Interactive Retrieval System for Multi-Stream Collections: multiXview at the CASTLE 2025 Interactive Grand Challenge

Omar Shahbaz Khan*
Reykjavik University
Reykjavík, Iceland
omark@ru.is

Ujjwal Sharma*
University of Amsterdam
Amsterdam, The Netherlands
u.sharma@uva.nl

Gonçalo Marcelino*
University of Amsterdam
Amsterdam, The Netherlands
g.barretoferreiramarcelino@uva.nl

Aaron Duane
Churney
Copenhagen, Denmark
aaron.duane24601@gmail.com

Stevan Rudinac
University of Amsterdam
Amsterdam, The Netherlands
s.rudinac@uva.nl

Marcel Worring
University of Amsterdam
Amsterdam, The Netherlands
m.worring@uva.nl

Björn Þór Jónsson
Reykjavik University
Reykjavík, Iceland
bjorn@ru.is

Abstract

We introduce multiXview, an interactive retrieval framework for synchronized multi-camera video collections. It features a multi-index search engine that supports natural-language queries over visual embeddings, speech transcripts, and scene descriptions. It supports a synchronized multi-stream player offering parallel playback, and a timeline-based navigation view for temporal scoping and faceted exploration. These components address the redundancy and fragmentation of overlapping egocentric and exocentric video feeds and enable users to locate, aggregate, and reconstruct events across partial perspectives. This paper focuses on system design and implementation, with quantitative and qualitative evaluation to take place at the CASTLE 2025 Grand Challenge Interactive Track.

CCS Concepts

• **Information systems** → **Search interfaces**; **Collaborative search**; **Video search**; • **Human-centered computing** → **Interactive systems and tools**.

Keywords

Multimedia Retrieval, Multimedia Analytics, Multi-Stream Interface, CASTLE

ACM Reference Format:
Omar Shahbaz Khan, Ujjwal Sharma, Gonçalo Marcelino, Aaron Duane, Stevan Rudinac, Marcel Worring, and Björn Þór Jónsson. 2025. Interactive Retrieval System for Multi-Stream Collections: multiXview at the CASTLE 2025 Interactive Grand Challenge. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3746027.3760243

*Equal contribution to this research.

MM '25, Dublin, Ireland

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3760243

1 Introduction

Today's video landscape is characterized by an explosion of multi-camera capture scenarios where the same events are simultaneously recorded from multiple perspectives. From surveillance networks to monitoring public spaces, and from multi-angle sports broadcasts to collaborative work environments and social gatherings documented across multiple devices by multiple participants, contemporary video collections increasingly feature synchronized streams that capture overlapping information about shared events and spaces. These multi-stream collections have found applications across various multimedia domains, including video analysis [42], multi-view event reconstruction [31], and video summarization [5].

The need for interactive search and exploration in these multi-stream environments is driven by the fundamental challenge that single events are now captured simultaneously across multiple video and multimedia streams, each providing partial, overlapping, or complementary information. Users must navigate not just temporal segments within videos, but also across multiple synchronized streams to piece together the full context of an event. The traditional goal of "finding the right video, event, or item of interest" transforms into the more complex task of "retrieving and aggregating information spread across multiple incomplete sources."

Existing interactive video search and exploration systems were developed during an era when video content was primarily captured by single cameras and curated for broad consumption. These well-established paradigms, optimized for collections with high topical, stylistic, and visual diversity, do not immediately generalize to multi-stream scenarios. The techniques that excel in established benchmarks, such as TRECVID [2], MediaEval [14], VBS [25, 33], and LSC [12]—which typically leverage discriminative visual features to arrive at relevant results rapidly—struggle when faced with the redundancy and fragmentation characteristic of synchronized multi-stream collections.

The widespread adoption of mobile and wearable cameras has created an explosion of first-person recordings of daily activities and environments. Unlike the curated, professionally-produced, multi-camera videos that populate traditional benchmarks, *egocentric* videos present a fundamentally different retrieval scenario. Egocentric videos are captured from a first-person perspective, showing the world as seen through the eyes of the camera wearer or actor, typically providing an immersive viewpoint of their direct interactions and immediate surroundings. In contrast, *exocentric* videos are recorded from a third-person perspective using external cameras positioned to observe the scene from outside, offering a broader contextual view of the environment and the subject's actions within it. The limited field of view inherent to wearable cameras provides only a partial view of any given scene or event. Consequently, critical contextual information—other participants, off-screen activities, or simultaneous events—may be scattered across multiple streams, requiring cross-video analysis to reconstruct the entire narrative.

Despite the growing prevalence of egocentric visual media in applications ranging from lifelogging [11] to workplace monitoring, interactive search and exploration methods designed for multi-stream scenarios remain significantly understudied. Existing video retrieval systems, optimized for diverse collections with comprehensive per-video coverage, struggle when faced with the redundancy and fragmentation characteristic of synchronized multi-stream collections.

In this work, we address this gap by presenting an interactive system designed explicitly for searching and exploring collections where multiple video streams capture overlapping information with varying degrees of redundancy. We extend Exquisitor [18–20, 35, 36], an established interactive visual search and exploration framework, with novel capabilities for handling multimodal egocentric and exocentric video collections. Our approach enables users to efficiently navigate the complex relationships between multiple partial information sources, leveraging both the redundancy and complementarity inherent in such collections.

Our system is designed for synchronized multi-stream video collections, such as the *CASTLE 2024* dataset [28], a new benchmark collection comprising over 600 hours of synchronized egocentric and exocentric video. This dataset, captured by multiple participants wearing cameras alongside fixed environmental cameras, exemplifies the retrieval challenges we address: finding and reconstructing events distributed across multiple partial views with varying degrees of overlap and complementarity. The dataset comprises multimodal content from 15 time-aligned sources, including the perspectives of 10 participants equipped with wearable cameras and 5 fixed cameras providing exocentric viewpoints, all recorded over four consecutive days in a fixed location without any partial censoring, such as blurred faces or distorted audio.

In this work, we develop a new interactive retrieval system *multiXview*, using Exquisitor as the foundation, with novel visualization and interaction capabilities specifically designed for complex multi-stream video collections, combining exocentric and egocentric streams to enable effective search and exploration within the challenging CASTLE 2024 dataset context. At a conceptual level, our contributions are as follows:

- *Multimodal Information Retrieval*: A unified search architecture that integrates the usage of multiple indexed information sources, i.e., speech-to-text transcripts, scene descriptions, and metadata.
- *Multi-Stream Player Interface*: An interface that enables parallel examination of multiple temporally-synchronized video streams, facilitating comprehensive event discovery and information retrieval across concurrent data sources. A location-aware visualization system enables coarse-grained participant positioning within the examination area, allowing for spatial filtering to reduce irrelevant content and enhance search efficiency.
- *Temporal Navigation Framework*: Timeline-based visualization that presents temporally-ordered views of heterogeneous streams (participant videos, fixed cameras) in a structured layout, enabling temporal scoping and faceted search capabilities without explicit query formulation.

By seamlessly merging dynamic visualizations with responsive video navigation, these enhancements deliver a fluid, feature-rich environment for exploring and pinpointing relevant content. The system's actual impact will be evaluated at the CASTLE 2025 Grand Challenge Interactive Track, where we will measure how effectively these integrated features accelerate discovery and decision-making.

2 Related Work

Over the past few decades, many successful approaches to (interactive) video search and exploration have been proposed. Some of the best-known examples from the early days of content-based image and video retrieval include Informedia [39] and MARS [32]. At that time, a typical method of analyzing and indexing video content for retrieval involved performing shot-boundary detection or uniform video frame sampling, followed by the extraction of low-level visual features, such as color, texture, and edge descriptors. Gradually, the quality of video representation has significantly improved to include, e.g., semantic concepts [30, 37, 41] and then, especially with the advent of deep learning, actions [7] and events [9, 13], facilitating increasingly accurate keyword-based video search. More recently, the rapid expansion of foundation models, such as VLMs [23, 26], has brought about a new paradigm shift, enabling effective video retrieval using natural language queries.

Alternatively, metadata related to videos, such as tags, name, creation time, and user description, can be used to find videos through keyword/text search or faceted browsing; however, the results will be full videos, not specific moments within the video. These techniques are not mutually exclusive, and many analytical multimedia applications utilize both. In addition, video summarizations, temporal querying, relevance feedback, and object-localization queries may also be present in such systems to further enhance the user's ability to interactively search and explore the video collection [29, 34, 36, 38].

The techniques used within a standard video retrieval system can be applied to egocentric video collections [1, 19, 21, 21]. However, the different perspective of the egocentric view is typically not present in large deep learning training datasets, or it is only represented in a small portion, such as sport activity videos. Thus,

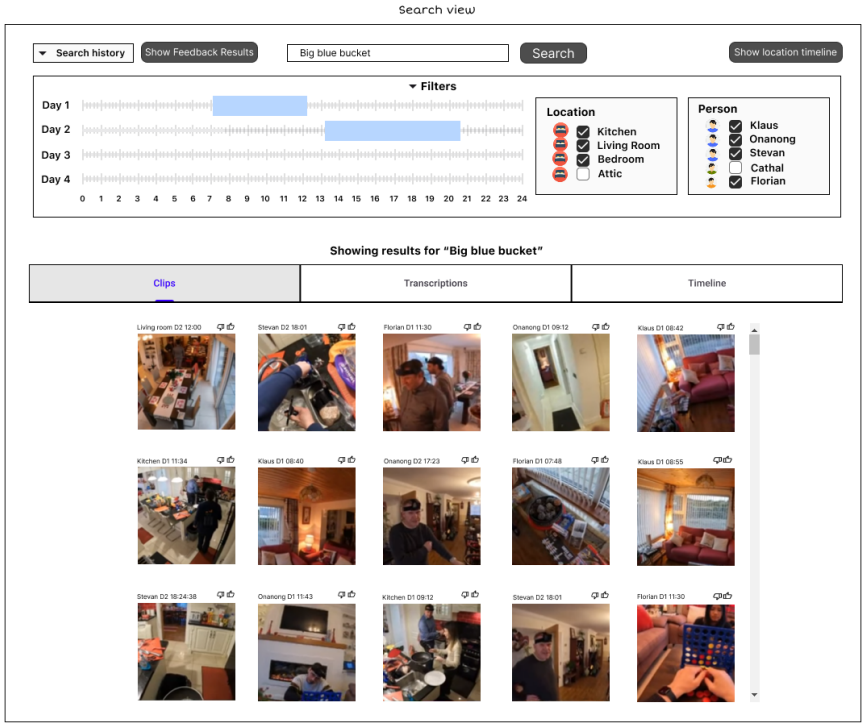

Figure 1: Mockup of the "Search view", the main view of the system. From top to bottom, it includes a search bar, a set of search filters, and a tabbed interface that presents the search results.

these models can fail to interpret actions from an egocentric perspective, while visual querying and non-action-related queries will still perform well. It is possible to use a specific model trained with egocentric data to better support queries related to action recognition [3, 4, 15] and person and object detection [22].

Multi-stream video collections consist of simultaneous streams captured by several cameras that share, and often overlap, the same field of view. A primary application for such collections is in surveillance, where fixed cameras capture viewpoints of an area (indoors and/or outdoors), with interfaces that allow operators to monitor activity or, at larger scales, run analytics that identify people, search for events, or track objects such as vehicles [6, 16, 24]. To enable efficient event retrieval, techniques such as video summarization [5, 8] and fused representations of the multi-stream space [40] are used.

Retrieval systems built for purely exocentric or egocentric footage assume a single viewpoint. They perform well within a single stream but struggle to merge perspectives or align events that co-occur in different sources. When a conversation is recorded simultaneously by a body-worn and a fixed camera, general retrieval systems often fail to link the clips or leverage their complementary content. These types of events are evident in the CASTLE 2024 collection, where many events can be observed in both egocentric and exocentric footage. Overcoming this challenge requires retrieval systems that can synchronize, correlate, and reason across multiple timelines and viewpoints.

3 multiXview

This section starts with a concise overview of Exquisitor, which serves as the foundation for multiXview, and then details the model and UI enhancements required for multi-video-stream support.

3.1 Exquisitor

Exquisitor is a comprehensive interactive search and exploration system designed for multimedia collections. The system requires data preprocessing, indexing, and analysis to transform the target multimedia collection to a form amenable to fast retrieval [35].

In the preprocessing and indexing phase, Exquisitor segments video collections using shot-boundary detection or fixed-length intervals, extracting keyframes from each segment. These keyframes are processed into semantic vector representations using deep learning models from the CLIP family to generate visual embeddings that capture static visual properties. Concurrently, the system generates event-level captions at 20-second intervals and converts them into text embeddings that capture long-range temporal context. These complementary multi-modal embeddings are indexed using approximate high-dimensional indexes to enable fast retrieval.

The online search phase supports natural language queries across modalities with optional late fusion for result combination. Query processing involves reformulation and expansion using Large Language Models, while interactive refinement is facilitated through relevance feedback mechanisms. Users can apply metadata-based filters or exclude specific media groups during search.

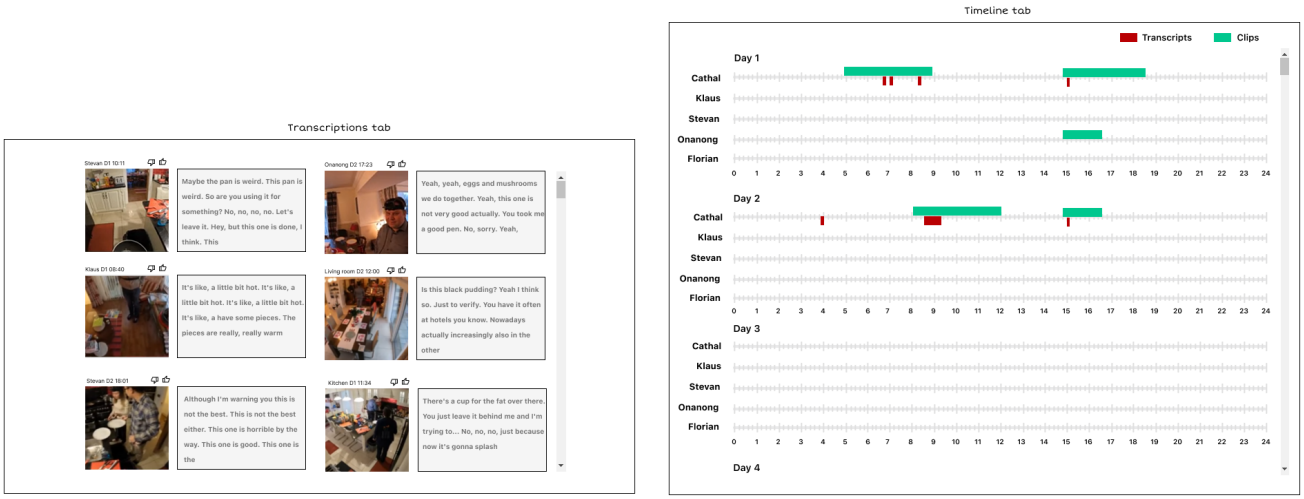

Figure 2: Mockups of the "Transcriptions" and "Timeline" tabs that are part of the "Search view".

The Exquisitor interface features a streamlined design with a search column and grid-based result view. Retrieved items are presented in ranked order, with media belonging to the same group (video keyframes, album images, or temporal sequences) displayed as rows to facilitate multi-event assessment. A left-side filter panel provides access to search constraints, while a summary view accessible through item selection enables group-level exclusion and detailed examination.

3.2 Search and Exploration

Several modalities and types of data can be extracted from a multi-video stream dataset. Beyond the visual modality, videos have audio from which transcripts can be obtained, typically with a model such as Whisper [27], and metadata that can be useful for filtering the search space. To enhance search and exploration, indexes are used to allow content-based and text-based retrieval, with filters that affect the search space. The filters can also be used to facilitate faceted search by browsing the collection and narrowing it by the selected filters.

3.2.1 Multi-Index Search. For the CASTLE dataset, we extract the visual features using the preprocessing and analysis approach of Exquisitor. Furthermore, since the CASTLE dataset contains transcript data, text features are also extracted. As such, each extracted feature representation will be held in its own high-dimensional ANN index. Given the size of the dataset, we opt to use a disk-based variant of the eCP index [10, 17] to reduce the memory footprint of less frequently used indexes. An added benefit of this index is that it is straightforward to extend it to support incremental retrieval, a crucial factor when restrictive filters are applied.

3.2.2 Metadata Filters. For constrained environment datasets, metadata is often enriched by knowledge beyond generic tags or video categories. They may contain information about the people present, such as schedules (e.g., a concert with performance times), office environments with records of when someone entered or exited, and so on. In the CASTLE dataset, the videos are captured within a constrained environment, specifically a house with a fixed layout. It incorporates the viewpoints of all participants and fixed room cameras, enabling the mapping of people to rooms based on visual modality. It even contains GPS data of the participants, which may be too coarse-grained to pinpoint them to the exact room, but can be helpful on rare occasions when they venture outside the house. We opt to use the following metadata for filtering and faceted search based on time and location: *day, time, location feeds, person feeds*. This allows users to quickly narrow down the search space if their task relates to a specific day, time period, location, or person.

3.3 User Interface

To support searches across modalities, the UI must clearly reflect each modality's results. Filters become central to navigating CASTLE's complexity, so they must be instantly understandable and easy to apply. The results from a search can convey different information depending on the visualization. For example, although a visual query is used, it may be more convenient to determine which locations or timespans the results cover. Finally, because CASTLE videos often require side-by-side viewpoint comparisons, we must augment the media summary and player view in multiXview with a new synchronized multi-stream viewer. These enhancements will transform it into a truly multi-stream search and exploration tool.

3.3.1 Search view. Figure 1 shows the interface design of the main view of our system, the "Search view". This view allows the user to enter a text query in the search bar at the top middle of the screen, select day and time filters by dragging a span across the desired area, and choose which location and person feeds should be included in the search. The search results are displayed below the filters box, which can be collapsed to provide more space for viewing results. The "Clips" tab displays the visual modality results, featuring keyframes of video segments. Above the image, there is basic metadata that states the camera feed (location or person) it is from, along with the day and time. To the right of the metadata information, there are the "thumbs-up" and "thumbs-down" buttons

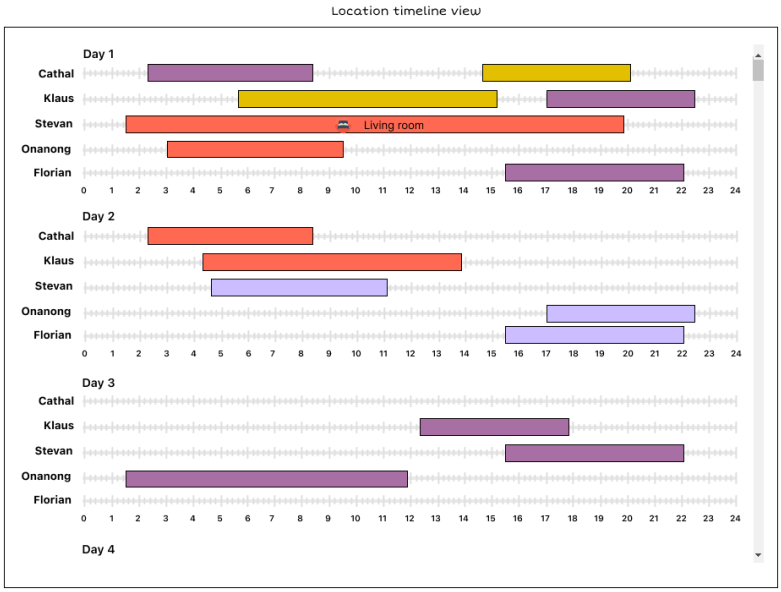

Figure 3: Mockup of the "Location timeline view". This view displays information regarding the location (color-coded) of each participant over the four days.

to perform relevance feedback if desired. The "Show Feedback Results" button, located to the left of the search bar, opens a pop-up box that displays the results. The search history drop-down (located at the top left of the view) can be used to navigate to results from previous searches. The top-right button opens the "Location timeline view" described below.

Figure 2 depicts the results displayed in the "Transcriptions" and "Timeline" tab from the "Search View". The results for transcripts display the keyframe closest to the corresponding transcript's time, along with the transcript text. The results for the timeline displays the streams from which the results are from. It groups the days from top to bottom, with each row in the grouping relating to a specific feed, indicated by the name on the left, with a 24-hour timeline next to the name. Green boxes in the timeline indicate results from the visual domain (clips tab), and the brown boxes indicate results from transcripts. This tab enables users to quickly identify which individuals or locations may be relevant to the task.

3.3.2 Location timeline view. Figure 3 depicts the "Location timeline view", which conveys results based on which locations a person was in during the day. Similarly to the "Timeline" tab, it groups the days. The rows in this view only show people and not videos that include both people and locations. The different colored spans indicate the location of the people based on the results of a query or full coverage, where the filters act as a faceted search. The name of the location is displayed when hovered.

3.3.3 Player view. The "Player view", depicted in Figure 4, opens whenever the user clicks a box in either timeline view or selects a keyframe in another tab. Its goal is to present rich contextual information about the chosen keyframe or time span. By default, it displays four video streams, with the selected feed in the top-left position. At the top left of the view is a global playback control that allows users to play or pause all videos, jump forward or backward by ten seconds, and adjust the playback speed. Each feed bears its name (person or location) in a header (for top-row videos) or footer (for bottom-row videos). Within that header/footer are five buttons:

(1) Transcript — Opens the feed's transcript in a pop-up box.
(2) Select — Allows the user to draw a selection box on a frame and run a visual-similarity search.
(3) Keyframes — Opens a scrollable pop-up gallery of all keyframes in the feed, enabling rapid navigation through its content.
(4) Mute — Toggle sound on and off.
(5) Fullscreen — View the feed in full screen.

The "Locations" panel, located to the left of the video grid, lists each filtered participant and their current location. A green dot beside a name indicates a live feed, while a red dot indicates that their camera feed is off. The eye icon marks which streams are actively playing. Above the videos, the current date and time are displayed. In the top-left corner, a Bounding Boxes toggle lets the interacting user overlay name labels around detected people. Beside the switch is a drop-down for choosing to view more or fewer streams simultaneously than the default. Below the grid, a timeline displays an aggregate version of the "Timeline" tab results (Figure 2), allowing users to jump to any point in time across all streams.

4 Use Case Scenarios

In this section, we outline potential workflows with the new interface for some hypothetical queries to the CASTLE 2024 dataset.

Query: Find a partially eaten apple. In the "Search view" (Figure 1), the user writes "eaten apple" in the search field and clicks the respective "Search" button. A grid of results appears on the

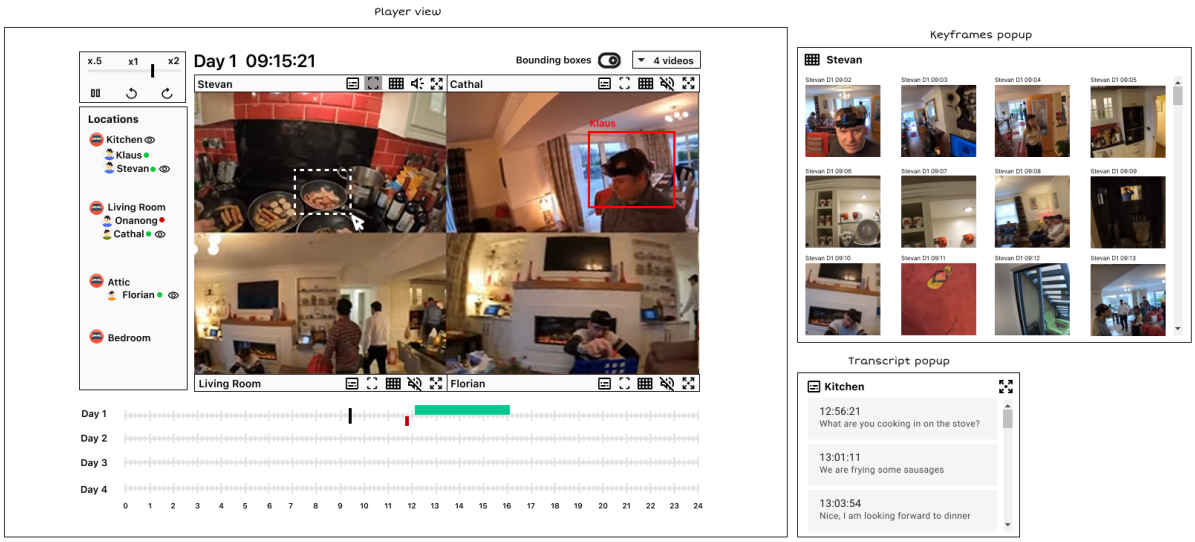

Figure 4: Mockup of the "Player view". It includes a video player capable of playing multiple video streams, video controls for the video player, and information regarding the location of people and the status of their streams.

"Clips" tab at the bottom of the "Search view". The user visually skims through the clips to find an instance of a partially eaten apple. If irrelevant clips are shown, the user can use the relevance feedback to influence the search process by clicking the "thumbs up" and "thumbs down" icons displayed alongside each clip. If the user finds an instance of an uneaten apple, they click the clip and are taken to the "Player view" (Figure 4). Here, the user views the video stream associated with the search result to discover when the apple is eaten, at what time, and by whom. The user can speed up and slow down the video using the video player controls.

Query: What was the ingredient not available (i.e., also not replaced) for the pumpkin risotto? The user selects only the kitchen in the filters and searches for "ingredient missing in risotto" using the "Search view". The user navigates to the "Transcriptions" (Figure 2) tab at the bottom of the "Search view". The user skims through the search results to find mentions of missing ingredients. When more context is needed to understand the transcripts, the user clicks on a particular search result and is directed to the "Player view" where they can inspect the transcripts in more detail by clicking on the respective icon in the video title bar. Alternatively, the user can watch the video stream with sound to better understand the clip's context and answer the query.

Query: What book did Florian spend the longest time reading? The user selects only Florian in the filters and searches for "book" using the "Search view". The user navigates to the "Timeline" tab (Figure 2) and sees at what times Florian's stream included a book. The user then clicks each of these times on the timeline, opening the "Player view" and then inspects the book Florian was reading. Having inspected all blocks of the timeline in which Florian was reading, the user can answer which book Florian read the most.

Query: Find winning hands of poker. The user searches for "cards" using the "Search view" and navigates to the "Timeline" tab to gain an understanding of how many times the person has played card games. Each game session is highlighted on the timeline of the card players as a consecutive block. The user then selects the first block on the timeline of one of the card players which opens the "Player view". The user clicks the different locations and people's names on the "Location" panel to view the respective video stream on the video player. Through this process, the user learns the placement of the poker players in the house. Having found the names of the poker players, the user drags their names from the "Location" panel into different sections of the video player to simultaneously view their egocentric video streams. The user then watches the poker game to find the winning poker hands. If multiple sessions of poker have been played, the user can return to the "Search view" to continue the search process by selecting another point on the "Timeline" tab associated with the query.

5 Conclusion

Unifying egocentric and exocentric feeds in a single interactive environment enables users to reconstruct fragmented events across multiple perspectives. By integrating multimodal search, spatially-aware filtering and browsing, and synchronized multi-stream views, multiXview aims to improve search and exploration of multi-stream collections such as CASTLE 2024. Overall performance will be assessed on-site at the CASTLE 2025 Grand Challenge Interactive Track, organized as part of the ACM Multimedia 2025 conference.

Acknowledgments

This work was partially supported by Icelandic Research Fund grant 239772-051.

References

[1] Ahmed Alateeq, Mark Roantree, and Cathal Gurrin. 2024. Voxento-Pro: An Advanced Voice Lifelog Retrieval Interaction for Multimodal Lifelogs. In *Proceedings of the 7th Annual ACM Workshop on the Lifelog Search Challenge*. 105–110.

[2] George Awad, Jonathan Fiscus, Afzal Godil, Lukas Diduch, Yvette Graham, and Georges Quénot. 2024. TRECVID 2024 - Evaluating video search, captioning, and activity recognition. In *Proceedings of TRECVID 2024*. NIST, USA.

[3] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. 2018. Scaling Egocentric Vision: The EPIC-KITCHENS Dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[4] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. 2022. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision* (2022), 1–23.

[5] Mohamed Elfeki, Liqiang Wang, and Ali Borji. 2022. Multi-stream dynamic video summarization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 339–349.

[6] Ching-Tang Fan, Yuan-Kai Wang, and Cai-Ren Huang. 2016. Heterogeneous information fusion and visualization for a large-scale intelligent video surveillance system. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 47, 4 (2016), 593–604.

[7] Basura Fernando, Efstratios Gavves, Jose M. Oramas, Amir Ghodrati, and Tinne Tuytelaars. 2015. Modeling Video Evolution for Action Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[8] Yanwei Fu, Yanwen Guo, Yanshu Zhu, Feng Liu, Chuanming Song, and Zhi-Hua Zhou. 2010. Multi-view video summarization. *IEEE Transactions on Multimedia* 12, 7 (2010), 717–729.

[9] Chuang Gan, Naiyan Wang, Yi Yang, Dit-Yan Yeung, and Alex G. Hauptmann. 2015. DevNet: A Deep Event Network for Multimedia Event Detection and Evidence Recounting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[10] Gylfi Þór Gudmundsson, Björn Þór Jónsson, and Laurent Amsaleg. 2010. A large-scale performance study of cluster-based high-dimensional indexing. In *Proceedings of the international workshop on Very-large-scale multimedia corpus, mining and retrieval*. 31–36.

[11] Cathal Gurrin, Alan F Smeaton, Aiden R Doherty, et al. 2014. Lifelogging: Personal big data. *Foundations and Trends® in information retrieval* 8, 1 (2014), 1–125.

[12] Cathal Gurrin, Liting Zhou, Graham Healy, Werner Bailer, Duc-Tien Dang Nguyen, Steve Hodges, Björn Þór Jónsson, Jakub Lokoč, Luca Rossetto, Minh-Triet Tran, and Klaus Schöffmann. 2024. Introduction to the Seventh Annual Lifelog Search Challenge, LSC'24. In *Proceedings of the 2024 International Conference on Multimedia Retrieval (ICMR '24)*. ACM, 1334–1335.

[13] Amirhossein Habibian, Thomas Mensink, and Cees G.M. Snoek. 2014. VideoStory: A New Multimedia Embedding for Few-Example Recognition and Translation of Events. In *Proceedings of the 22nd ACM International Conference on Multimedia (MM '14)*. ACM, 17–26.

[14] Steven Hicks, Andreas Lommatzsch, Ali Hürriyetoglu, Romain Vuillemot, Mihai Gabriel Constantin, Vajira Thambawita, and Martha A. Larson (Eds.). 2024. *Working Notes Proceedings of the MediaEval 2023 Workshop, Amsterdam, The Netherlands and Online, 1-2 February 2024*. CEUR Workshop Proceedings, Vol. 3658. CEUR-WS.org. https://ceur-ws.org/Vol-3658

[15] Thomas Hummel, Shyamgopal Karthik, Mariana-Iuliana Georgescu, and Zeynep Akata. 2024. Egocvr: An egocentric benchmark for fine-grained composed video retrieval. In *European Conference on Computer Vision*. Springer, 1–17.

[16] Muhammad Attique Khan, Kashif Javed, Sajid Ali Khan, Tanzila Saba, Usman Habib, Junaid Ali Khan, and Aaqif Afzaal Abbasi. 2024. Human action recognition using fusion of multiview and deep features: an application to video surveillance. *Multimedia Tools and Applications* 83, 5 (2024), 14885–14911.

[17] Omar Shahbaz Khan, Gylfi Þór Guðmundsson, and Björn Þór Jónsson. 2025. The Curious Case of High-Dimensional Indexing as a File Structure: A Case Study of eCP-FS. *arXiv preprint arXiv:2507.21939* (2025).

[18] Omar Shahbaz Khan, Björn Þór Jónsson, Stevan Rudinac, Jan Zahálka, Hanna Ragnarsdóttir, Þórhildur Þorleiksdóttir, Gylfi Þór Guðmundsson, Laurent Amsaleg, and Marcel Worring. 2020. Interactive Learning for Multimedia at Large. In *Proc. European Conference on IR (ECIR)*.

[19] Omar Shahbaz Khan, Ujjwal Sharma, Hongyi Zhu, Stevan Rudinac, and Björn Þór Jónsson. 2024. Exquisitor at the Lifelog Search Challenge 2024: Blending Conversational Search with User Relevance Feedback. In *Proc. of the 7th Annual ACM Workshop on the Lifelog Search Challenge (LSC '24)*. ACM, 5 pages.

[20] Omar Shahbaz Khan, Hongyi Zhu, Ujjwal Sharma, Evangelos Kanoulas, Stevan Rudinac, and Björn Þór Jónsson. 2024. Exquisitor at the Video Browser Showdown 2024: Relevance Feedback Meets Conversational Search. In *MultiMedia Modeling*. Springer Nature Switzerland, 347–355.

[21] Hoang-Bao Le, Thao-Nhu Nguyen, Tu-Khiem Le, Minh-Triet Tran, Thanh-Binh Nguyen, Van-Tu Ninh, Liting Zhou, and Cathal Gurrin. 2024. LifeSeeker 6.0: Leveraging the linguistic aspect of the lifelog system in LSC'24. In *Proceedings of the 7th Annual ACM Workshop on the Lifelog Search Challenge*. 53–57.

[22] Yong Jae Lee, Joydeep Ghosh, and Kristen Grauman. 2012. Discovering important people and objects for egocentric video summarization. In *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 1346–1353.

[23] 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. In *Proceedings of the 40th International Conference on Machine Learning (ICML'23)*. JMLR.org, Article 814, 13 pages.

[24] Yang Liu, Dingkang Yang, Gaoyun Fang, Yuzheng Wang, Donglai Wei, Mengyang Zhao, Kai Cheng, Jing Liu, and Liang Song. 2023. Stochastic video normality network for abnormal event detection in surveillance videos. *Knowledge-Based Systems* 280 (2023), 110986.

[25] Jakub Lokoč, Stelios Andreadis, Werner Bailer, Aaron Duane, Cathal Gurrin, Zhixin Ma, Nicola Messina, Thao-Nhu Nguyen, Ladislav Peška, Luca Rossetto, et al. 2023. Interactive video retrieval in the age of effective joint embedding deep models: lessons from the 11th VBS. *Multimedia Systems* (2023), 1–24.

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning*. PMLR, 8748–8763.

[27] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*. PMLR, 28492–28518.

[28] Luca Rossetto, Werner Bailer, Duc-Tien Dang-Nguyen, Graham Healy, Björn Þór Jónsson, Onanong Kongmeesub, Hoang-Bao Le, Stevan Rudinac, Klaus Schöffmann, Florian Spiess, et al. 2025. The CASTLE 2024 Dataset: Advancing the Art of Multimodal Understanding. *arXiv preprint arXiv:2503.17116* (2025).

[29] Luca Rossetto and Ralph Gasser. 2025. Feature-Driven Video Segmentation and Advanced Querying with vitrivr-Engine. In *MultiMedia Modeling - 31st International Conference on Multimedia Modeling, MMM 2025, Nara, Japan, January 8-10, 2025, Proceedings, Part V (Lecture Notes in Computer Science, Vol. 15524)*. Springer, 272–277.

[30] Stevan Rudinac, Martha Larson, and Alan Hanjalic. 2012. Leveraging visual concepts and query performance prediction for semantic-theme-based video retrieval. *International Journal of Multimedia Information Retrieval* 1, 4 (2012), 263–280.

[31] Viktor Rudnev, Gereon Fox, Mohamed Elgharib, Christian Theobalt, and Vladislav Golyanik. 2025. Dynamic EventNeRF: Reconstructing General Dynamic Scenes from Multi-view RGB and Event Streams. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR) Workshops*. 4866–4876.

[32] Yong Rui, Thomas S. Huang, and Sharad Mehrotra. 1997. Content-based image retrieval with relevance feedback in MARS. *Proceedings of International Conference on Image Processing (ICIP)* (1997).

[33] Klaus Schoeffmann. 2019. Video browser showdown 2012-2019: A review. In *2019 International conference on content-based multimedia indexing (CBMI)*. IEEE, 1–4.

[34] Klaus Schoeffmann and Mario Leopold. 2025. AI-based Video Content Understanding for Automatic and Interactive Multimedia Retrieval. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 3750–3758.

[35] Ujjwal Sharma, Omar Shahbaz Khan, Stevan Rudinac, and Björn Þór Jónsson. 2025. Can Relevance Feedback, Conversational Search and Foundation Models Work Together for Interactive Video Search and Exploration?. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 3740–3749.

[36] Ujjwal Sharma, Omar Shahbaz Khan, Stevan Rudinac, and Björn Þór Jónsson. 2025. Exquisitor at the Video Browser Showdown 2025: Unifying Conversational Search and User Relevance Feedback. In *International Conference on Multimedia Modeling*. Springer, 264–271.

[37] Cees Snoek, Kvd Sande, OD Rooij, Bouke Huurnink, J Uijlings, M v Liempt, M Bugalhoy, I Trancosoy, F Yan, M Tahir, et al. 2009. The MediaMill TRECVID 2009 semantic video search engine. In *TRECVID workshop*.

[38] Michael Stroh, Vojtěch Kloda, Benjamin Verner, Zuzana Vopálková, Raphael Buchmüller, Bastian Jäckl, Jakub Hajko, and Jakub Lokoč. 2025. PraK Tool V3: Enhancing Video Item Search Using Localized Text and Texture Queries. In *MultiMedia Modeling: 31st International Conference on Multimedia Modeling, MMM 2025, Nara, Japan, January 8–10, 2025, Proceedings, Part V*. Springer-Verlag, 326–333.

[39] H.D. Wactlar, T. Kanade, M.A. Smith, and S.M. Stevens. 1996. Intelligent access to digital video: Informedia project. *Computer* 29, 5 (1996), 46–52.

[40] Chenggang Yan, Biao Gong, Yuxuan Wei, and Yue Gao. 2020. Deep multi-view enhancement hashing for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 4 (2020), 1445–1451.

[41] Jun Yang, Yu-Gang Jiang, Alexander G. Hauptmann, and Chong-Wah Ngo. 2007. Evaluating bag-of-visual-words representations in scene classification. In *Proceedings of the International Workshop on Workshop on Multimedia Information Retrieval (MIR '07)*. ACM, 197–206.

[42] Zhenguo Yang, Qing Li, Wenyin Liu, and Jianming Lv. 2020. Shared Multi-View Data Representation for Multi-Domain Event Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42, 5 (2020), 1243–1256.