# T-EAGLE: Capturing Temporal Narratives via Sequence Captioning and Text Matching

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

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

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

Cathal Gurrin
School of Computing
Dublin City University
Dublin, Ireland
cgurrin@computing.dcu.ie

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

## Abstract

There is a growing need to retrieve specific events or information from personal lifelog data, but this is particularly challenging due to the massive scale and the passive nature of data capture by lifelogging devices. Current systems typically rely on image similarity for single, isolated images, which struggle to capture the user intent expressed in natural language and the semantic links between the images and activities occurring over time. To address this issue, we propose a novel lifelog retrieval framework that explicitly combines both visual and temporal similarity in a multi-stage process, shifting the focus from single images to coherent sequences of actions. Our approach uses image embeddings to initialize a set of candidate images. Importantly, the system then re-evaluates the query similarity based on action descriptions which contain temporal information across image sequences. Action captioning, integrated into the indexing process, captures richer temporal and semantic context, allowing the system to distinguish between visually similar but semantically distinct events. Additionally, the system incorporates an evidence-based question answering mechanism, in which the narratives of the retrieved sequences provide contextual grounding for the answering model. The paper proposes a hybrid retrieval framework that combines image similarity for candidate initialization and visual-textual similarity for event retrieval. The integration of action descriptions enables language-based temporal representation of events. These are extracted offline through semantic content analysis and serve as the basis for building an evidence-based Question Answering module using these narratives as context. This approach helps bridge the gap between user intent and the multimodal, temporally structured nature of lifelog data.

## CCS Concepts

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

*Corresponding author


*LSC '25, Chicago, IL, USA*

ACM ISBN 979-8-4007-1857-1/25/06
https://doi.org/10.1145/3729459.3748691

## Keywords

lifelog, interactive retrieval systems, semantic embedding

**ACM Reference Format:**
Thang-Long Nguyen-Ho, Allie Tran, Minh-Triet Tran, Cathal Gurrin, and Graham Healy. 2025. T-EAGLE: Capturing Temporal Narratives via Sequence Captioning and Text Matching. In *Proceedings of the 8th Annual ACM Workshop on the Lifelog Search Challenge (LSC '25), June 30–July 03, 2025, Chicago, IL, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3729459.3748691

## 1 Introduction

Lifelogging devices have the potential to generate massive personal archives, capturing our daily experiences in incredible detail. However, figuring out the best way to structure this data for efficient storage and retrieval is still a challenge. Even with recent advances in multimedia analysis and interactive embeddings (discussed in Section 2), a significant challenge remains: understanding the intent of the users, often subtly expressed in natural language. Specifically, bridging the gap between the raw visual content and the meaningful actions or events it represents is a core issue. Users frequently search for memories linked to specific activities unfolding over time, rather than merely seeking visually similar snapshots. This necessitates a search system sensitive to both visual and temporal understanding.

Our previous method (EAGLE [12]) mainly focused on retrieving the most relevant *'shots'* (single images) based on visual similarity. To address this limitation, we group the results by part of the day (morning, afternoon, evening) and by locations. While this improved the interaction, it still relied on the user's own knowledge to find relevant results faster. To address this limitation more systematically, in this work, we propose a system that explicitly combines visual and textual similarity in a multi-stage retrieval process. We shift the focus from retrieving single images to identifying coherent action segments, or *'events'*, represented by groups of temporally consecutive images and the generated textual descriptions for them. This allows for richer temporal context to be captured; for example, while multiple images may depict the same airport scene, analyzing a sequence and accompanying description can help distinguish between 'arriving at the airport' and 'leaving the airport'.

Our system aims to bridge the gap between user intent and lifelog content by using a combination of image and text similarity. The system first interprets the user's intent through natural language input, then retrieves relevant candidates through a two-stage process designed to identify specific events:

> Stage 1: Candidate Retrieval: The system uses image similarity via image embeddings along with metadata filters to efficiently identify a set of *candidate images* that are visually relevant to the query. This is the initial search step to locate images that are likely to contain results.
> Stage 2: Event Retrieval as Document Retrieval: The system then treats the textual descriptions (narratives) of action sequences (events) as documents. It then performs document retrieval by measuring the textual similarity between the user query and these narratives to retrieve the most relevant *action sequences*. These 'documents' contain both narrative and timing information for each event.

For answer-type questions, our method can provide **Evidence-Based Answers**: At this point, the narratives of the retrieved action sequences act as contextual evidence, which a Large Language Model uses to determine the final answer.

In conclusion, this work presents a framework that explicitly integrates image similarity for candidate identification and text similarity for retrieving specific actions or events. A key contribution of this approach is treating temporal action as documents, which allows more contextual information to be compressesed into each image representation. To facilitate the temporal representation, our system incorporates action description directly into the indexing process. Finally, the effectiveness of this methodology is demonstrated through an evidence-based question-answering mechanism, which uses these textual descriptions as contextual grounding for a large language model to perform reasoning and answer user queries.

## 2 Related Work

Research in the area of lifelog search often focuses on locating specific moments or recalling information within personal data archives. Benchmarking activities such as the Lifelog Search Challenge (LSC) [4–7] serve as an annual evaluation platform for state-of-the-art interactive lifelog retrieval techniques. This section reviews the main approaches used in modern lifelog search systems and how they have been improved over time to enhance the search process.

Many of the systems at LSC'24 have leveraged the capabilities of Vision Language Models (VLMs) such as CLIP [14] and BLIP [11] to generate rich semantic representations for both images and text. By embedding these two modalities into a shared latent space, systems such as LifeInsight 2.0 [19] (which combines CLIP and BLIP2) and LifeGraph 4 [15](which integrates VLMs into the knowledge graph) can perform semantic similarity-based search, going beyond the limits of simple metadata or keyword matching. LifeSeeker 6.0 [9] enriches their data using VLMs to automatically generate metadata captions (CapMeta) to increase the level of content understanding.

However, some argue that understanding individual moments is not enough; lifelogs are inherently a continuous flow of activities and events. This perspective was first introduced at the LSC workshops by MyEachtra [16], which proposed treating 'events' as the atomic units of retrieval. In this approach, event features were constructed by aggregating the visual embeddings of the images that belong to the same event. Building on the content understanding foundation provided by VLMs, another key development at LSC'24 was the segmentation of lifelog data into more structured and meaningful units. Instead of looking at each image in isolation, systems like Memento 4.0 [1] apply hierarchical event segmentation to group related moments together, enabling retrieval based on events rather than single images. MyEachtraX [17], a later version of MyEachtra, further reinforces the value of retrieving related event segments to answer complex multi-step questions. MEMORIA [3] also aims to suggest and manage important events. Structuring lifelog data into events not only makes management and retrieval more efficient, but also aligns more closely with the way humans organize and recall their memories, enabling more complex forms of interaction. To enable users to interact intuitively and effectively with both structured lifelog data (into events) or semantically rich raw data, Large Language Models (LLMs) have emerged as a key technology. LLMs enable systems to deeply understand complex queries in natural language, engage in conversations, and even reason to answer questions based on lifelog content. MemoriEase 2.0 [18] and Memento 4.0 [1] are prominent examples of conversational systems that use LLMs to interpret user intent and orchestrate the search or question answering process based on segmented events. MyEachtraX [17] adopts an advanced RAG [10] architecture, where LLM plays a central role in combining information from retrieved events to generate answers.

## 3 System

To address the challenges of lifelog retrieval based on user intent and event context, we propose a system that integrates multimodal processing and event-based inference. The system operates through two main phases: an offline Indexing Phase, which prepares and represents the data, and an online Retrieval Phase, which processes user queries through a two-step procedure that combines visual and textual information.

### 3.1 Data Representation and Indexing

This phase focuses on converting raw lifelog data into a structured, semantically rich format suitable for efficient retrieval. Figure 1 describes the data preparation process.

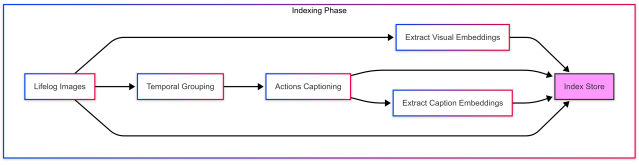


**Figure 1: Overview of the process flow before indexing data into the database. Lifelog images are split into days, which are then further divided into smaller segments. For each segment, a narrative of actions is generated. Finally, both the visual embedding and the action narrative embedding are stored in the database.**

*3.1.1 Visual Feature Extraction:* All images in the lifelog dataset are transformed into a visual embedding vector using a pretrained CLIP model, which captures rich semantic content. To ensure consistency in metrics at query time, we normalize features to remove the influence of magnitude when calculating similarity.

*3.1.2 Temporal Grouping and Event Segmentation:* Recognizing that events are temporally continuous and occur during the day, we group temporally adjacent images into potential segments within a day. Each group consists of 96[1] consecutive keyframes with high information density [8], which has been shown to be effective in capturing action sequences and identifying potential event boundaries. Each of these groups represents a basic 'action sequence'.
For each action sequence, multiple action descriptions are generated. For each chunk, we use Tariser [20] model that is capable of describing each action segment based on the visual content. These natural language captions aim to represent both temporal information and semantic context of the image sequence. At the same time, we also compute and store a text embedding vector for each caption using Stella Embedding [21] because of its balance between model size and performance.

*3.1.3 Indexing Storage:* The visual embeddings extracted from each image and the text embeddings corresponding to each generated action sequence are stored in the Milvus vector database, which is optimized for high-performance similarity search. At the same time, relevant metadata such as timestamps and location information are indexed and stored in Elasticsearch to support filtering and composing general-purpose queries based on these attributes. Using two databases enables efficient monitoring and scoring.

## 3.2 Retrieval Process

When a user issues a query $Q$ in natural language, the system performs the following multi-stage retrieval process, illustrated in Figure 2.

*3.2.1 Query Processing:* The natural language query $Q$ is processed to generate an embedding vector using the image-text model, which will be used for Candidate Selection Stage 3.2.2, and an embedding vector using the action embedding model, which will be used for Retrieval Stage 3.2.3. To understand the user intent and specific entities, we use an Name Entity Recognition in English [13] to extract specific entities such as locations, actions, and times of day to be used as constraints and to score the returned results.

*3.2.2 Stage 1: Visual Candidate Retrieval:* The first stage focuses on finding visually relevant images. The system performs a similarity search in the image embedding space. Specifically, it retrieves the top $K$ images whose visual embeddings are closest to the query embedding in the latent space of the model such as CLIP. This initial candidate set is further filtered based on the metadata constraints inferred from the query. The result of this stage is a set of potential candidate images.

*3.2.3 Stage 2: Text-based Event Retrieval:* The candidate images from the previous stage are mapped to the corresponding action sequences to which they belong. The core of the method lies in this stage, where we perform text-based retrieval, which is a form of document retrieval. System then uses relevance scoring algorithms [2] to retrieve $k$ corresponding action sequences that are likely to contain the answer. This process treats each action sequence as a 'document' and retrieves the 'documents' whose descriptions best match the user query, thereby integrating the temporal and contextual contexts described in the narrative.

[1]We use 96 consecutive keyframes, consistent with our video captioning model's training [20], a length shown to effectively capture action sequences and event boundaries.

## 3.3 Evidence-Based Question Answering

In the final phase of the system, we focus on Question and Answer (QA) generation. This phase operates on a Retrieval-Augmented Generation architecture. The top-k action sequence narratives from 3.2 are considered most relevant to the user's query as the context for answer generation.

These narratives typically contain a sentence describing the event segment and related structured metadata (JSON format) extracted and indexed from the dataset (Listing 1). Each action sequence is separated and typically accompanied by metadata.

This phase focuses on interaction with a Large Language Model (LLM), which we use here as GPT-4. The input to the LLM is carefully constructed from the original user question, along with narrative representations of the top-k retrieved action sequences.

```
User query: [User query]

Evidence fragment 1 (Timestamp: [timestamp]):
[Text of action sequence 1]
[json containing metadata of action sequence 1]

Evidence fragment 2 (Timestamp: [timestamp]):
[Text of action sequence 2]
[json containing metadata of action sequence 2]

...

Evidence fragment k (Timestamp: [timestamp]):
[Text of action sequence k]
[json containing metadata of action sequence k]


Instruction Prompt: Only based on the evidence fragments
provided above, answer the user's question. Think
carefully before giving the answers. If the answer cannot
be determined from the evidence, state so.
```

**Listing 1: Structured Evidence Input Format**

The number of pieces of evidence (K) represents a trade-off: higher 'k' values provide more potential evidence but also increase the risk of introducing irrelevant information and may exceed the answering model's contextual processing or attentional capacity. Conversely, lower 'k' provides more attention but may miss the necessary context. The LLM is asked to generate an answer based solely on the provided context, ensuring that the answer is grounded in the retrieved lifelog segments.

## 3.4 User interface

Inspired by the concept Natural Language Interface that our team proposed previously [12], we continue to apply this philosophy to the current interface to improve the user experience. The core philosophy is to minimize active interactions and instead prioritize passive feedback from users. The system interface is improved by

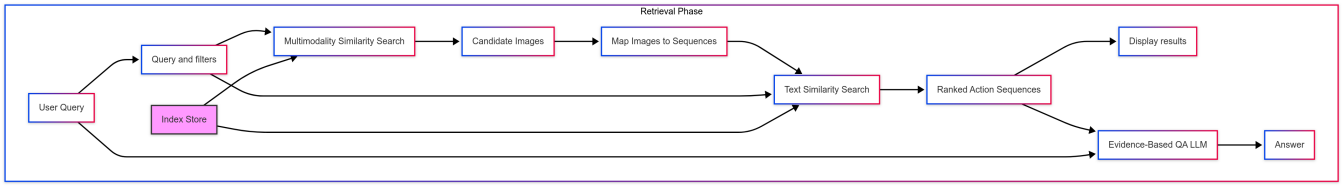


**Figure 2: System retrieval pipeline. After extracting entities using the parser, the system identifies potential candidates that may contain the answer. Next, it maps these candidates to action sequences that include them and retrieves the complete segment. Subsequently, it uses action narratives embeddings to rerank the candidates and select the top ones for answer composition.**

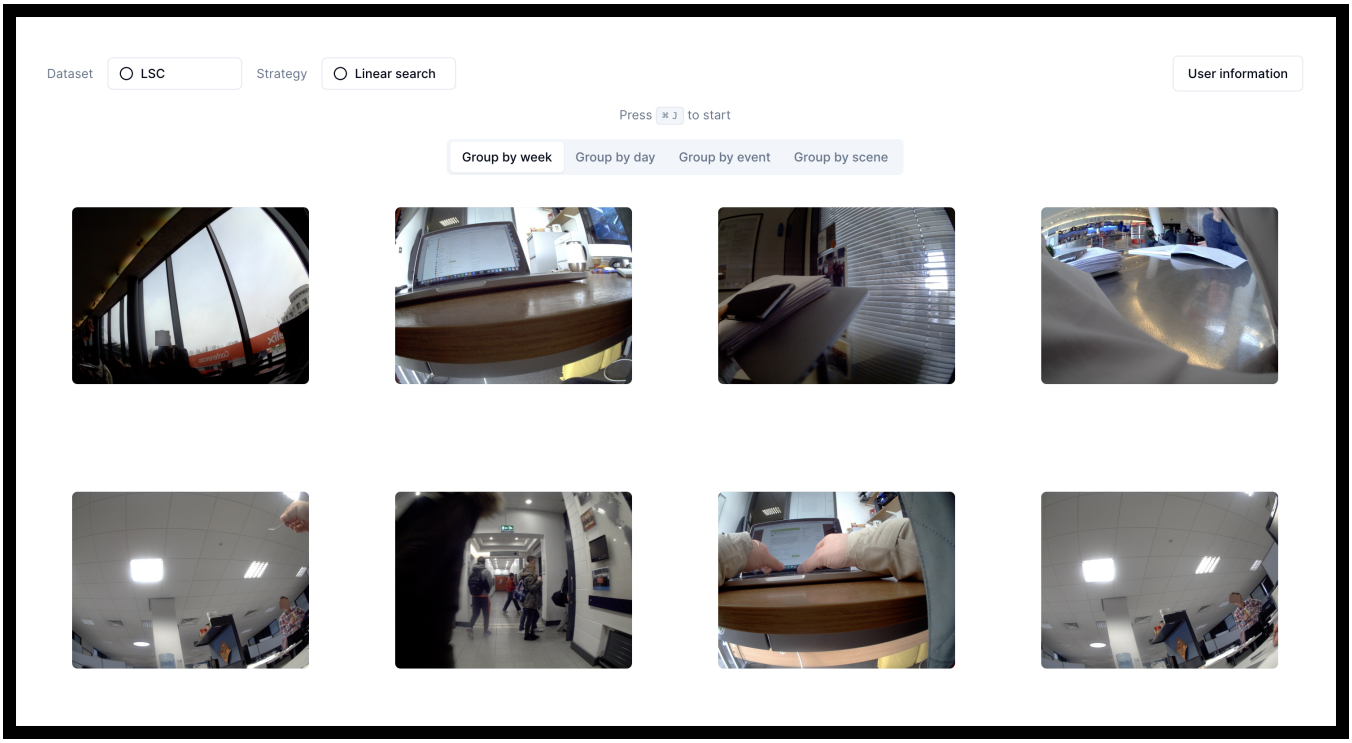


**Figure 3: Our system interface displayed eight moments on the screen at once. Users can group these moments in various ways, including by week, day, event segment, or location. Searching begins with a search bar where users can input queries in natural language, which are then converted into system language to retrieve the results.**

eliminating functions that require complex direct input, aiming for users to mainly control the system through simple text interactions.

Unlike traditional interfaces that often contain many input fields (e.g. searching by metadata, entering multiple fields to determine time) that require users to analyze and become familiar with how the system operates. Our new user interface focuses only on viewing and grouping images. To minimize interactions, the user interface only provides a single text input field. After the user enters a query, the system automatically executes the algorithm to find the results, thereby completely eliminating the need for the user to have any background knowledge about the system. Figure 3 illustrates the main interface of our system.

This concept provides a higher level of satisfaction than traditional natural language interfaces. The increased focus on algorithm optimization proves to be robust, works well with a wide range of queries, and is suitable for different user preferences.

## 4 Conclusion

We proposed a temporal query-focused version of EAGLE, intended to improve understanding of user intent expressed in natural language, and to capture temporal links between images by shifting the focus from single images to coherent action sequences, combining visual and temporal similarity in a multi-stage process.

The system utilizes image similarity for initial candidates. Then refines the results using textual similarity based on action narratives for image sequences. The system considers these sequences as 'documents' for retrieval. Integrating action captioning into the indexing process captures richer temporal and semantic context,

distinguishing between visually similar but contextually different events.

Furthermore, the system incorporates an evidence-based question-answering mechanism where the narratives of retrieved sequences provide context for a Large Language Model to generate answers grounded in the lifelog data. This approach effectively bridges the gap between user intent and the complex, multimodal, temporal nature of lifelog archives.

## Acknowledgments

This publication has emanated from research conducted with the financial support of or supported in part by a grant from Science Foundation Ireland under Grant numbers 18/CRT/6223 and 13/RC/2106_P2 at the ADAPT SFI Research Centre at Dublin City University and the support of the Faculty of Engineering & Computing, DCU.

## References

[1] Naushad Alam, Yvette Graham, and Cathal Gurrin. 2024. Memento 4.0: A Prototype Conversational Search System for LSC'24. In *Proceedings of the 7th Annual ACM Workshop on the Lifelog Search Challenge* (Phuket, Thailand) *(LSC '24)*. Association for Computing Machinery, New York, NY, USA, 82–87. https://doi.org/10.1145/3643489.3661126

[2] Giambattista Amati. 2009. *BM25*. Springer US, Boston, MA, 257–260. https://doi.org/10.1007/978-0-387-39940-9_921

[3] Alexandre Gago, Bernardo Kaluza, Eva Bartolomeu, Josefa Pandeirada, Ricardo Ribeiro, and António J. R. Neves. 2024. MEMORIA: A Memory Enhancement and MOment RetrIeval Application at the LSC2024. In *Proceedings of the 7th Annual ACM Workshop on the Lifelog Search Challenge* (Phuket, Thailand) *(LSC '24)*. Association for Computing Machinery, New York, NY, USA, 99–104. https://doi.org/10.1145/3643489.3661129

[4] Cathal Gurrin, Björn Þór Jónsson, Duc Tien Dang Nguyen, Graham Healy, Jakub Lokoc, Liting Zhou, Luca Rossetto, Minh-Triet Tran, Wolfgang Hürst, Werner Bailer, and Klaus Schoeffmann. 2023. Introduction to the Sixth Annual Lifelog Search Challenge, LSC'23. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval* (Thessaloniki, Greece) *(ICMR '23)*. Association for Computing Machinery, New York, NY, USA, 678–679. https://doi.org/10.1145/3591106.3592304

[5] 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* (Phuket, Thailand) *(ICMR '24)*. Association for Computing Machinery, New York, NY, USA, 1334–1335. https://doi.org/10.1145/3652583.3658891

[6] Cathal Gurrin, Liting Zhou, Graham Healy, Björn Þór Jónsson, Duc-Tien Dang-Nguyen, Jakub Lokoć, Minh-Triet Tran, Wolfgang Hürst, Luca Rossetto, and Klaus Schöffmann. 2022. Introduction to the Fifth Annual Lifelog Search Challenge, LSC'22. In *Proceedings of the 2022 International Conference on Multimedia Retrieval* (Newark, NJ, USA) *(ICMR '22)*. Association for Computing Machinery, New York, NY, USA, 685–687. https://doi.org/10.1145/3512527.3531439

[7] Cathal Gurrin, Liting Zhou, Graham Healy, Allie Tran, Luca Rossetto, Werner Bailer, Duc-Tien Dang-Nguyen, Steve Hodges, Björn Þór Jónsson, Minh-Triet Tran, and Klaus Schöffmann. 2025. Introduction to the 8th Annual Lifelog Search Challenge, LSC'25. In *Proceedings of the 2025 International Conference on Multimedia Retrieval* (Chicago, IL, USA) *(ICMR '25)*. Association for Computing Machinery, New York, NY, USA, 2143–2144. https://doi.org/10.1145/3731715.3734579

[8] Feiyan Hu and Alan F Smeaton. 2018. Image aesthetics and content in selecting memorable keyframes from lifelogs. In *MultiMedia Modeling: 24th International Conference, MMM 2018, Bangkok, Thailand, February 5-7, 2018, Proceedings, Part I 24*. Springer, 608–619.

[9] 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* (Phuket, Thailand) *(LSC '24)*. Association for Computing Machinery, New York, NY, USA, 53–57. https://doi.org/10.1145/3643489.3661121

[10] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv:2005.11401 [cs.CL] https://arxiv.org/abs/2005.11401

[11] 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.

[12] 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* (Phuket, Thailand) *(LSC '24)*. Association for Computing Machinery, New York, NY, USA, 18–23. https://doi.org/10.1145/3643489.3661115

[13] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. https://nlp.stanford.edu/pubs/qi2020stanza.pdf

[14] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020 [cs.CV] https://arxiv.org/abs/2103.00020

[15] Luca Rossetto, Athina Kyriakou, Svenja Lange, Florian Ruosch, Ruijie Wang, Kathrin Wardatzky, and Abraham Bernstein. 2024. LifeGraph 4 - Lifelog Retrieval using Multimodal Knowledge Graphs and Vision-Language Models. In *Proceedings of the 7th Annual ACM Workshop on the Lifelog Search Challenge* (Phuket, Thailand) *(LSC '24)*. Association for Computing Machinery, New York, NY, USA, 88–92. https://doi.org/10.1145/3643489.3661127

[16] Ly Duyen Tran, Binh Nguyen, Liting Zhou, and Cathal Gurrin. 2023. MyEachtra: Event-based interactive lifelog retrieval system for lsc'23. In *Proceedings of the 6th Annual ACM Lifelog Search Challenge*. 24–29.

[17] Ly Duyen Tran, Thanh-Binh Nguyen, 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* (Phuket, Thailand) *(LSC '24)*. Association for Computing Machinery, New York, NY, USA, 93–98. https://doi.org/10.1145/3643489.3661128

[18] Quang-Linh Tran, Binh Nguyen, Gareth J. F. 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* (Phuket, Thailand) *(LSC '24)*. Association for Computing Machinery, New York, NY, USA, 12–17. https://doi.org/10.1145/3643489.3661114

[19] Gia Huy Vuong, Van-Son Ho, Tien Thanh Nguyen Dang, Xuan-Dang Thai, Thang-Long Nguyen-Ho, Minh-Khoi Pham, Tu-Khiem Le, Van-Tu Ninh, Graham Healy, Cathal Gurrin, and Minh-Triet Tran. 2024. 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* (Phuket, Thailand) *(LSC '24)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3643489.3661112

[20] Liping Yuan, Jiawei Wang, Haomiao Sun, Yuchen Zhang, and Yuan Lin. 2025. Tarsier2: Advancing Large Vision-Language Models from Detailed Video Description to Comprehensive Video Understanding. arXiv:2501.07888 [cs.CV] https://arxiv.org/abs/2501.07888

[21] Dun Zhang, Jiacheng Li, Ziyang Zeng, and Fulong Wang. 2025. Jasper and Stella: distillation of SOTA embedding models. arXiv:2412.19048 [cs.IR] https://arxiv.org/abs/2412.19048