Extending Lifelog Retrieval to Multi-stream Video Retrieval at the CASTLE Challenge 2025

Quang-Linh Tran
quang-linh.tran2@mail.dcu.ie
ADAPT Centre, School of Computing,
Dublin City University
Dublin, Ireland

Hoang-Bao Le
bao.le2@mail.dcu.ie
Dublin City University
Dublin, Ireland

Thang-Long Nguyen Ho
thanglong.nguyenho27@mail.dcu.ie
Dublin City University
Dublin, Ireland

Graham Healy
graham.healy@dcu.ie
Dublin City University
Dublin, Ireland

Liting Zhou
liting.zhou@dcu.ie
Dublin City University
Dublin, Ireland

Allie Tran
allie.tran@dcu.ie
School of Computing, Dublin City
University
Dublin, Ireland

Abstract

We present the DCU team's system for the CASTLE Challenge at ACM Multimedia 2025, which explores video retrieval and question answering in egocentric, multi-user environments. Our system adapts techniques developed for lifelogging, particularly event-based semantic retrieval and QA pipelines, to the CASTLE dataset with minimal architectural changes. It combines vision-language embeddings, transcript-based retrieval, and person tracking to support both automatic and interactive search workflows. In the interactive track, we introduce a modular interface for narrative reconstruction and exploratory search. Qualitative results show that the system can generate plausible, evidence-based answers to complex multimodal queries. These findings suggest that lifelog retrieval systems offer a viable foundation for broader egocentric video analysis.

CCS Concepts

• **Information systems → Users and interactive retrieval**.

Keywords

Egocentric Video Retrieval, Interactive Retrieval System, Multimedia

ACM Reference Format:
Quang-Linh Tran, Hoang-Bao Le, Thang-Long Nguyen Ho, Graham Healy, Liting Zhou, and Allie Tran. 2025. Extending Lifelog Retrieval to Multi-stream Video Retrieval at the CASTLE Challenge 2025. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3746027.3760244

1 Introduction

The CASTLE Challenge at ACM Multimedia 2025 presents a novel benchmark for evaluating interactive and automatic video retrieval systems on egocentric video data collected from multiple subjects in a shared environment. Unlike conventional video retrieval tasks, CASTLE [15] requires systems to reason across long temporal spans, handle multimodal signals, and support naturalistic, user-driven search patterns. This paper presents the DCU team's contribution to the challenge, which adapts and extends lifelogging [5] retrieval methods to the CASTLE dataset using lightweight but robust techniques for multimodal retrieval, interaction design, and question answering.

Our approach builds on the retrieval-first paradigm of systems which were originally developed for lifelogging scenarios with wearable cameras [18]. We evaluate their transferability to CASTLE's egocentric video setting by reusing key components such as event grouping, visual-semantic retrieval, and question answering workflows. We also introduce a modular interactive interface to support exploratory analysis during the live interactive track. The system supports object and event retrieval, question answering, and analytics through a unified backend powered by modern embedding models and search infrastructure.

We describe our approaches to both the fully automatic and interactive tracks and evaluate their effectiveness through case studies and qualitative analysis. Our system was able to produce plausible and evidence-based answers to challenging multimodal queries, despite the lack of ground truth answers. The findings support our hypothesis that lifelog QA systems can be meaningfully reused in egocentric settings, with minimal architectural changes.

2 Related Works

2.1 Multimodal Retrieval

Recent vision–language models (VLMs) have transformed cross-modal retrieval by embedding images and text in shared semantic spaces. CLIP [13] introduced contrastive learning for zero-shot tasks, followed by SigLIP [25] and SigLIP 2 [21], which enhanced alignment via sigmoid loss, decoder pretraining, and multilingual data—while remaining compatible with CLIP pipelines.

BLIP and BLIP-2 [7, 8] adopted a retrieval-first approach combining captioning, matching, and grounding. BLIP-2 links frozen vision encoders to LLMs via a trainable Q-Former, outperforming larger models like Flamingo-80B [1] across retrieval and captioning

MM '25, Dublin, Ireland

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3760244

benchmarks. Our CASTLE QA pipeline uses SigLIP and BLIP-2 to encode frames and queries.

Modern video retrieval systems now emphasize multimodal fusion. VISIONE blends CLIP, CLIP2Video [4], and ALADIN [10] via hybrid indexing; vitrivr [16] and Verge [12] integrate multilingual and timeline-aware features. These trends reflect a shift toward interactive, semantically grounded pipelines that support exploratory and narrative queries.

Our work follows this trajectory, adapting state-of-the-art VLMs and indexing methods to the challenges of egocentric, multi-user video data in CASTLE.

2.2 From Lifelogging to Egocentric Video QA

Lifelogging systems [5] aim to retrieve personal moments from continuous first-person data. The Lifelog Search Challenge (LSC)[6] has advanced QA pipelines that support spatial, temporal, and contextual reasoning. Systems like MyEachtraX[19], MemoriEase[20], and T-EAGLE[11] combine event segmentation, VLMs, and LLM-based query interpretation to support natural language search.

CASTLE shares lifelogging's continuous, egocentric nature but adds multi-user complexity: overlapping interactions, shared environments, and social entanglement. Our system examines how lifelog QA methods can be extended to support narrative and analytical queries in this more dynamic, collective context.

3 Fully-Automatic Track

This section outlines the distinct approaches used to solve the three sub-tasks, including Event Instance Search, Object Instance Search, and Question Answering, in the automatic track of the CASTLE Challenge. Despite considerable overlap in techniques, each task is addressed independently, using tailored pipelines optimised for its specific requirements.

3.1 Event Instance Search

From the CASTLE dataset of over 600 hours of videos, we extract a static image for each second in the videos. There are some images inserted to replace undesired published content, so we use visual similarity to remove them. The total number of keyframes extracted from the dataset is 2,067,676 images.

We use the BLIP2 [7] embedding model trained on the COCO dataset [9] to extract the visual embedding of images. The embedding size of the images is 768. We also tested on two other variants of this model, including pretrained and pretrained_ViT, but the match of query and images is low, so we chose the model trained on the COCO version.

After the data processing, we index the data to Elasticsearch[1]. The indexed data includes ImageID (ID of representative images), VideoID (ID includes name of owner, day, and hour), owner, day, hour, start time, end time, transcript between start and end time, and BLIP2 embedding vector. The vector is enabled for search with cosine similarity.

For each query of the Event Instance Search task, we select the best image as an anchor image to search for the whole event of the corresponding event. Based on the information of the anchor image (owner, day), we filter all the images on the same day, excluding those from fixed cameras. Then, for each person's POV data, we use the aforementioned embeddings to calculate the similarity and use an adjustable sliding window to group all the images with a high ranking compared to the anchor image.

3.2 Object Instance Search

We use the same set of keyframes from the Event Instance search task, but to avoid the large amount of data for indexing, we group keyframe images into events by comparing visual similarity from BLIP2 embedding. For each video, we group keyframes by comparing the cosine similarity of consecutive keyframes and choosing a threshold of 0.9 to split events. After the grouping, we obtain 204034 events and choose the first image in the event as the representative image. The start time of the event is the time of the first image, and the end time is the time of the last image. We index event segmentation into an Elasticsearch index and perform a search on embeddings.

For each query in the Object Instance Search sub-task, we pass the query to the BLIP2 embedding model to get the query embedding and search on Elasticsearch to find the top 100 relevant keyframes. We also try to use Sora to generate an image for each query and search on image embedding, but the result is not sufficiently relevant. This is because the objects that Sora generated are big and centered in the image, while in the dataset, they are only small objects and placed around a lot of other objects. The results and some special cases are presented in the section 5.2.

3.3 Question Answering

For the Question Answering (QA) sub-task, we adapt the MyEachtraX pipeline [19] with minimal changes in order to assess its generalisability to this dataset. To maintain architectural continuity, we continue to use SigLIP [25] as the embedding model for both visual frames and text queries. Event-level grouping is also preserved from the original system, allowing us to retrieve semantically coherent lifelog segments in response to natural language questions.

To handle questions involving people, such as 'Who was I with?' or 'What was person X doing?', we introduce a face-tracking module. First, people are detected in keyframes using YOLOv8 [22] and tracked over time using Deep SORT [24]. We cluster detected face crops and use the first hour of the dataset (8–9 am on Day 1) as training data for a FaceNet [17] model, which enables consistent person identification across the full video timeline.

For questions grounded in spoken language or audio context, we transcribe the videos using Whisper [14], applying a 30-second sliding window with a 10-second stride. These transcript windows are embedded using MiniLM [23], enabling parallel retrieval of relevant speech segments alongside visual events.

Our QA pipeline operates in four stages:

- Query parsing: We use Gemini [2] to extract intent, relevant modalities, and any temporal or named entity cues from the user query.
- Dual retrieval: The system retrieves top-k visual events using SigLIP and top-k transcript segments using MiniLM embeddings.
- Answer generation: Gemini is used to generate separate answers from each source (events and transcripts).

[1] https://www.elastic.co/

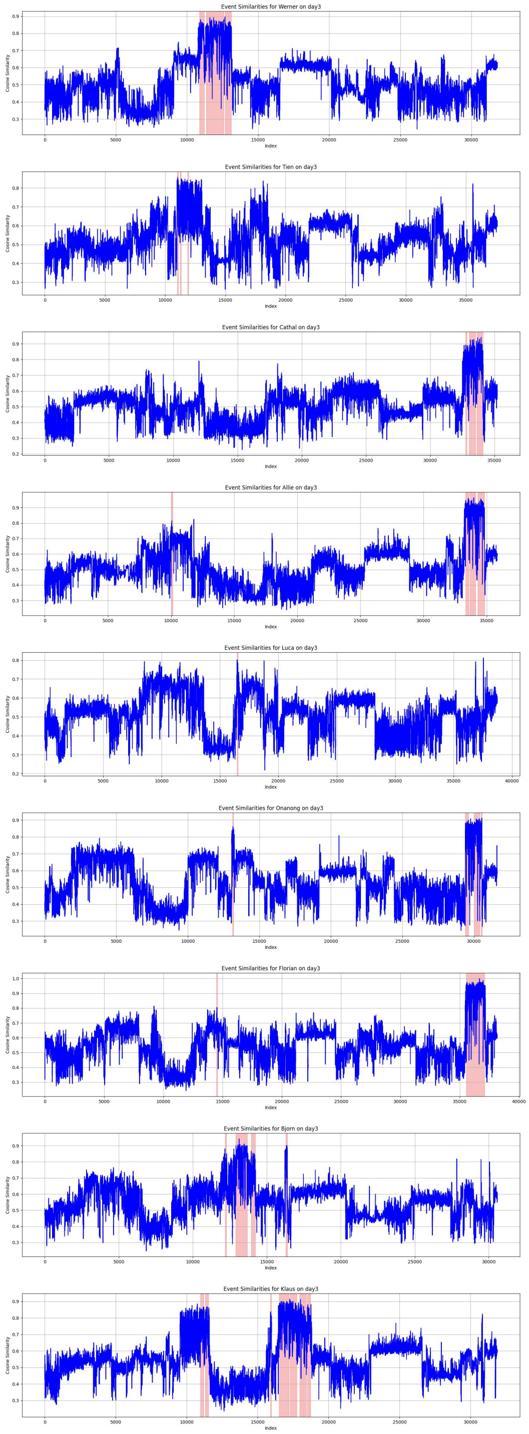


Figure 1: Event Similarity Finding by an anchor image. We use a key frame that is highly similar to the query and search for it by an adjustable sliding window. These examples for the query MM25-EIS04: Find winning hands of poker. In the figure, we can see that using Florian as an anchor frame helps to detect other participants in the event, such as Cathal, Allie, and Onanong.

- Answer consolidation: A final Gemi prompt compares the answers and returns a unified response.

This modular QA system is designed to support a broad spectrum of questions, from visually grounded to speech- or identity-based queries. By reapplying the MyEachtraX framework to the CASTLE dataset, we test its transferability to multi-user, socially complex video contexts.

4 Interactive track

Our interactive approach is built on the philosophy that information retrieval is often not a simple image-seeking task but a process of knowledge exploration and analysis. Users start with an imprecise or incomplete understanding of their information needs and the underlying data structures. Therefore, our system is designed not only to answer precise questions but also to guide users through an iterative cycle of exploration, hypothesis generation, and refinement. Each feature is a step toward increasing the user's knowledge, gradually converting vague intentions into specific queries.

1. Semantic Search. In our system, semantic search is a starting point in a user's retrieval process. It supports vague natural language queries, allowing users to explore a dataset with an initial, often vague, concept of an event. This step captures a large set of potentially relevant moments that serve as raw information for further exploration. This feature is implemented as described in the 3.1 section. The search bar in Figure 2 illustrates how users can enter natural language queries to search for events.

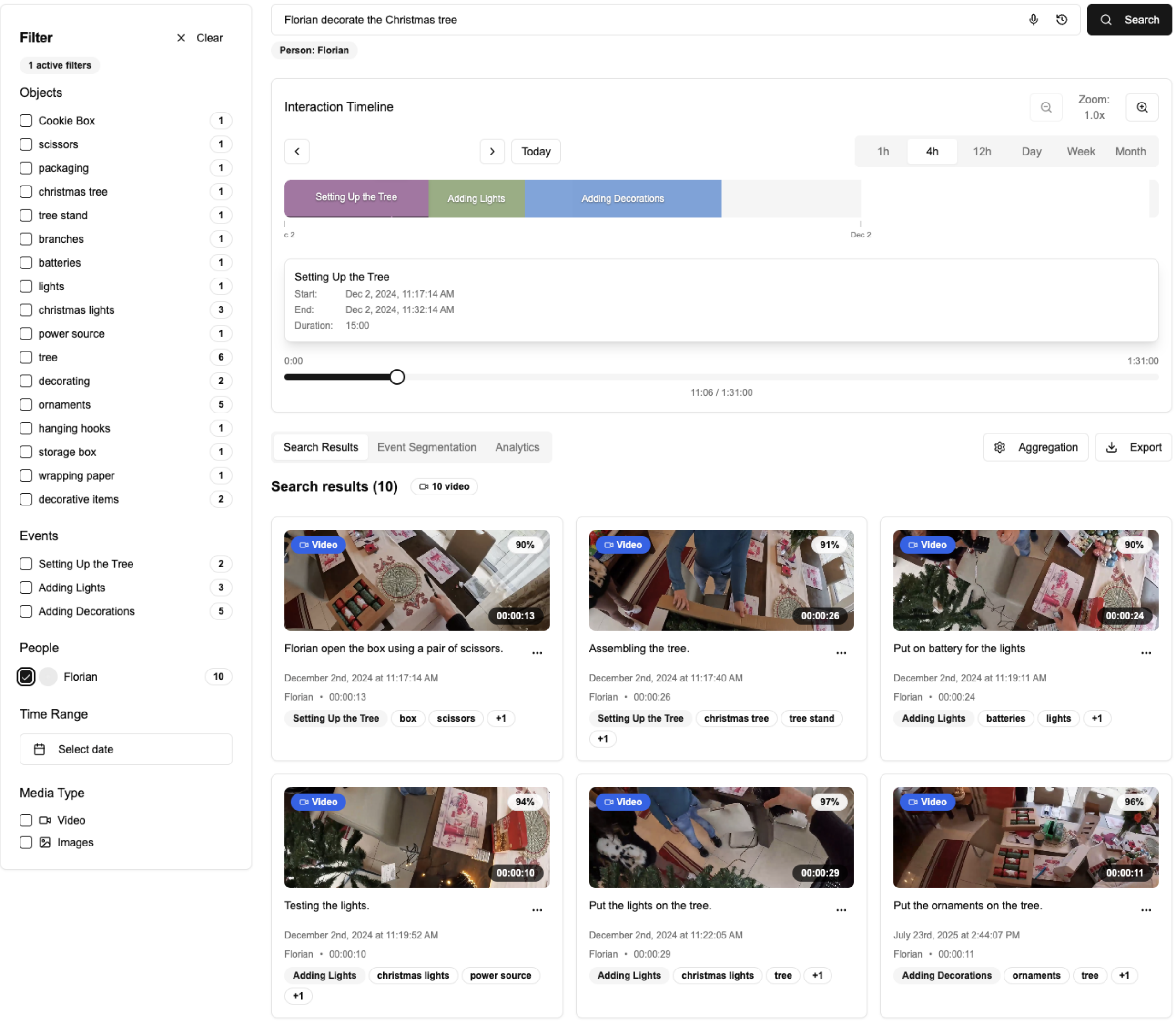
2. Dynamic Filtering for Guided Knowledge Refinement. Based on the initial search, our query-aware dynamic filtering mechanism demonstrates the principle of guided exploration. Instead of requiring the user to know in advance which filters are available, the system analyses the initial results and displays the tags that appear in the results. The observations from the user tags will provide useful information to refine (e.g., specific locations, objects). This step helps users gradually build a clearer target for their query. The sidebar in the user interface illustrates how tags exist in the query.

3. Temporal Event Segments. To improve user understanding of the flow of events, search results are classified into coherent temporal event segments rather than independent frames. Each segment, along with an event short description generated using a similar approach to the Section 3.3, acts as a narrative 'block', reducing the user's cognitive load.

This schema structure is central to helping users quickly understand complex scenes and action sequences, allowing them to construct a timeline of events and identify patterns more effectively than simply viewing the raw video.

Events are mapped to a list of activities inspired by Activities of Daily Living (ADLs/IADLs) [3] categories. This classification provides a concept that makes it easier for users to link the events to familiar real-world concepts, namely filtering by actions based on their understanding of events and forming search paths.

4. Subject-Centred Timeline for Narrative Exploration. For queries that focus on a specific individual, the system generates an interactive timeline. This timeline arranges all relevant event segments generated in the previous section in chronological order, allowing



Florian decorate the Christmas tree

Person: Florian

Setting Up the Tree
Start: Dec 2, 2024, 11:17:14 AM
End: Dec 2, 2024, 11:32:14 AM
Duration: 15:00

Search results (10) 10 video

Florian open the box using a pair of scissors.
December 2nd, 2024 at 11:17:14 AM
Florian • 00:00:13
Setting Up the Tree, box, scissors, +1

Assembling the tree.
December 2nd, 2024 at 11:17:40 AM
Florian • 00:00:26
Setting Up the Tree, christmas tree, tree stand, +1

Put on battery for the lights
December 2nd, 2024 at 11:19:11 AM
Florian • 00:00:24
Adding Lights, batteries, lights, +1

Testing the lights.
December 2nd, 2024 at 11:19:52 AM
Florian • 00:00:10
Adding Lights, christmas lights, power source, +1

Put the lights on the tree.
December 2nd, 2024 at 11:22:05 AM
Florian • 00:00:29
Adding Lights, christmas lights, tree, +1

Put the ornaments on the tree.
July 23rd, 2025 at 2:44:07 PM
Florian • 00:00:11
Adding Decorations, ornaments, tree, +1

Objects

□ Cookie Box	1
□ scissors	1
□ packaging	1
□ christmas tree	1
□ tree stand	1
□ branches	1
□ batteries	1
□ lights	1
□ christmas lights	3
□ power source	1
□ tree	6
□ decorating	2
□ ornaments	5
□ hanging hooks	1
□ storage box	1
□ wrapping paper	1
□ decorative items	2

Events

□ Setting Up the Tree	2
□ Adding Lights	3
□ Adding Decorations	5

People

☑ Florian	10

Time Range

Select date

Media Type

□ Video
□ Images

Figure 2: Interactive user interface of the system.

users to scroll through the subject's activities over time (Figure 2). This tool is specifically built for narrative exploration, allowing for a comprehensive view of the events relevant to the query. By visualizing the sequence of activities, users can jump to a time location with a likelihood of having the desired information.

5. Time Analysis. The final component in the retrieval loop is the analytics engine, which allows users to move from individual observations to comprehensive observations. This algorithm uses a sliding window algorithm over user-defined time frames (e.g., 1 hour, 12 hours, 1 day) to aggregate statistics from segmented events. It can help users synthesize or answer complex analytics questions such as 'How many specific days did *Audience A* engage in *reading*?' or 'What is the longest continuous duration of a *reading* event?'.

5 Case studies

This section illustrates the system's performance on selected queries from each sub-task. While sharing architectural components, each pipeline exhibits different strengths and limitations depending on the retrieval goal.

5.1 Event Instance Search

Using only BLIP2 [7] pre-trained Vision Language Model for the embedding stage is a shortcoming in this task. Although the strength of BLIP2 is exploited in searching and grouping all of the satisfied frames, it faces the lack of a clear understanding of the meaning of the events. For example, in Figure 1, the query having ID 'MM25-EIS04' is *'Find winning hands of poker'*, and by using BLIP as the embedding model and Elasticsearch as the searching tool, the outputs are below:

The keywords *winning hands"* and *poker"* suggest a multi-step reasoning process: *Cards* $\longrightarrow$ *Poker game* $\longrightarrow$ *Players* $\longrightarrow$ *Round context* $\longrightarrow$ *Card comparison* $\longrightarrow$ *Outcome*. However, the model retrieves frames broadly related to cards or poker without distinguishing 'winning' outcomes or their temporal context.

This reveals two key challenges:

- **Model placement:** We lack guidelines for deciding at which stage VLMs should be applied and which models are most suitable for specific sub-tasks.
- **Sub-event segmentation:** An event $\mathcal{E} = \{t_i | i = 1, \ldots, T\}$ may contain sub-events (e.g., a single poker round) lasting only $T' < T$ seconds. Isolating these requires synchronising frames across multiple participants and identifying transitions.

Future work should explore structured representations and temporal reasoning to address these gaps.

5.2 Object Instance Search

This task demonstrates strong performance for well-defined visual targets. For example, the system successfully retrieves accurate results for objects such as a *bird-shaped cookie cutter*, *thermal image camera*, and *set of coloured pens*.

More abstract or fine-grained object queries present greater difficulty. For instance, in the query *'Ace of Spades'*, the system retrieves multiple card images but struggles to consistently distinguish the correct suit. Figure 3 shows the top three retrieved keyframes: two show the Ace of Spades, but the third shows the Ace of Diamonds.

Similar issues arise for queries like *"partially eaten apple"* and *"squirrel-shaped tree ornament"*, where the system returns generic matches (whole apples, generic ornaments) but fails to capture distinctive object attributes. This may reflect limited object-level granularity in the BLIP2 training set.

Future versions of the system could benefit from post-retrieval filters or dedicated object detectors to enhance precision.

5.3 Question Answering

The Question Answering system was tested on a set of multimodal queries requiring integration across vision, transcripts, and identity tracking. Since no ground truth answers were available, we evaluated outputs based on plausibility and the strength of supporting evidence.

In the *Connect Four drawing* query, the system correctly identified Florian's POV as containing visual evidence of the board creation. Although Allie referenced the drawing verbally, the system prioritised visual frames—demonstrating its ability to align observable actions with the query intent.

The *cookie-counting* query required approximate temporal reasoning. While repetitive scenes made it difficult to segment batches cleanly, the system aggregated footage from Luca's POV and inferred a plausible count of three baking rounds.

The most illustrative case was the *missing ingredient* query. Here, the system synthesised ingredient list visuals from Werner's footage and cooking discussions from Stevan's transcript to infer that *caraway seeds* were missing. This ingredient was repeatedly mentioned but never seen, highlighting the system's capacity to reason across modalities using weak but complementary signals.

6 Conclusion

This paper presents the DCU team's system for the CASTLE Challenge at ACM MM 2025. We demonstrated that retrieval techniques from the lifelogging domain, particularly event-based semantic retrieval and question answering pipelines, can be adapted to the CASTLE egocentric dataset with minimal changes. Our fully-automatic pipeline supports object and event instance search using BLIP2 embeddings and event clustering, while our QA module extends the MyEachtraX architecture with transcript-based retrieval and person tracking.

In the interactive track, we designed an interface that promotes exploratory search, narrative reconstruction, and light analytics through semantic filtering, subject-based timelines, and event-level clustering. Qualitative results from the QA task show that the system can combine weak visual and verbal cues to produce plausible, multimodal answers, even when no definitive ground truth is available.

Future work will focus on improving temporal aggregation, cross-subject correlation, and retrieval transparency, particularly in queries requiring comparative reasoning or complex visual states. Overall, the results validate our hypothesis that lifelog retrieval systems offer a transferable and extensible foundation for broader egocentric video analysis.

Acknowledgments

This research was conducted with the financial support of Research Ireland at ADAPT, the Research Ireland Centre for AI-Driven Digital Content Technology at Dublin City University [13/RC/21-06_P2, 18/CRT/6223]. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems* 35 (2022), 23716–23736.

[2] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261* (2025).

[3] Peter F Edemekong, Deb Bomgaars, Sukesh Sukumaran, and Shoshana B Levy. 2019. Activities of daily living. (2019).

[4] Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. 2021. Clip2video: Mastering video-text retrieval via image clip. *arXiv preprint arXiv:2106.11097* (2021).

[5] 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.

[6] Cathal Gurrin, Liting Zhou, Graham Healy, Allie Tran, Luca Rossetto, Werner Bailer, Duc-Tien Dang-Nguyen, Steve Hodges, Björn Þór Jónsson, Minh-Triet Tran, et al. 2025. Introduction to the 8th Annual Lifelog Search Challenge, LSC'25. In *Proceedings of the 2025 International Conference on Multimedia Retrieval*. 2143–2144.

[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:2301.12597 [cs.CV] https://arxiv.org/abs/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] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft COCO: Common Objects in Context. arXiv:1405.0312 [cs.CV] https://arxiv.org/abs/1405.0312

[10] Nicola Messina, Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Fabrizio Falchi, Giuseppe Amato, and Rita Cucchiara. 2022. Aladin: distilling fine-grained

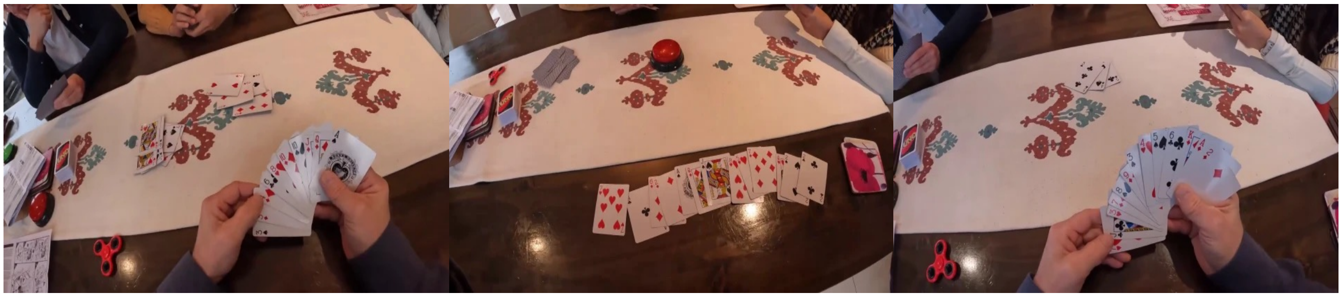
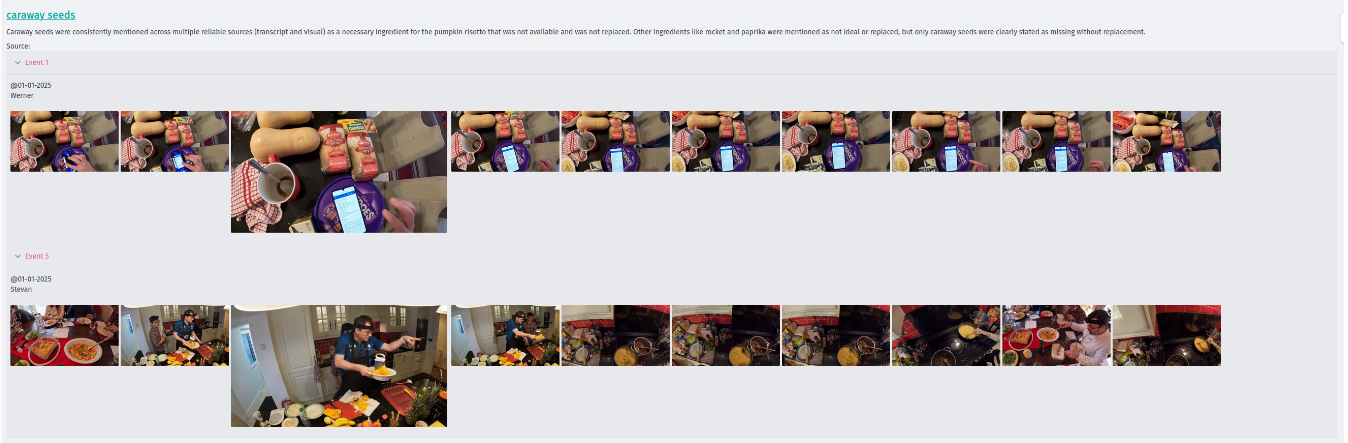


Figure 3: Top 3 retrieved keyframes for the query ‘Ace of Spades.’ Two correct matches and one wrong (Ace of Diamonds).



caraway seeds

Caraway seeds were consistently mentioned across multiple reliable sources (transcript and visual) as a necessary ingredient for the pumpkin risotto that was not available and was not replaced. Other ingredients like rocket and paprika were mentioned as not ideal or replaced, but only caraway seeds were clearly stated as missing without replacement.

Source:

Event 1

@01-01-2025
Werner

Event 5

@01-01-2025
Stevan

Figure 4: QA05 – Missing Ingredient Reasoning. Evidence pointing to the absence of caraway seeds, extracted from Werner and Stevan's perspectives.

alignment scores for efficient image-text matching and retrieval. In *Proceedings of the 19th international conference on content-based multimedia indexing*. 64–70.

[11] Thang-Long Nguyen-Ho, Onanong Kongmeesub, Minh-Triet Tran, Dongyun Nie, Graham Healy, and Cathal Gurrin. 2024. Eagle: Eyegaze-assisted guidance and learning evaluation for lifeloging retrieval. In *Proceedings of the 7th Annual ACM Workshop on the Lifelog Search Challenge*. 18–23.

[12] Nick Pantelidis, Maria Pegia, Damianos Galanopoulos, Konstantinos Apostolidis, Klearchos Stavrothanasopoulos, Anastasia Moumtzidou, Konstantinos Gkountakos, Ilias Gialampoukidis, Stefanos Vrochidis, Vasileios Mezaris, et al. 2024. Verge in vbs 2024. In *International Conference on Multimedia Modeling*. Springer, 356–363.

[13] 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.

[14] 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.

[15] Luca Rossetto, Werner Bailer, Duc-Tien Dang-Nguyen, Graham Healy, Björn Þór Jónsson, Onanong Kongmeesub, Hoang-Bao Le, Stevan Rudinac, Klaus Schoeffmann, Florian Spiess, Allie Tran, Minh-Triet Tran, Quang-Linh Tran, and Cathal Gurrin. 2025. The CASTLE 2024 Dataset: Advancing the Art of Multimodal Understanding. *CoRR* abs/2503.17116 (2025). arXiv:2503.17116 doi:10.48550/ARXIV.2503.17116

[16] Loris Sauter, Ralph Gasser, Silvan Heller, Luca Rossetto, Colin Saladin, Florian Spiess, and Heiko Schuldt. 2023. Exploring effective interactive text-based video search in vitrivr. In *International Conference on Multimedia Modeling*. Springer, 646–651.

[17] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 815–823.

[18] Allie Tran, Werner Bailer, Duc-Tien Dang-Nguyen, Graham Healy, Steve Hodges, Björn Þór Jónsson, Luca Rossetto, Klaus Schoeffmann, Minh-Triet Tran, Lucia Vadicamo, and Cathal Gurrin. 2025. The State-of-the-Art in Lifelog Retrieval: A Review of Progress at the ACM Lifelog Search Challenge Workshop 2022-24. *CoRR* abs/2506.06743 (2025). arXiv:2506.06743 doi:10.48550/ARXIV.2506.06743

[19] Ly-Duyen Tran, Nguyen Thanh Binh, Cathal Gurrin, and Liting Zhou. 2024. MyEachtraX: Lifelog Question Answering on Mobile. In *Proceedings of the 7th Annual ACM Workshop on the Lifelog Search Challenge, LSC 2024, Phuket, Thailand, 10 June 2024*. ACM, 93–98. doi:10.1145/3643489.3661128

[20] Quang-Linh Tran, Binh Nguyen, Gareth JF Jones, and Cathal Gurrin. 2024. MemoriEase 2.0: A Conversational Lifelog Retrieve System for LSC'24. In *Proceedings of the 7th Annual ACM Workshop on the Lifelog Search Challenge*. 12–17.

[21] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. 2025. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786* (2025).

[22] Rejin Varghese and M Sambath. 2024. Yolov8: A novel object detection algorithm with enhanced performance and robustness. In *2024 International conference on advances in data engineering and intelligent computing systems (ADICS)*. IEEE, 1–6.

[23] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems* 33 (2020), 5776–5788.

[24] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. 2017. Simple Online and Realtime Tracking with a Deep Association Metric. *CoRR* abs/1703.07402 (2017). arXiv:1703.07402 http://arxiv.org/abs/1703.07402

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