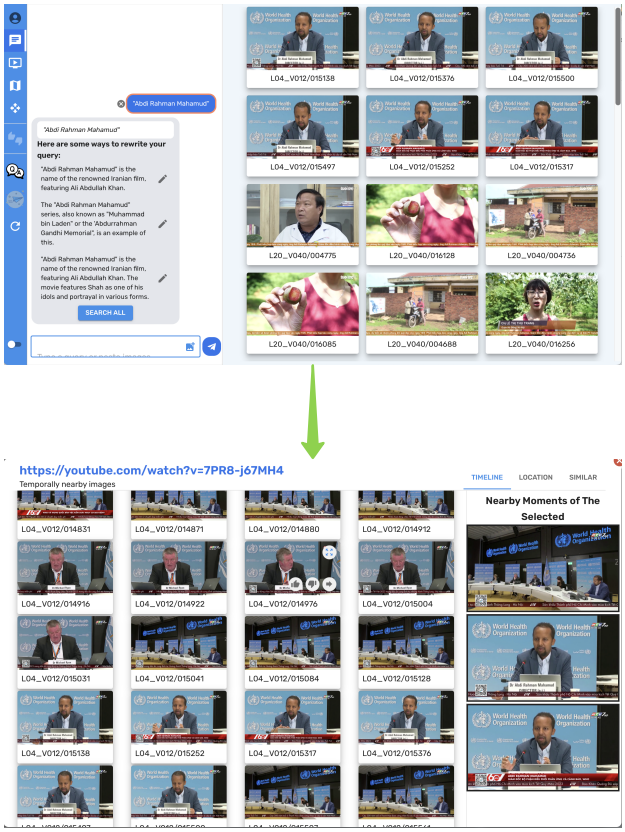
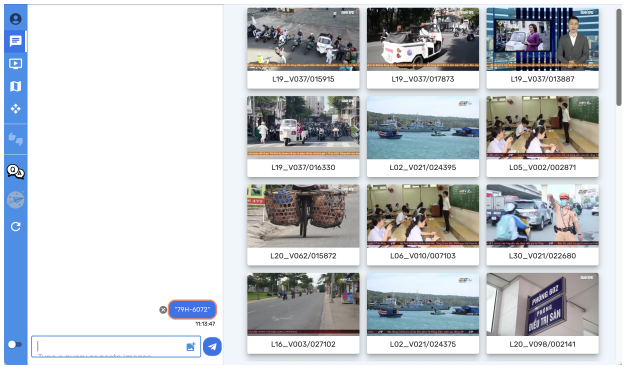
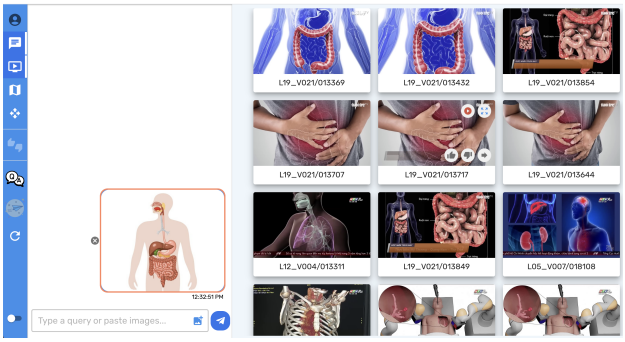


Figure 11: Apply filter with Optical Character Recognition(OCR) to find relevant frames.



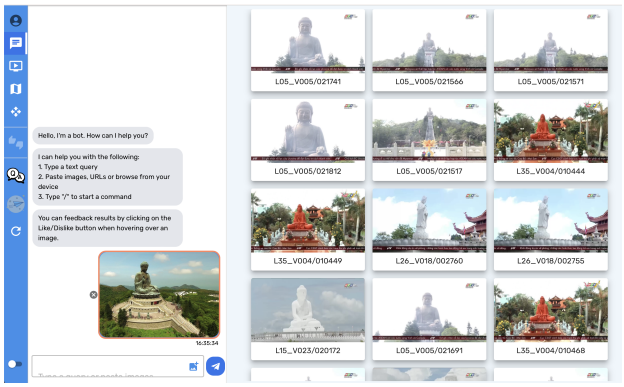
Given the substantial details pertaining to the colossal Buddha statue in Hong Kong, specifically its height and weight, it is unequivocally apparent that the implementation of visual search would be advantageous for this query. Consequently, procuring an image

Figure 12: Utilize visual search to identify the matched frames.



from a search engine such as Google could yield effective results, as illustrated in Figure 13.

Figure 13: Utilize visual search to identify the matched frames.



6 CONCLUSION

In conclusion, the NewsInsight video event retrieval system is a comprehensive system that uses a variety of mechanisms and features to provide insightful and relevant search results for video event data. It incorporates semantic search mechanisms from state-of-the-art systems and focuses on using spatial information to retrieve relevant information from video event data. The use of the BLIP model for zero-shot image-text retrieval and Elasticsearch for filtering irrelevant images has been shown to improve the precision and recall scores of the system in previous experiments. The integration of visual similarity search functionality and explicit relevance feedback further enhances the accuracy of the search results, while AI-based query description rewriting and visual example generation features provide additional support for end-users. Our system, NewsInsight, has demonstrated its efficacy in retrieving pertinent images corresponding to a given query.

ACKNOWLEDGMENTS

This research was funded by Vingroup and supported by Vingroup Innovation Foundation (VINIF) under project code VINIF.2019.DA19.

REFERENCES

[1] Naushad Alam, Yvette Graham, and Cathal Gurrin. 2022. Memento 2.0: An Improved Lifelog Search Engine for LSC'22. In *Proceedings of the 5th Annual on Lifelog Search Challenge* (Newark, NJ, USA) *(LSC '22)*. Association for Computing Machinery, New York, NY, USA, 2–7. https://doi.org/10.1145/3512729.3533006

[2] Ahmed Alateeq, Mark Roantree, and Cathal Gurrin. 2022. Voxento 3.0: A Prototype Voice-Controlled Interactive Search Engine for Lifelog. In *Proceedings of the 5th Annual on Lifelog Search Challenge* (Newark, NJ, USA) *(LSC '22)*. Association for Computing Machinery, New York, NY, USA, 43–47. https://doi.org/10.1145/3512729.3533009

[3] Trong-Le Do et al. 2023. News Event Retrieval from Large Video Collection in Ho Chi Minh City AI Challenge 2023. In *The 12th International Symposium on Information and Communication Technology, SoICT 2023, Ho Chi Minh City, Vietnam, December 7-8, 2023*. ACM.

[4] Cathal Gurrin, Björn Þór Jónsson, Duc Tien Dang Nguyen, Graham Healy, Jakub Lokoc, Liting Zhou, Luca Rossetto, Minh-Triet Tran, Wolfgang Hürst, Werner Bailer, and Klaus Schoeffmann. 2023. Introduction to the Sixth Annual Lifelog Search Challenge, LSC'23. In *Proc. International Conference on Multimedia Retrieval (ICMR'23)* (Thessaloniki, Greece) *(ICMR '23)*. New York, NY, USA. https://doi.org/10.1145/3591106.3592304

[5] Silvan Heller, Ralph Gasser, Mahnaz Parian-Scherb, Sanja Popovic, Luca Rossetto, Loris Sauter, Florian Spiess, and Heiko Schuldt. 2021. Interactive Multimodal Lifelog Retrieval with Vitrivr at LSC 2021. In *Proceedings of the 4th Annual on Lifelog Search Challenge* (Taipei, Taiwan) *(LSC '21)*. Association for Computing Machinery, New York, NY, USA, 35–39. https://doi.org/10.1145/3463948.3469062

[6] Nhat Hoang-Xuan, Hoang-Phuc Trang-Trung, E.-Ro Nguyen, Thanh-Cong Le, Mai-Khiem Tran, Tu-Khiem Le, Van-Tu Ninh, Cathal Gurrin, and Minh-Triet Tran. 2022. Flexible Interactive Retrieval SysTem 3.0 for Visual Lifelog Exploration at LSC 2022. In *LSC@ICMR 2022: Proceedings of the 5th Annual on Lifelog Search Challenge, Newark, NJ, USA, June 27 - 30, 2022*, Cathal Gurrin, Graham Healy, Liting Zhou, Björn Þór Jónsson, Duc-Tien Dang-Nguyen, Jakub Lokoc, Minh-Triet Tran, Wolfgang Hürst, Luca Rossetto, and Klaus Schoeffmann (Eds.). ACM, 20–26. https://doi.org/10.1145/3512729.3533013

[7] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. arXiv:cs.CV/2301.12597

[8] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*. PMLR, 12888–12900.

[9] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557* (2019).

[10] Jakub Lokoč, Patrik Veselý, František Mejzlík, Gregor Kovalčík, Tomáš Souček, Luca Rossetto, Klaus Schoeffmann, Werner Bailer, Cathal Gurrin, Loris Sauter, Jaeyub Song, Stefanos Vrochidis, Jiaxin Wu, and Björn þóR Jónsson. 2021. Is the Reign of Interactive Search Eternal? Findings from the Video Browser Showdown 2020. *ACM Trans. Multimedia Comput. Commun. Appl.* 17, 3, Article 91 (jul 2021), 26 pages. https://doi.org/10.1145/3445031

[11] Thao-Nhu Nguyen, Tu-Khiem Le, Van-Tu Ninh, Minh-Triet Tran, Thanh Binh Nguyen, Graham Healy, Sinéad Smyth, Annalina Caputo, and Cathal Gurrin. 2022. LifeSeeker 4.0: An Interactive Lifelog Search Engine for LSC'22. In *Proceedings of the 5th Annual on Lifelog Search Challenge* (Newark, NJ, USA) *(LSC '22)*. Association for Computing Machinery, New York, NY, USA, 14–19. https://doi.org/10.1145/3512729.3533014

[12] Thao-Nhu Nguyen, Tu-Khiem Le, Van-Tu Ninh, Minh-Triet Tran, Nguyen Thanh Binh, Graham Healy, Annalina Caputo, and Cathal Gurrin. 2021. LifeSeeker 3.0: An Interactive Lifelog Search Engine for LSC'21. In *Proceedings of the 4th Annual on Lifelog Search Challenge* (Taipei, Taiwan) *(LSC '21)*. Association for Computing Machinery, New York, NY, USA, 41–46. https://doi.org/10.1145/3463948.3469065

[13] Tien-Thanh Nguyen-Dang, Xuan-Dang Thai, Gia-Huy Vuong, Van-Son Ho, Minh-Triet Tran, Van-Tu Ninh, Minh-Khoi Pham, Tu-Khiem Le, and Graham Healy. 2023. LifeInsight: An Interactive Lifelog Retrieval System with Comprehensive Spatial Insights and Query Assistance. In *Proceedings of the 6th Annual ACM Lifelog Search Challenge* (Thessaloniki, Greece) *(LSC '23)*. Association for Computing Machinery, New York, NY, USA, 59–64. https://doi.org/10.1145/3592573.3593106

[14] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*. 2641–2649.

[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. 2021. Learning Transferable Visual Models From Natural Language Supervision. *CoRR* abs/2103.00020 (2021). arXiv:2103.00020 https://arxiv.org/abs/2103.00020

[16] 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. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:cs.CV/2103.00020

[17] Ricardo Ribiero, Alina Trifan, and Antonio J. R. Neves. 2022. MEMORIA: A Memory Enhancement and MOment RetrIeval Application for LSC 2022. In *Proceedings of the 5th Annual on Lifelog Search Challenge* (Newark, NJ, USA) *(LSC '22)*. Association for Computing Machinery, New York, NY, USA, 8–13. https://doi.org/10.1145/3512729.3533011

[18] Joseph John Rocchio Jr. 1971. Relevance feedback in information retrieval. *The SMART retrieval system: experiments in automatic document processing* (1971).

[19] Ly-Duyen Tran, Manh-Duy Nguyen, Nguyen Thanh Binh, Hyowon Lee, and Cathal Gurrin. 2021. Myscéal 2.0: A Revised Experimental Interactive Lifelog Retrieval System for LSC'21. *Proceedings of the 4th Annual on Lifelog Search Challenge* (2021).

[20] Ly-Duyen Tran, Manh-Duy Nguyen, Binh Nguyen, Hyowon Lee, Liting Zhou, and Cathal Gurrin. 2022. E-Myscéal: Embedding-Based Interactive Lifelog Retrieval System for LSC'22. In *Proceedings of the 5th Annual on Lifelog Search Challenge* (Newark, NJ, USA) *(LSC '22)*. Association for Computing Machinery, New York, NY, USA, 32–37. https://doi.org/10.1145/3512729.3533012

[21] Minh-Triet Tran, Thanh-An Nguyen, Quoc-Cuong Tran, Mai-Khiem Tran, Khanh Nguyen, Van-Tu Ninh, Tu-Khiem Le, Hoang-Phuc Trang-Trung, Hoang-Anh Le, Hai-Dang Nguyen, Trong-Le Do, Viet-Khoa Vo-Ho, and Cathal Gurrin. 2020. FIRST - Flexible Interactive Retrieval SysTem for Visual Lifelog Exploration at LSC 2020. In *Proceedings of the Third ACM Workshop on Lifelog Search Challenge, LSC@ICMR 2020, Dublin, Ireland, June 8-11, 2020*, Cathal Gurrin, Klaus Schöffmann, Björn Þór Jónsson, Duc-Tien Dang-Nguyen, Jakub Lokoc, Minh-Triet Tran, and Wolfgang Hürst (Eds.). ACM, 67–72. https://doi.org/10.1145/3379172.3391726

[22] Maxim Kuznetsov Vladimir Vorobev. 2023. A paraphrasing model based on ChatGPT paraphrases.