Memento 4.0: A Prototype Conversational Search System for LSC'24

Naushad Alam
SFI Insight Centre for Data Analytics, Dublin City University
Dublin, Ireland
naushad.alam2@mail.dcu.ie

Yvette Graham
ADAPT Centre, Trinity College Dublin, Ireland
ygraham@tcd.ie

Cathal Gurrin
ADAPT Centre, Dublin City University
Dublin, Ireland
cathal.gurrin@dcu.ie

ABSTRACT

The practice of lifelogging, capturing one's daily experiences through wearable devices, has evolved significantly over the last decade, presenting both challenges and opportunities in information retrieval. This paper presents an early prototype of a conversational lifelog retrieval system designed to address the open challenges in this domain. Our system integrates a hierarchical event segmentation approach to automatically organize lifelog data into meaningful events, facilitating event-based retrieval over traditional image retrieval. Moreover, we incorporate a question-answering pipeline, leveraging large language models such as GPT-3.5 Turbo and Mistral7B, to enable free-form natural language interaction with the lifelog dataset. Moreover, we enhance our system's user interface by building on previous versions to streamline event-based retrieval and question-answering functionalities.

CCS CONCEPTS

• **Information systems** → **Retrieval models and ranking**; **Search interfaces**.

KEYWORDS

lifelog, information retrieval, semantic image representation, conversational search

ACM Reference Format:
Naushad Alam, Yvette Graham, and Cathal Gurrin. 2024. Memento 4.0: A Prototype Conversational Search System for LSC'24. In *The 7th Annual ACM Lifelog Search Challenge (LSC '24), June 10, 2024, Phuket, Thailand.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3643489.3661126

1 INTRODUCTION

The practice of documenting one's life has undergone a range of transformations throughout history. Earlier approaches centered around writing journals and diaries, while more contemporary methods include capturing moments through photographs and tracking activities via digital wearable devices. Lifelogging is one such contemporary means of documenting oneself with the help of devices such as wearable cameras and wearable activity trackers which aim to capture everyday life experience of a person in a passive manner, i.e. the documentation process is automatic and continuous without requiring any human intervention. The dataset subsequently collected is one containing a large in-the-wild multimodal archive consisting of egocentric images and metadata including location, date-time, activity data, amongst others.

The goal of digitally documenting oneself could be many. It could be done by enthusiasts who would want to record themselves to allow them to revisit life events at a later time and/or keep track of activities over a longer period. It could also be used to help people suffering from health ailments such as dementia where such an archive could be leveraged during treatment, e.g. for memory aides or reminiscence therapy. Retrieving information from lifelogs however has its own set of challenges given it is a noisy archive as pictures are captured from a first-person perspective at regular intervals using the wearable camera which leads to a set of observable scenes, that on occasion are blurry or frequent shift in terms of point-of-view due to motion. The interaction methodology is also a crucial aspect of lifelog retrieval systems. Over the years, several systems have been proposed demonstrating novel and effective interaction modes such as concept-based retrieval methodology using keywords, virtual-reality based interaction, and more recently free-form natural language based methods. There also has been a shift towards building conversational retrieval systems for lifelogs which can handle free-form question-answering over the dataset in natural language. The annual benchmarking challenge for lifelog information retrieval, the Lifelog Search Challenge [10] also introduced question-answering (QA) query type last year, in addition to the pre-existing query types to push research in this direction.

In this work, we present an upgraded version of our previous systems [1–3] which have been participating in the Lifelog search challenge since 2021. Our proposed system aims to address many open challenges in this research domain. Since the goal of information retrieval from lifelogs is to retrieve relevant events or moment from the dataset which poses a unique challenge of organizing and indexing the dataset by segmenting them into events. To address this, we devised a hierarchical event segmentation approach to automatically segment lifelog data into events and transition towards 'event' retrieval as opposed to image retrieval. Furthermore, we incorporate a question-answering pipeline to handle QA query types which leverages event summaries generated using large language models like GPT 3.5 [7] and Mistral7B [13] to answer free-form natural language questions over the lifelog dataset.

2 RELATED WORK

2.1 Lifelog Search

Lifelog search is defined as searching the multimodal corpus consisting of images, textual metadata as well as sensor data to retrieve moments or events from a person's life. Over the recent few years

This work is licensed under a Creative Commons Attribution 4.0 International License.
LSC '24, June 10, 2024, Phuket, Thailand
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0550-2/24/06
https://doi.org/10.1145/3643489.3661126

several novel retrieval systems have been proposed for lifelog information retrieval especially to participate in the search challenges such as Lifelog Search Challenge [9, 10] and NTCIR-Lifelog tasks [8, 22, 23]. The 2023 Lifelog Search Challenges [10] also attracted multiple system from across the world which included returning participants as well as many first time participants.

MyEachtra [20] proposed a event-based user-interface and incorporated question-answering models based on FrozenBiLM which is a video question-answering model to address question-answering queries. Voxento [5] used CLIP [17] embeddings for image retrieval and made improvements in their user interface to better support the end-user. E-LifeSeeker [15] also used CLIP embeddings to build their search backend and improved their core engine with the latest pre-trained embedding models. They also used differential networks to address the question-answering queries in the challenge. Memoria [18] also used CLIP embeddings to build their search engine. However, they did not used the embeddings for image-text similarity but instead leveraged them to generate image captions for lifelog images which are then used to build the search engine. They also propose a methodology to segment the lifelog dataset into events as well as integrate free-text search into their system.

LifeInsight [16] used the BLIP model as their backend model and had features such as visual similarity search, relevance feedback function and AI-based query description rewriting mechanism to better support the end-user during the search process. LifeLens [12] presented a novel minimalist user interface design to improve the usability and ease of use of an interactive lifelog retrieval system. They incorporate features such as user feedback mechanism to search for similar images by selecting a group of images, faceted filtering based on time, location, people as well as grid view for viewing search results. LifeXplore [19] used image-text embeddings derived from the OpenCLIP model to build their search backend. MemoriEase [21] employed an embedding-based retrieval approach with BLIP as the main search engine and combined it with a concept-based retrieval approach to further improve performance. FIRST [11] adopted generative models to equip the system with predictive ability rather than entirely relying on the user to input the query. Our system Memento 3.0 [3] which participated in LSC 2023 also like many other participating system used embeddings from CLIP and OpenCLIP models to build the search backend. We implemented a feature allowing users to switch between various models from the CLIP and OpenCLIP model suite to conduct search, granting them greater flexibility based on their needs. Additionally, we designed separate search interfaces tailored to different query types, seamlessly switchable from the home screen for enhanced usability.

3 DATASET AND CHALLENGE FORMAT

3.1 Dataset Overview

The Lifelog Search Challenge 2024 reuses the dataset from previous challenges held in 2022 and 2023. The dataset consists of ~724K first-person images collected using a narrative clip device from a single lifelogger for an 18-month period during 2019-2020. All the images in the dataset are fully redacted and anonymized as per GDPR norms.

- **Visual Concepts:** For each image in the dataset, the visual concepts consist of information such as detected objects within the image, image caption along with caption confidence score, and text detected from images using off-the-shelf OCR models.
- **Metadata:** The metadata for LSC'24 is similar to previous years data consisting of information like biometrics (calories burnt, heart rate, step count, etc.), location, timezone, sleep information such as sleep stages as well as sleep efficiency and music data.

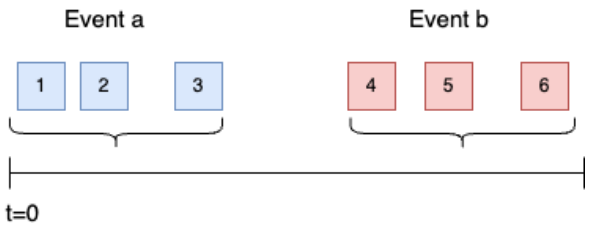

Figure 1: High level approach of event segmentation

Additional information regarding the dataset can be found in reference [9].

3.2 Challenge Overview

The Lifelog Search Challenge is a live annual benchmarking competition for lifelog information retrieval. The challenge comprises of 3 types of queries to evaluate the participating systems.

- **Known Item Search Query:** For this query type, users are prompted to pinpoint a particular moment within the lifelog dataset. Clues are progressively unveiled at 30-second intervals, requiring users to accurately submit one image within a set timeframe.
- **Ad-Hoc Query:** For Ad-Hoc queries, participants are tasked with submitting as many correct images as possible within a specified time frame for a given query. Submissions are assessed in real-time during the competition for this query type.
- **Question-Answering Query:** For this task, the objective is to provide a natural language response when presented with a query expressed in natural language.

4 SYSTEM OVERVIEW

This section discusses the primary components of our proposed system.

4.1 Event Segmentation

As discussed previously, the Lifelog dataset records a person's daily life experiences with the help of devices such as wearable cameras, bio-activity trackers, etc. The captured dataset at a granular level looks very similar to a camera roll which consists of images captured in chronological order accompanied by relevant metadata such as location, date-time, and(or) other bio-statistics. However, at a conceptual level, the dataset is a collection of events or moments from a person's life, where each event can consist of multiple images, and convey a single activity or group of related activities. Formally,

Subevent 1
Time: 10:00AM – 10:03 AM

Subevent 2
Time: 10:03AM – 10:05AM

Subevent 3
Time: 10:05AM – 11:30AM

Event Summary:
The lifelogger's day started with a visit to a coffee shop where he stood in front of the counter while a woman and a coffee machine were visible in the background. He then walked through a large brick building with a clock tower, passing by another person and through a hallway with white walls and a door. After that, the person attended a series of meetings and discussions with a group of people in various rooms, using laptops and computers to collaborate and work together. Throughout the day, he was seen enjoying a cup of coffee while working on his laptop, suggesting that it was a part of his daily routine.

Figure 2: Our event segmentation approach where initially the dataset in broken down into 'subevents' based on visual similarity. The subevents are then combined together heuristically resulting in larger coherent event consisting of inter-related activities. The image also displays the high-level event summary generated by the GPT 3.5 model.

an event in a person's life can be defined as a specific instance or occurrence within a defined spatiotemporal domain that has a broader objective or goal such as eating food, working in the office, cooking food, etc.

The larger objective of building retrieval systems for lifelog datasets is also retrieving events/moments and not simply retrieving images. We therefore devised an event segmentation approach that adopts a hierarchical methodology to segregate the dataset into individual events in a bottom-up fashion. Our approach solely leverages images captured from the wearable camera to establish event boundaries. A naive approach to determining event boundaries could be to chronologically parse the image dataset and determine the similarity of the adjacent images, where two adjacent dissimilar images could indicate the start of a new event. This simplistic approach fails to account for the nuances in the lifelog dataset and the major issue of constantly shifting point-of-view. For instance, an event that consists of having lunch with friends might have multiple POVs captured from the wearable camera such as chatting with friends, view of eating food or even view of just sitting.

Our event segmentation approach tries to tackle such individual POVs initially which we refer to as 'subevents' and then heuristically combine them into a single, larger coherent event. We use CLIP [17] embeddings to derive similarities between adjacent images. Figure 1 presents at a high level the methodology to establish event boundaries where the blue and red boxes numbered 1-6 are sub-events happening chronologically. The sub-events 1-3 are heuristically combined to form a single larger event based on the time difference between the subevents, in a similar way subevents 4-6 are combined together to form a single event. The time difference between subevents 3 and 4 is higher than our empirically defined threshold, hence they are not clubbed together and rather act as the event boundary between Event a and Event b.

4.2 Narrative Generation from Lifelogs

We use the events obtained in Section 4.1 to generative event-wise summary/narrative of the lifelog dataset leveraging large language models. The narrative generation pipeline is 3-step process as outlined below:

- **Caption Generation from Lifelog images:** We initially generate captions for all the images the dataset using BLIP-2 [14].
- **Generate activity summaries for subevents:** The captions serve as inputs for the large language model, prompting it to produce a concise summary of the subevent. The prompt includes background information and task instructions. We generate two distinct summaries: one using OpenAI's GPT 3.5 turbo model [7] and the other using MistralAI's Mistral-7B model [13]. This approach aims to increase diversity in the final event summary, thereby enhancing the efficiency of question-answering capabilities on our system.
- **Generate events summary from using individual event summaries:** Ultimately, we utilize the sub-event summaries as input for the model to generate a comprehensive event summary, amalgamating the sub-events into a cohesive and logical narrative.

4.3 Retrieval Augmented Generation (RAG) for Question-Answering over Lifelogs

We implement a Retrieval Augmentation Generation pipeline leveraging the event summaries generated from Section 4.2. We leverage

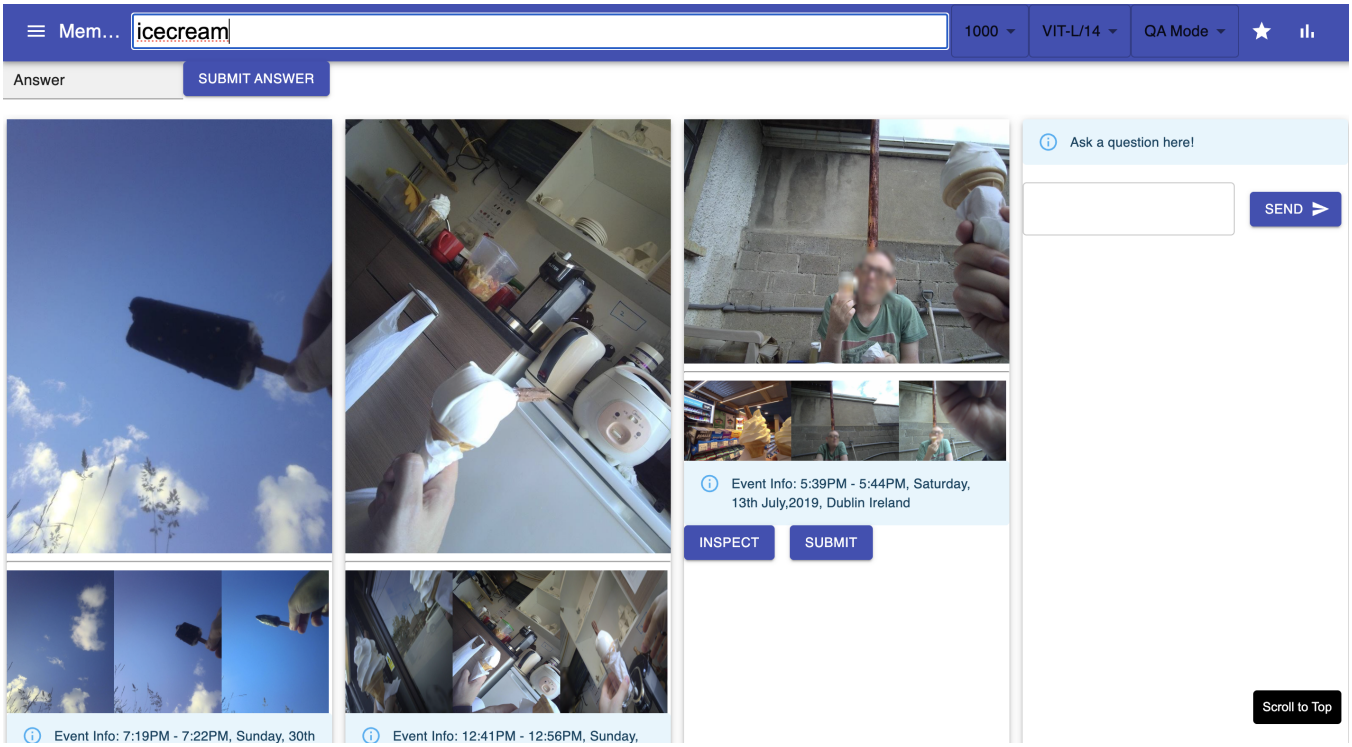

Figure 3: User Interface for question-answer queries

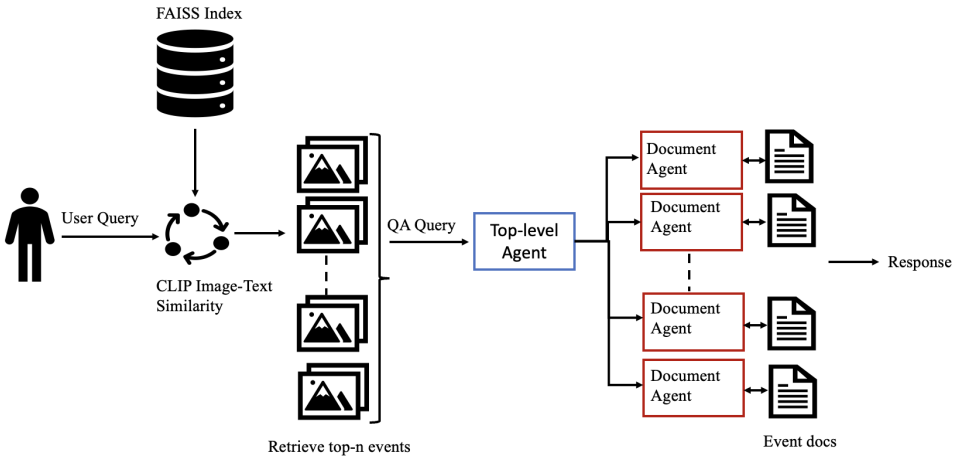

Figure 4: Memento 4.0: Search flow for question-answering query types.

the LlamaIndex framework to setup the pipeline to enable question-answering over the lifelog dataset.

The RAG pipeline's architecture models the task of question-answering over lifelogs as question-answering over a large corpus of individual documents. The documents in our use-case are the summarized events containing information such as date-time, location, activity summary etc. The pipeline follows a hierarchical architecture where initially each event document is assigned its own language model agent, which utilizes various tools to tackle specific problems. A language model agent can be defined as a model which uses a large language model as its central computational engine to reason through a problem, plan to solve the problem and use a set of tools to solve it [6]. The agents for individual events (documents) are supervised by a top-level agent which interacts with all the agents lower down in the hierarchy to get the desired answer. Each of the document agent in turn have access to a set

of tools enabling them to carry out their intended tasks efficiently, details about which are discussed below,

- **Summary Query Engine:** The summary query engine is responsible for summarizing the contents of the document (event) and is useful where the query requires summarizing the document in order to generate an answer.
- **Vector Query Engine:** The vector query engine is useful when answering factual questions about a particular event or answering questions related to specific aspects of an event e.g. the date, start and end time, or what happened during that event etc.
- **Sub Question Query Engine:** This tool handles the problem of answering a complex query using multiple data sources. It first breaks down the complex query into sub questions for each relevant data source, then gather all the intermediate responses and synthesizes a final response.

4.4 Search Flow and User Interface

The search flow for our proposed system largely remains similar to last year's system with the exception of question-answering queries. The newer system, like its predecessor systems, use CLIP [17] image-text embeddings for its search and ranking functionality. Figure 4 shows a high-level overview of the search flow for question answering queries. For QA queries the search flow is a a 2-step process as shown in Figure 4 : initially, the CLIP model-based search backend handles the query, narrowing down the relevant subset of data likely to contain the target information. Subsequently, this subset is processed by the QA system, which takes the QA query and produces the final response. However, the two-step process may be omitted in cases where users are confident about their information needs, such as querying about specific events like *What happened early morning on Christmas?*. In such instances, a vector search is initiated using only event documents, utilizing information such as date, time, and activity summaries to reach the final answer.

The user interface of the system has further been modified to display the search results in the form of 'events' unlike the previous versions of our systems Memento [1, 2, 4], where the unit of search was a single image from the lifelog dataset. Additionally for QA query types, the UI provides a chat interface to facilitate question-answering over the fetched results.

5 CONCLUSION AND FUTURE WORK

In this study, we introduce Memento 4.0, a prototype conversational search system tailored for lifelog information retrieval. Our approach involves the development of a hierarchical event segmentation technique to autonomously partition the lifelog dataset into manageable chunks or events. Additionally, we employ large language models to create activity summaries for these events, which in turn serve as the foundation for constructing a question-answering pipeline across the dataset. In future work, we aim to enhance this system into a seamless end-to-end conversational search solution, eliminating the need for manual intervention entirely.

ACKNOWLEDGMENTS

This publication has emanated from research supported by Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289_P2, co-funded by the European Regional Development Fund.

REFERENCES

[1] Naushad Alam and Yvette Graham. 2023. Memento: a prototype search engine for LSC 2021. *Multimedia Tools and Applications* (April 2023). https://doi.org/10.1007/s11042-023-15067-9

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

[3] Naushad Alam, Yvette Graham, and Cathal Gurrin. 2023. Memento 3.0: An Enhanced Lifelog Search Engine for LSC'23. In *Proceedings of the 6th Annual ACM Lifelog Search Challenge*. ACM, Thessaloniki Greece, 41–46. https://doi.org/10.1145/3592573.3593103

[4] Naushad Alam, Yvette Graham, and Cathal Gurrin. 2023. Memento 3.0: An Enhanced Lifelog Search Engine for LSC'23. In *Proceedings of the 6th Annual ACM Lifelog Search Challenge*. ACM, Thessaloniki Greece, 41–46. https://doi.org/10.1145/3592573.3593103

[5] Ahmed Alateeq, Mark Roantree, and Cathal Gurrin. 2023. Voxento 4.0: A More Flexible Visualisation and Control for Lifelogs. In *Proceedings of the 6th Annual ACM Lifelog Search Challenge*. ACM, Thessaloniki Greece, 7–12. https://doi.org/10.1145/3592573.3593097

[6] Jacob Andreas. 2022. Language Models as Agent Models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 5769–5779. https://doi.org/10.18653/v1/2022.findings-emnlp.423

[7] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. http://arxiv.org/abs/2005.14165 arXiv:2005.14165 [cs].

[8] Cathal Gurrin, Hideo Joho, Frank Hopfgartner, Liting Zhou, Van-Tu Ninh, Tu-Khiem Le, Rami Albatal, Duc-Tien Dang-Nguyen, and Graham Healy. 2019. Overview of the NTCIR-14 Lifelog-3 Task. (2019), 13.

[9] 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 *Proc. International Conference on Multimedia Retrieval (ICMR'24)* (Phuket, Thailand) *(ICMR '24)*. New York, NY, USA. https://doi.org/10.1145/3652583.3658891

[10] Cathal Gurrin, Björn Þór Jónsson, Klaus Schöffmann, Duc-Tien Dang-Nguyen, Jakub Lokoč, Minh-Triet Tran, Wolfgang Hürst, Luca Rossetto, and Graham Healy. 2023. Introduction to the Sixth Annual Lifelog Search Challenge, LSC'23. In *Proc. International Conference on Multimedia Retrieval (ICMR'23)* (Thessaloniki, Greece) *(ICMR '23)*. Association for Computing Machinery, New York, NY, USA.

[11] Nhat Hoang-Xuan, Thang-Long Nguyen-Ho, Cathal Gurrin, and Minh-Triet Tran. 2023. Lifelog Discovery Assistant: Suggesting Prompts and Indexing Event Sequences for FIRST at LSC 2023. In *Proceedings of the 6th Annual ACM Lifelog Search Challenge*. ACM, Thessaloniki Greece, 47–52. https://doi.org/10.1145/3592573.3593104

[12] Maria Tysse Hordvik, Julie Sophie Teilstad Østby, Manoj Kesavulu, Thao-Nhu Nguyen, Tu-Khiem Le, and Duc-Tien Dang-Nguyen. 2023. LifeLens: Transforming Lifelog Search with Innovative UX/UI Design. In *Proceedings of the 6th Annual ACM Lifelog Search Challenge*. ACM, Thessaloniki Greece, 1–6. https://doi.org/10.1145/3592573.3593096

[13] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. http://arxiv.org/abs/2310.06825 arXiv:2310.06825 [cs].

[14] 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. http://arxiv.org/abs/2301.12597 arXiv:2301.12597 [cs].

[15] Thao-Nhu Nguyen, Tu-Khiem Le, Van-Tu Ninh, Cathal Gurrin, Minh-Triet Tran, Thanh Binh Nguyen, Graham Healy, Annalina Caputo, and Sinead Smyth. 2023. E-LifeSeeker: An Interactive Lifelog Search Engine for LSC'23. In *Proceedings of the 6th Annual ACM Lifelog Search Challenge*. ACM, Thessaloniki Greece, 13–17.

https://doi.org/10.1145/3592573.3593098

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

[17] 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]* (Feb. 2021). http://arxiv.org/abs/2103.00020 arXiv: 2103.00020.

[18] Ricardo Ribeiro, Luísa Amaral, Wei Ye, Alina Trifan, António J. R. Neves, and Pedro Iglésias. 2023. MEMORIA: A Memory Enhancement and MOment RetrIeval Application for LSC 2023. In *Proceedings of the 6th Annual ACM Lifelog Search Challenge*. ACM, Thessaloniki Greece, 18–23. https://doi.org/10.1145/3592573.3593099

[19] Klaus Schoeffmann. 2023. lifeXplore at the Lifelog Search Challenge 2023. In *Proceedings of the 6th Annual ACM Lifelog Search Challenge*. ACM, Thessaloniki Greece, 53–58. https://doi.org/10.1145/3592573.3593105

[20] 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*. ACM, Thessaloniki Greece, 24–29. https://doi.org/10.1145/3592573.3593100

[21] Quang-Linh Tran, Ly-Duyen Tran, Binh Nguyen, and Cathal Gurrin. 2023. MemoriEase: An Interactive Lifelog Retrieval System for LSC'23. In *Proceedings of the 6th Annual ACM Lifelog Search Challenge*. ACM, Thessaloniki Greece, 30–35. https://doi.org/10.1145/3592573.3593101

[22] Liting Zhou, Cathal Gurrin, Graham Healy, Hideo Joho, Thanh-Binh Nguyen, Rami Albatal, and Frank Hopfgartner. 2022. Overview of the NTCIR-16 Lifelog-4 Task. In *Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies (NTCIR-16)*. Tokyo, Japan.

[23] Liting Zhou, Graham Healy, Cathal Gurrin, Ly-Duyen Tran, Naushad Alam, Hideo Joho, Longyue Wang, Tianbo Ji, Chenyang Lyu, and Duc-Tien Dang-Nguyen. 2023. Overview of the NTCIR-17 Lifelog-5 Task. https://doi.org/10.20736/0002001329