An Optimized And Interactive Video Event Retrieval System With An Improved Temporal Algorithm

Kiet Pham Gia ✉[1 [0009−0003−6261−6177]], Nhi Nguyen Truong[1 [0009−0001−6334−6944]], Long Nguyen Huynh[1 [0009−0009−1919−6545]], Vinh Nguyen Quoc[1 [0009−0003−9046−2742]], Phuong Le Tran[1 [0009−0003−0248−6171]], Binh Tran Le[1 [0009−0008−0938−3585]], Tri Pham Xuan[1 [0009−0008−1679−4737]], Duong Tran Ham[1 [0009−0008−2088−8023]], Tin Huynh[1 [0000−0002−9139−9891]], and Kiem Hoang[1 [0009−0007−4003−6736]]

The Saigon International University, Ho Chi Minh City, Viet Nam
{phamgiakietk14, nguyentruongcongnhik15, nguyenhuynhphilongk14, nguyenquocvinhk15, letransongphuongk14, tranlehaibinhk12, phamxuantri, tranhamduong, huynhngoctin, hoangkiem}@siu.edu.vn

Abstract. Video event retrieval is the process of identifying and extracting specific events or actions from a video collection based on a given query or description. Effective retrieval systems must adeptly manage the storage, indexing, searching, and delivery of such information. However, current approaches often focus on speed or accuracy, while user interactivity and scalability are not the most focused keywords. Therefore, this paper introduces PumpkinV2, an interactive video retrieval system with high precision and scalability. Capturing a visually balanced user interface, the system deployed a temporal-enabled visual-text association search pipeline with adaptive reranking methods. The retrieval results are enhanced by applying multiplicative and additive approaches in the integrated temporal algorithm. Furthermore, the system is built on top of a highly scalable vector database, combined with vector quantization, format-optimized data and a production-grade gateway web server. PumpkinV2 achieved outstanding results at AI Challenge HCMC 2024, an annual video event retrieval competition, with 94 percent accuracy during qualifying rounds and ranked top 10 amongst finalists, proving its capability of as a robust, scalable and highly interactive system for video event retrieval.

Keywords: Information System · Video Event Retrieval · Temporal Retrieval · Interactive Retrieval.

1 Introduction

Video content serves as a dynamic and powerful source of information [29], combining various forms of content such as text, images, audio, etc in a single instance of data. The explosion of video content has revolutionized how the digital

world operates. As an example, a recent survey of big data in Astronomy [9] revealed a single telescope project can generate up to hundreds of petabytes of captured videos annually. As the volume of such data continues to surge across various sectors, there is an escalating demand for effective retrieval systems, as a droplet of water is often all one needs in the vast ocean of digital data.

Seeking an answer to this demand, various competitions and workshops are being hold all over the globe. TREC Video Retrieval Evaluation (TRECVID) [1], Video Browser Showdown (VBS) [2], Lifelog Search Challenge (LCS) [3] and for the last few years, AI Challenge Ho Chi Minh City (AIC HCMC) [4] has attracted many researchers and corporations in Vietnam.

The AI Challenge HCMC is an annual Video Event Retrieval competition where participants must retrieve visual events from a 300-hour video collection of television news programs. The content, ranging from weather reports to interviews and sports events, poses challenges due to its volume, variety, and complexity. The competition evaluates systems based on both accuracy and response time, making it a credible test for building scalable, accurate, and interactive Video Event Retrieval Systems.

Emerging from these competition were capable video event retrieval systems, handling massive data collections with ease. These systems allowed a variety of querying method, from visual-text association [22], object-based search [23], etc to even geographical information-based retrieval [23]. However, a majority of these methods can only retrieve information based on a single instance, while temporal-enabled methods are still being a premium feature [25] amongst such systems. More importantly, many of these systems are lacking many characteristics needed in reality. Aside from speed and accuracy, a video event retrieval system also has to leverage scalability and interactivity to be valuable for practical usages [24].

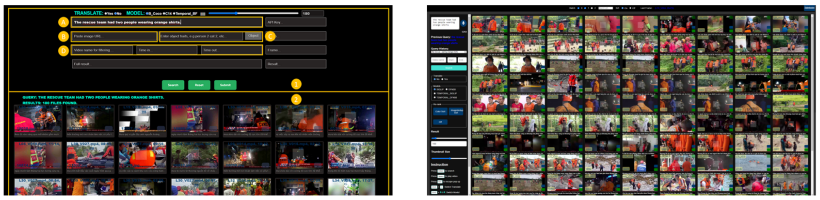

Fig. 1. Comparison between Pumpkin User Interface at AIC 2023 (Left) and PumpkinV2 at AIC 2024 (Right).

[1] https://trecvid.nist.gov/
[2] https://videobrowsershowdown.org/
[3] http://lifelogsearch.org/lsc/
[4] https://aichallenge.hochiminhcity.gov.vn/

Having also participated in the AIC 2023, PumpkinV2's predecessor [17] achieved remarkable results with temporal search and a variety of functionality, but it's not without its flaws. Therefore, various optimization techniques for data and vectors were deployed to speed up the retrieval process. PumpkinV2 also experimented on new Shot Boundary Detection methods, as well as proposing a method to "group" shots in structured video content. Additionally, PumpkinV2 adopted reportedly state-of-the-art visual-text embedding models [21]. For temporal algorithm, we proposed a score-based additive and multiplicative approach to improve the accuracy of the retrieval targets. We also deployed a number of post-retrieval operation, such as filter by metadata, transcription content and reranking methods which are based on shot collections, colors and visual hashes. Additionally, the system's User Interface had been redesigned, as shown in Figure 1, offering a suitable visual composition. Finally, PumpkinV2's pipeline is deployed on a production-grade WSGI server and a basic distributed file system between remote server and local storage. Using AI Challenge HCMC 2024 as a testing ground, our system achieved an average accuracy of 94 percent during qualifying rounds and ranked top 10 amongst finalists with an average time to find the correct answer being under a minute.

In short, utilizing the AI Challenge HCM 2024 as a credible testing ground, the 2024 PumpkinV2 system inherited the core principles of its predecessor - scalability, precision and user-friendliness - while improving upon its flaws and:

- Proposing an improved temporal retrieval algorithm with multiplicative and additive approaches.
- Employing a robust visual-text association search pipeline with state-of-the-art embedding models, new Shot Boundary Detection methods, database optimization, built on top of a scalable vector database and production-grade gateway web server as well as a basic distributed file system to promote scalability.
- Proposing a more visually balanced User Interface as well as adding rerank methods for a more efficient and interactive retrieval process.

2 Related Work

The rapid growth of multimedia content, especially videos, has made managing and retrieving data from large datasets a tremendous challenge. Competitions like TRECVID, Lifelog Search Challenge (LSC), Video Browser Showdown (VBS), and the AI Challenge Ho Chi Minh City (AIC HCM) encourage innovation in scalable video retrieval systems, with AIC HCM focusing on AI-based solutions for practical use in Vietnam.

To address these challenges, various systems have exploited innovative approaches to enhance the retrieval process [24]. As a foundation, nearly all of these systems utilize joint text-visual embeddings such as OpenCLIP [8], CLIP [18], BLIP [15], and ALADIN [16], etc. Aside from the utilization of these embeddings, addtional types of information can make retrieval process more intuitive

and effective. For instance, Localized Search enables the identification of specific locations within video content [23]. Automatic Speech Recognition (ASR) and Object Detection have significantly enhanced the searchability of spoken and visual content [23]. Systems such as AGAIN [22] employ dual image-to-text representations, whereas Voxento-Pro [1] facilitates voice-based searches. TRECVID 2023 systems, such as BUPT_MCPRL [20], Doshisha University [7] and RUC_AIM3 [13], leverage ASR and text-based video conversion to improve retrieval outcomes. Event search with systems like LifeInsight 2.0 [27] and Exquisitor [14] focusing on identifying specific actions, scenes, or dialogues. Technologies such as ASR and Optical Character Recognition (OCR) further enhance non-visual content retrieval [19].

As video content retrieval evolves, temporal queries are crucial to enhancing systems' search capabilities [24]. Systems like AGAIN [22] adopt a multimodal approach, integrating image, text, and audio to enhance retrieval. On the other hand, Vibro [19] introduces a hierarchical graph for temporal navigation, while LifeInsight 2.0 [27] and PraK Tool V2 [25] leverage temporal search mechanisms for lifelog and distant activity retrieval.

In terms of user experience, VISIONE 5.0 [3] emphasizes a minimalistic design, while Vibro [19] and ViewsInsight [26] incorporate advanced navigation and interaction tools. Furthermore, query assistance has become a critical feature, with systems like NewsInsight [28] and Exquisitor [14] leveraging AI-driven reformulation techniques to enhance retrieval accuracy.

While these systems demonstrate significant advancements, they often only focused on one aspect or another, leading to missed opportunities for a more holistic solution. A key limitation is the insufficient focus on retrieval efficiency, with many systems not prioritizing retrieval and content delivery speed. Additionally, although some interfaces are user-friendly, they frequently lack the flexibility and advanced functionalities needed to support users in navigating complex video datasets.

PumpkinV2 addresses these gaps with improved temporal search algorithms, optimized system performance, and an enhanced user interface. The system used a novel multiplicative and additive scoring mechanism for temporal event detection, along with Shot Boundary Detection and reranking based on color, visual hashes, and shot collections. This approach improves retrieval accuracy, scalability, and ease of use in navigating complex video data.

3 PumpkinV2's System with An Improved Temporal Algorithm

Figure 2 illustrated the system's overall pipeline. In Data Processing, PumpkinV2 performed Shot Boundary Detection, Video Transcription Extraction and CLIP Embeddings Extraction to extract various features from the video collection, these features are then optimized and stored in a vector database as vectors and corresponding payloads. In Query Processing part, PumpkinV2's User Interface allows the user to retrieve a single instance or a sequence of visual instance

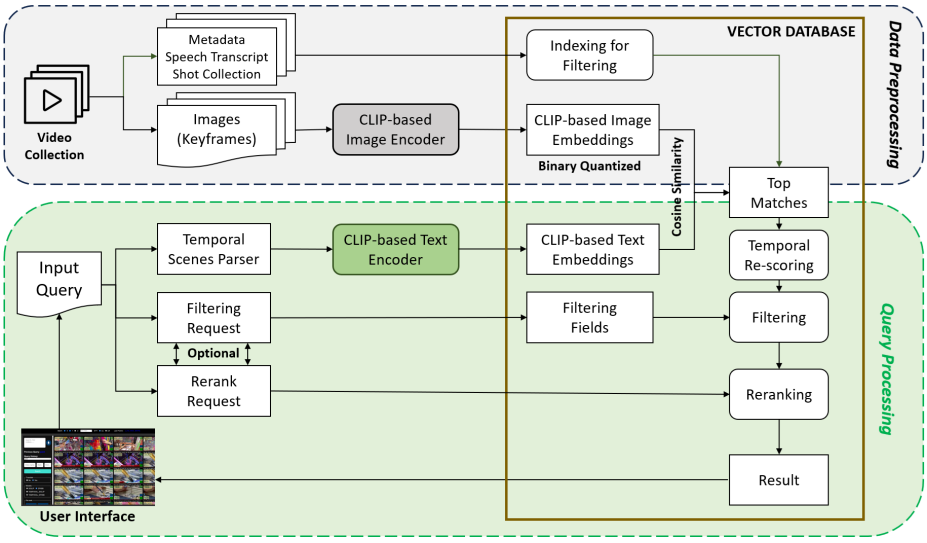

Fig. 2. PumpkinV2's Pipeline as a Video Event Retrieval System.

by virtue of the temporal algorithm. Furthermore, the results obtained can be filtered or reranked to enhance interactivity. These processes are deployed on a production-grade WSGI server with distributed file system method, promoting stability and scalability.

3.1 Data Preprocessing

Shot Boundary Detection and Collections of Shots: Shot Boundary Detection involves identifying transitions between shots in a video sequence. A "shot" is a series of consecutive frames captured by a single camera without interruption and keyframes are often used to represent a shot. Two Shot Boundary Detection methods experimented in PumpkinV2: PySceneDetect [5] with traditional algorithms, and an Autoshot [33] model that was pretrained on the manually labeled AI Challenge 2022 dataset [6].

For AIC 2024 dataset, PySceneDetect detected 1550572 keyframes, nearly twice as many keyframes as Autoshot with a figure of 762654, as shown in Table 1. After detection, we applied Imagedupe [7] to remove duplicate keyframes. Surprisingly, 467423, about one-third of PySceneDetect's keyframes were considered duplicates, while Autoshot only had 92730 duplications. Additionally, in the AIC HCMC 2024 queries, PySceneDetect missed several keyframes, particularly those that were blurry or dimly lit. Autoshot, however, successfully detected the ones

[5] https://github.com/Breakthrough/PySceneDetect
[6] https://github.com/PhucNguyenLamp/Shot_Detection
[7] https://github.com/idealo/imagededup

PySceneDetect missed. Based on this observation, PumpkinV2 mainly utilized the results from Autoshot method for further operations.

Fig. 3. An example of using the cue of MC talking (green) as the start and end of each piece of news (red) to create a collection of shots.

For the AI Challenge HCMC 2024 alone, since most of the video content are of news programs, they have a structure where a piece of news starts and ends with scene of MC discussing the news. Based on this observed fact, we marked all the scenes of MC talking and from that, PumpkinV2 managed to have a collection of each news piece, as shown in Figure 3. We believe this approach can be of high value when structured video content are present in a video collection.

Table 1. Number of keyframes and duplications of each SBD method and the number of shot collections for AIC 2024 dataset.

Method	Number of Keyframes	Number of Duplications
PySceneDetect	1550572	467423
Autoshot	762654	92730
Number of Shot Collections: 28059		

Video Transcript Extraction: PumpkinV2 also deployed a pipeline to extracting speech content from videos. Firstly the raw audio files are extracted, these raw audio files have very long duration, in which speech content don't appear every second. A Voice Activity Dectector (VAD) is needed for this part in order to split the audio data into separated speech segments. PumpkinV2 utilized Pyannote [6] for VAD and our own Speech-To-Text (STT) model - wav2vec2 [5] fine-tuned on VLSP datasets [8] - as the STT models.

CLIP-based Models For Visual-Text Embeddings: Based on the results reported in [21] as well as our own observation, we deployed SIGLIP [30] and DFN5B [10] as our main visual-text embeddings models. From the experience during AIC 2024 rounds, SIGLIP has a very broad understanding of visual terms (for example it can recognizes the back of a white fluffy bear lying down, but DFN5B to bring that keyframes to top results), but DFN5B handles long query much more robust (which is often the case in AIC).

[8] https://vlsp.org.vn/

Vector Database: This iteration of PumpkinV2 utilized Qdrant [9] as its main vector database framework. Qdrant is an open-source vector search engine and database designed for efficient and scalable similarity search on high-dimensional data. For each vector of keyframes embedding, PumpkinV2 also stores the following information as its payload: Video Name, Keyframe Name, Video FPS, Speech Transcript Content and Near-duplicated Frames List (The list of frame with vector similarity of 94 percent or greater which belong to different shots compared to the current keyframes). Each field of information is indexed so that it can be use a filter field for retrieval operations later on.

3.2 Query Processing

Additive Approach and Multiplicative Approaches in Score-based Temporal Search: Temporal Search refers to the process of retrieving a target based on or time-related features of the content [2]. The temporal algorithm of Pumpkin 2023 processed a query by narrowing the search space based on results within a specified time range. However, this way, the final result is based on the score of the final scene only, making results obtained heavily skewed towards it.

In order to improve upon this problem, we experimented on a way to perform a multiplicative and additive approaches on the scores of relevant results in a given range. The final score Y for a keyframe F of video v for the query Q that has n scenes q_i is calculated as follows:

– Additive Approach:

$$Y(v, F, Q) = \sum_{i=0}^{n} X(v, f_i \in [f + i\Delta, f + (i+1)\Delta], q_i) \tag{1}$$

– Multiplicative Approach:

$$Y(v, F, Q) = \prod_{i=0}^{n} X(v, f_i \in [f + i\Delta, f + (i+1)\Delta], q_i) \tag{2}$$

Where Δ denotes the time range (in second) allowed for the keyframes f_i to be considered to be relevant to the keyframes f_{i-1} and X is the cosine similarity score of keyframes (v, f_i) for scene q_i.

Using these approaches for temporal retrieval, we observed a significant improvement in accuracy, as results matching multiple scenes received higher scores than those matching only one or two scenes. Overall, the Additive Approach is more stable for all temporal queries, but for dense queries, like finding shots in fast-paced sports programs, the Multiplicative Approach is better at pushing the target to the top results. In AIC 2024, since each news program typically lasts up to a minute, we set $\Delta = 1500$, meaning shots within a minute are considered relevant to each other.

[9] https://qdrant.tech/

Reranking Methods: PumpkinV2 system also employed rerank methods as a mean of post-retrieval processing, encompassing three main methods:

- **Shot Collection-based Rerank**: The system groups related shots into their corresponding collections, and retrieval results are organized based on the score of the collections' elements. This method was favored during AIC 2024 for simplifying result viewing.
- **Color-based Rerank**: Keyframes are reordered using the Step Color Sorting algorithm [10]. This enhances visual clarity, helping users locate their target more efficiently.
- **Visual hashes-based Rerank**: Results are sorted by perceptual hash similarity [11], which aids users in finding their retrieval target more comfortably.

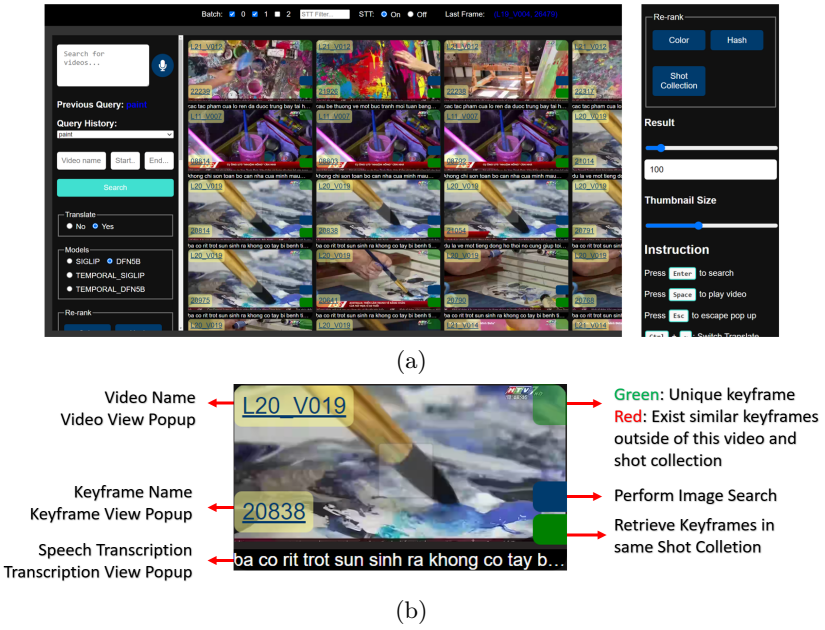

Fig. 4. (a) The overall view of PumpkinV2's redesigned User Interface and (b) The information displayed on each retrieval result and the operations' shortcuts.

A Redesigned User Interface: Designing the PumpkinV2's UI, we focused on the keywords: *user-centered*, *responsiveness*, *minimalism*, and *scalability*. The UI maintains a clean layout, focusing on critical information. The left and top bars contain the query section, where users can choose embedding models, translate query text, perform voice queries, and filter results by metadata (video name,

[10] https://www.alanzucconi.com/2015/09/30/colour-sorting/

time, dataset batch), transcription, or rerank methods. Users can adjust the quantity of retrieval results, thumbnail size, and toggle the transcription field to improve loading speed. Keyboard shortcuts allow quick toggling of these operations. The right section displays retrieval results, where each keyframe includes essential data (video name, keyframe, transcription). Pop-up features (video, keyframe, transcription views) enable quick actions like playing videos or retrieving similar keyframes.

3.3 Optimization Techniques

Since AIC 2023, "optimization" has been a key focus in improving the PumpkinV2 system. Firstly, scalar quantization [32], specifically Binary Quantization has been implemented, reduced memory usage by a factor of 32 and accelerated vector retrieval by 27 times when applied to CLIP vectors of size 1024 (DFN5B) and 1152 (SIGLIP). Torch Compile [4] was also utilized to compile embedding models ahead of time, reducing embedding extraction inference time by 30 percent. For video content, video data was re-encoded using the AV1 codec [12], cutting file size by half while preserving quality. Keyframes extracted were reformatted to ".avif" [11] with 5 percent quality, reducing average size from 152KB to 4KB. Additionally, we distributed the data on both the remote server and local storage, forming a basic distributed file system, specifically moving static files (e.g., images, videos) to local machines, while keeping vectors, JSONs, and logs on remote servers, greatly reduced network usage to ensure concurrency and scalability. Lastly, rebuilding APIs on the production Gunicorn WSGI server [12] improved API response time by efficiently managing multiple requests through worker processes. Applying these techniques, PumpkinV2 is capable of performing vector search in 50 millisecond and the content can be delivered to UI in approximately 800 millisecond.

4 The AI Challenge HCMC 2024 as a benchmark for Video Event Retrieval System

4.1 Datasets

The video dataset for the AIC 2024 is sourced from reputable YouTube channels "60 Giay Official", "HTV Sports", "Bao Thanh Nien", "ViVU TV", "HTV Giai Tri", "HTV Entertainment", and "Bao Tuoi Tre". They consist of diverse video content categorized into distinct themes, reflecting a broad spectrum of societal interests. The distribution of videos across these themes are shown in Figure 5, and the statistical data for the datasets provided are reported in Table 2.

With this dataset, challenges include missing or inaccurate metadata, inconsistent image quality that affects recognition tasks, obscured scenes that hinder

[11] barman2020evaluation
[12] https://gunicorn.org/

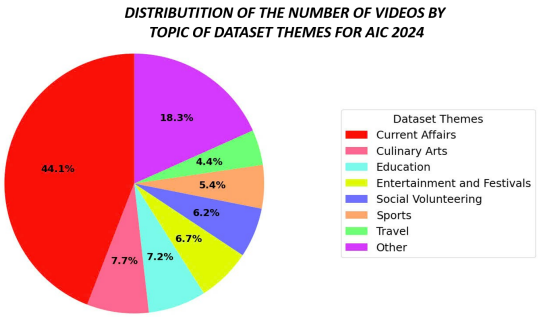

Fig. 5. Distribution of content theme among AIC 2024 dataset's videos.

Table 2. Number of videos, duration and number of keyframes extracted by Autoshot method from the AIC 2024 datasets. Bacth 1 and 2 were used in qualifying rounds, while all three were included during final.

Batch	Number of videos	Duration	Number of keyframes	Round
1	363	121h:14m:52s	364856	Qualifying, Final
2	363	103h:25m:10s	288440	Qualifying, Final
3	346	67h:45m:54s	109358	Final
Total	**1072**	**292h:26m:56s**	**762654**	

detection of key elements, and the massive volume of content, which complicates segmentation and classification of relevant sections. Developing effective solution for processing, managing, and categorizing this diverse and sometimes flawed data was one of our most important goal for PumpkinV2.

4.2 PumpkinV2 in AIC 2024

Textual Know-Item Search: In Textual Know-Item Search (KIS) task, participants are required to use natural language descriptions to retrieve specific video segments. By analyzing the relationship between image data and text, PumpkinV2 identifies scenes that match the query, for example as seen in Figure 6 involving snowy scenes. The system processes detailed multi-point descriptions, such as the example of snowfall, trees, and cars covered in snow, and returns highly relevant results. The color rerank function provided a more intuitive display of results and the transcription viewing function enables accurate manual matching between the query description and spoken dialogue in the video, significantly improving retrieval accuracy.

Visual Know-Item Search: In the Visual Known Item Search task, teams must locate and submit an exact video segment which is shown earlier. The challenge involves identifying a 20-second portion, for example, of a 30-second bakery news clip that focuses on bread and store details, as shown in Figure 7. Thanks to the strategic approach of Shot Collection, PumpkinV2 was able

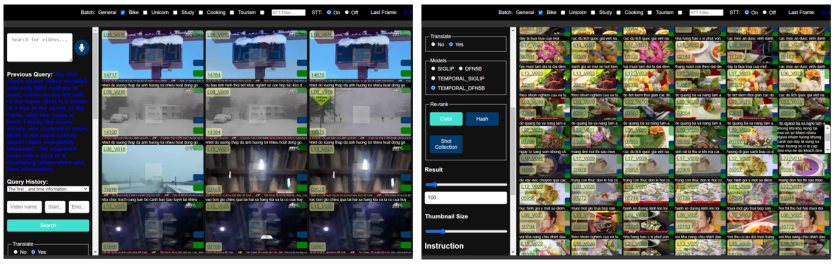

Fig. 6. PumpkinV2 temporal search retrieving the AIC 2024's query *"Heavy snowfall obscures trees on the left. A tree with two ropes appears next, followed by snow-covered cars and roads. The final shot shows temperature and time."* (left) and using color rerank and transcription viewing on an AIC 2024's dish-related query (right).

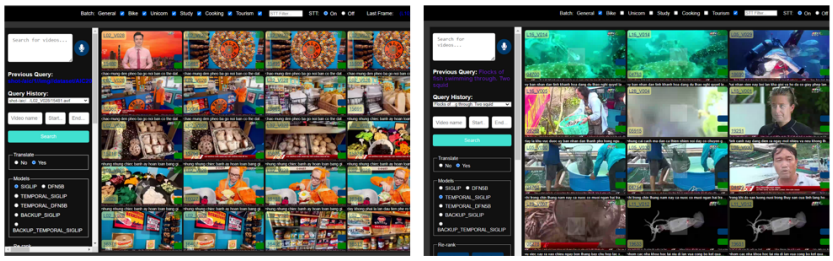

Fig. 7. PumpkinV2 retrieving results for two of Visual KIS queries in Final Round: a scene with within a bread store (left) and an underwater scene (right).

to classify the whole enclosed scenes. Additionally, our temporal approach was critical for more complex queries requiring specific sequences of events, such as locating consecutive scenes of a flattened coral reef, fish swimming, and two squid in the ocean, ensuring the results meet the temporal and sequential criteria. PumpkinV2 trivialized this task, in which the average time for our team to answer each query was under 15 seconds.

Visual Question Answering: In Visual Question Answering, many queries require reasoning that goes beyond surface-level recognition, where even state-of-the-art Large VLMs often struggle [31]. For example, in the AIC 2024 semi-round, one task focused on a football match between Uzbekistan and North Korea, asking how many Uzbekistan players were visible at the moment a penalty kick was taken. Another query involved counting yellow balloons in a crowded street scene, where the balloons often overlapped or were obscured, requiring manual frame-by-frame inspection and temporal awareness skills that come naturally to human viewers but are challenging for current models. In both cases, PumpkinV2's ability to perform temporal textual search proved invaluable, allowing our team to quickly retrieve and review video segments, with interactive

popup viewing features further enhancing efficiency when we need to manually draft the answers.

In the final round, our team used a divide-and-conquer approach, assigning each member specific tools to maximize PumpkinV2's capabilities. The improved temporal algorithm proved essential for most queries, while reranking methods, particularly shot collection rerank, efficiently grouped results. The transcription filter was also valuable for challenging queries, as it accurately predicted speech content in several instances. Additionally, PumpkinV2's redesigned UI and optimizations ensured stable, competitive performance for the team.

Using these tasks in AI Challenge HCMC 2024 as a credible testing ground. PumpkinV2 succesfully identified 80 out of 85 queries in qualifying rounds, achieving an accuracy of 94 percent. Amongst the finalists, PumpkinV2 outperformed many other systems when it comes to Visual KIS task, as our temporal trivialized the task when we have abundant temporal information. Additionally, the optimization enabled PumpkinV2 to processed a query in approximately 800 millisecond to display top 200 results. even in poor internet environment. As a result, PumpkinV2 ranked top 10 amongst finalists, with an average time needed to find the correct answer being under a minute, proving the system's abilities as an effective Video Event Retrieval System.

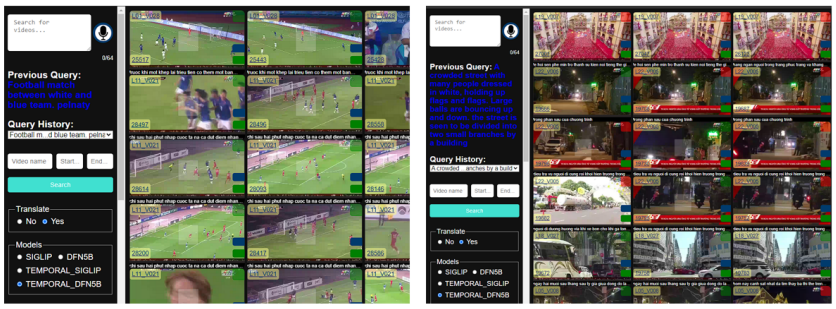

Fig. 8. AIC 2024's tricky VQA query for soccer players counting (left) and yellow ball number counting (right).

5 Conclusion

In conclusion, PumpkinV2 stands as a candidate for significant advancements in the realm of video event retrieval systems, addressing the dual challenges of scalability and interactivity while maintaining high precision. PumpkinV2 integrates state-of-the-art visual-text embedding models, advanced temporal algorithms, and database optimization. Its redesigned user interface further enhances interactivity, making the system more user-friendly. The system's remarkable performance at AI Challenge HCMC 2024, with 94 percent accuracy in qualifying rounds and ranked top 10 amongst finalists, solidifies its position as a robust solution for video event retrieval task.

References

1. Alateeq, A., Roantree, M., Gurrin, C.: Voxento-pro: An advanced voice lifelog retrieval interaction for multimodal lifelogs. In: Proceedings of the 7th Annual ACM Workshop on the Lifelog Search Challenge. p. 105–110. LSC '24, Association for Computing Machinery, New York, NY, USA (2024)
2. Alonso, O., Strötgen, J., Baeza-Yates, R., Gertz, M.: Temporal information retrieval: Challenges and opportunities. Twaw **11**, 1–8 (2011)
3. Amato, G., Bolettieri, P., Carrara, F., Falchi, F., Gennaro, C.e.a.: Visione 5.0: Enhanced user interface and ai models for vbs2024. In: MultiMedia Modeling. pp. 332–339. Springer Nature Switzerland, Cham (2024)
4. Ansel, J., Yang, E., He, H., Gimelshein, N., et al, J.: Pytorch 2: Faster machine learning through dynamic python bytecode transformation and graph compilation. In: Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2. pp. 929–947 (2024)
5. Baevski, A., Zhou, H., Mohamed, A., Auli, M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. CoRR **abs/2006.11477** (2020), https://arxiv.org/abs/2006.11477
6. Bredin, H., Laurent, A.: End-to-end speaker segmentation for overlap-aware resegmentation. In: Proc. Interspeech 2021. Brno, Czech Republic (August 2021)
7. Chen, Z., Li, F., Seibel, M.S., Brügge, N.S., Ohsaki, M., Handels, H.e.a.: Doshisha university, universität zu lübeck and german research center for artificial intelligence at trecvid 2023: Miqg task (2023)
8. Cherti, M., Beaumont, R., Wightman, R., et al, M.W.: Reproducible scaling laws for contrastive language-image learning. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 2818–2829 (2022), https://api.semanticscholar.org/CorpusID:254636568
9. Faaique, M.: Overview of big data analytics in modern astronomy. International Journal of Mathematics, Statistics, and Computer Science **2**, 96–113 (2024)
10. Fang, A., Jose, A.M., Jain, A., Schmidt, L., Toshev, A., Shankar, V.: Data filtering networks. arXiv preprint arXiv:2309.17425 (2023)
11. Farid, H.: An overview of perceptual hashing. Journal of Online Trust and Safety **1**(1) (2021)
12. Han, J., Li, B., Mukherjee, D., Chiang, C.H., et al, G.: A technical overview of av1. Proceedings of the IEEE **109**(9), 1435–1462 (2021)
13. Kaiwen Wei, Zihao Yue, L.Z.Q.J.: Ruc aim3 at trecvid 2023: Video to text description (2023)
14. Khan, O.S., Sharma, U., Zhu, H., Rudinac, S., Jónsson, B.T.: Exquisitor at the lifelog search challenge 2024: Blending conversational search with user relevance feedback. In: Proceedings of the 7th Annual ACM Workshop on the Lifelog Search Challenge. p. 117–121. LSC '24, Association for Computing Machinery, New York, NY, USA (2024)
15. Li, J., Li, D., Xiong, C., Hoi, S.C.H.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning (2022), https://api.semanticscholar.org/CorpusID:246411402
16. Messina, N., Stefanini, M., Cornia, M., Baraldi, L., et al, F.: Aladin: Distilling fine-grained alignment scores for efficient image-text matching and retrieval. In: Proceedings of the 19th International Conference on Content-Based

Multimedia Indexing. p. 64–70. CBMI '22, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3549555.3549576, https://doi.org/10.1145/3549555.3549576

17. Pham Gia, K., Tran Le, H.B., Nguyen Huynh, P.L., Le Tran, S.P., Pham Hoang, L., Pham Xuan, T., Tran Ham, D., Huynh Ngoc, T., Hoang, K.: An interactive system for multimedia retrieval in video collection with temporal integration. In: Proceedings of the 12th International Symposium on Information and Communication Technology. pp. 989–996 (2023)
18. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (2021), https://api.semanticscholar.org/CorpusID:231591445
19. Schall, K., Hezel, N., Barthel, K.U., Jung, K.: Optimizing thenbsp;interactive video retrieval tool vibro fornbsp;thenbsp;video browser showdown 2024. In: MultiMedia Modeling: 30th International Conference, MMM 2024, Amsterdam, The Netherlands, January 29 – February 2, 2024, Proceedings, Part IV. p. 364–371. Springer-Verlag, Berlin, Heidelberg (2024)
20. Song, Y., Zhang, H., Ma, Z., Jiang, S., Cui, Z., Zhao, Y.: Bupt_mcprl at trecvid 2023: Video to text description and actev srl challenge (2023)
21. Sun, Q., Wang, J., Yu, Q., Cui, Y., Zhang, F., Zhang, X., Wang, X.: Eva-clip-18b: Scaling clip to 18 billion parameters. arXiv preprint arXiv:2402.04252 (2024)
22. Tran, M.N., To, T.A., Thai, V.N., Cao, T.D., Nguyen, T.T.: Again: A multimodal human-centric event retrieval system using dual image-to-text representations. In: Proceedings of the 12th International Symposium on Information and Communication Technology. p. 931–937. SOICT '23, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3628797.3628975, https://doi.org/10.1145/3628797.3628975
23. Tran Gia, B., Bui Cong Khanh, T., Tran Nhat, K., Luu Trung, K., et al, T.D.: Integrating multiple models for effective video retrieval and multi-stage search. In: Proceedings of the 12th International Symposium on Information and Communication Technology. p. 1003–1010. SOICT '23, Association for Computing Machinery, New York, NY, USA (2023)
24. Vadicamo, L., Arnold, R., Bailer, W., Carrara, F., et al, G.: Evaluating performance and trends in interactive video retrieval: Insights from the 12th vbs competition. IEEE Access **12**, 79342–79366 (2024). https://doi.org/10.1109/ACCESS.2024.3405638
25. Vopálková, Z., Yaghob, J., Stroh, M., Schlegel, U., Lokoc, J.: Searching temporally distant activities in lifelog data with prak tool v2. In: Proceedings of the 7th Annual ACM Workshop on the Lifelog Search Challenge. p. 111–116. LSC '24, Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3643489.3661131, https://doi.org/10.1145/3643489.3661131
26. Vuong, G.H., Ho, V.S., Nguyen-Dang, T.T., Thai, X.D., et al, L.: Viewsinsight: Enhancing video retrieval for vbs 2024 with a user-friendly interaction mechanism. In: MultiMedia Modeling. pp. 400–406. Springer Nature Switzerland, Cham (2024)
27. Vuong, G.H., Ho, V.S., Nguyen Dang, T.T., Thai, X.D., et al, N.H.: Lifeinsight2.0: An enhanced approach for automated lifelog retrieval in lsc'24. In: Proceedings of the 7th Annual ACM Workshop on the Lifelog Search Challenge. p. 1–6. LSC '24, Association for Computing Machinery, New York, NY, USA (2024)

28. Vuong, G.H., Ho, V.S., Nguyen-Dang, T.T., Thai, X.D., Ninh, V.T., et al, P.: Newsinsight: A comprehensive video event retrieval system with spatial insights and query assistance. In: Proceedings of the 12th International Symposium on Information and Communication Technology. p. 893–900. SOICT '23, Association for Computing Machinery, New York, NY, USA (2023)
29. Westerman, D., Spence, P.R., Van Der Heide, B.: Social Media as Information Source: Recency of Updates and Credibility of Information*. Journal of Computer-Mediated Communication **19**(2), 171–183 (01 2014). https://doi.org/10.1111/jcc4.12041, https://doi.org/10.1111/jcc4.12041
30. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11975–11986 (2023)
31. Zhao, Y., Pang, T., Du, C., Yang, X., Li, C., Cheung, N.M.M., Lin, M.: On evaluating adversarial robustness of large vision-language models. Advances in Neural Information Processing Systems **36** (2024)
32. Zhou, W., Lu, Y., Li, H., Tian, Q.: Scalar quantization for large scale image search. In: Proceedings of the 20th ACM international conference on Multimedia. pp. 169–178 (2012)
33. Zhu, W., Huang, Y., Xie, X., Liu, W., Deng, J., Zhang, D., Wang, Z., Liu, J.: Autoshot: A short video dataset and state-of-the-art shot boundary detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2023)