OpenLifelogQA: An Open-Ended Multimodal Lifelog Question-Answering Dataset

Quang-Linh Tran[1(✉)], Hoang-Bao Le[1], Tuong-Nghiem Diep[2], Binh Nguyen[2], Gareth J. F. Jones[1], and Cathal Gurrin[1(✉)]

[1] ADAPT Centre, School of Computing, Dublin City University, Dublin, Ireland
{quang-linh.tran2,bao.le2}@mail.dcu.ie,
{gareth.jones,cathal.gurrin}@dcu.ie
[2] University of Science, Vietnam National University, Ho Chi Minh City, Vietnam
dtnghiem21@apcs.fitus.edu.vn, ngtbinh@hcmus.edu.vn

Abstract. We introduce **OpenLifelogQA**, a large-scale open-ended lifelog QA dataset constructed from 18 months of multimodal lifelog data. Lifelogging is the passive collection and analysis of personal daily activities using wearable devices, producing rich multimodal data such as images, locations, and biometrics. Question answering (QA) over lifelog data enables users to interactively query their own experiences, supporting applications in memory support, lifestyle analysis, and personal assistance. OpenLifelogQA contains 14,187 Q&A pairs spanning multiple question types and difficulty levels, designed to support robust evaluation in realistic settings. Compared with prior resources, OpenLifelogQA offers greater diversity and practicality for real-world applications. To establish baselines, we evaluate the LLaVA-NeXT-Interleave 7B model, achieving 89.7% BERTScore, 25.87% ROUGE-L, and an average LLM Score of 3.97. By releasing OpenLifelogQA, we aim to promote future research on lifelog technologies, paving the way for personal lifelog assistants capable of memory augmentation, healthcare support, and lifestyle coaching.

Keywords: Lifelog Question Answering · Multi-modal Question Answering Dataset · Large Language Models

1 Introduction

Lifelogging refers to the passive collection, storage, and analysis of personal daily life data using wearable devices [8]. Such data typically includes point-of-view (PoV) images from wearable cameras, biometrics, location traces, and other digital records such as emails or internet activity. Lifelog data has been widely studied for applications including memory archiving and enhancement [12,25], health monitoring [13], and lifestyle analysis [14] [18]. Among these, lifelog retrieval—finding specific moments in lifelog data—has been an active

W. Buntine et al. (Eds.): SOICT 2025, CCIS 2912, pp. 189–201, 2027.
https://doi.org/10.1007/978-981-92-2584-2_15

research topic for many years [9,22,23], helping users recover forgotten events. However, retrieval alone is insufficient to satisfy users' information needs, since many insights remain hidden within the retrieved data. This motivates the development of resources to enable question answering (QA) over lifelog data.

Lifelog QA aims to provide precise answers to user questions based on lifelog data. Unlike retrieval, which only identifies relevant content, QA must extract and reason over specific details within that content. For example, the question "What did I have for lunch on May 01, 2025, and for how long?" requires first identifying lunch-related activities, then analyzing associated images and metadata to determine both the food consumed and its duration. This task poses several challenges: (i) the long-term, large-scale nature of lifelog data requires robust retrieval to locate relevant events; (ii) many queries demand complex, multi-hop reasoning; and (iii) the inherently multimodal nature of lifelog data—images, timestamps, and locations—requires models to interpret heterogeneous information effectively. Despite these challenges, lifelog QA has enormous potential. It could allow users to query their lifelogs for purposes such as fact verification, lifestyle analysis, or memory enhancement [21]. For example, "Did I take any medicine yesterday?" is a practical reminder-type query, while "Tell me what I discussed with J. during our last meeting" points toward future context-aware assistants capable of richer personal interactions.

Given these potential benefits, lifelog QA promises to significantly enhance how users interact with their personal data. However, progress has been constrained by the lack of high-quality datasets. Prior work has produced either small-scale collections [20] or large but synthetic lifelog QA datasets [19], both of which limit diversity and realism for real-world applications. Moreover, with the rise of generative AI, open-ended QA—where answers are expressed naturally rather than as spans or multiple-choice—offers a more realistic interaction paradigm. To address these gaps, this paper introduces a novel open-ended lifelog QA dataset, built on a large-scale, realistic collection from the Lifelog Search Challenge [9]. Our dataset contains diverse question types reflecting practical use cases, laying the foundation for robust evaluation and the development of models that can more effectively understand and interact with real-world lifelog data.

2 Related Work

Before discussing datasets specifically in the field of lifelog QA, we outline details of some datasets in Egocentric Video QA. Video QA shares some characteristics with lifelog QA in using PoV data and focusing on life-related questioning. QaEgo4D [2] is the largest human-curated Egocentric Video QA dataset with 1,325 videos and 14,507 QA pairs. This dataset focuses on videos of long duration and is constructed based on the Ego4D dataset [6]. There are other short video duration datasets such as OpenEQA [17], EgoVQA [5], and EgoTaskQA [11]. However, even the longest video in these datasets cannot be compared to the length of a set of daily lifelog images, if these images were assembled into

a video. In addition, lifelog data includes not only images, but also other data such as time, location, and biometrics, so lifelog QA can be both visual-related and metadata-related. These are the key differences between Egocentric Video QA and Lifelog QA.

In the domain of lifelog QA, three notable datasets have been developed, each with distinct approaches and limitations. The LLQA dataset [20] created a lifelog QA task by augmenting the Lifelog Search Challenge 2020 (LSC'20) dataset [7] with around 15,000 semi-automatically generated QA pairs. This dataset had only 85 days of lifelog data and was constructed on a closed QA format with two main types of questions (multiple-choice and yes/no). Meanwhile, TimelineQA [19] takes a template-based approach to synthetically generate a massive dataset of 128 million lifelog entries. The approach first creates an imaginary persona with key information such as gender, profession, and then generates life episodes (events) with different time densities, such as daily, weekly, monthly and yearly. Finally, the QA pairs are generated based on the information of episodes. While impressive in scale, its synthetic and text-only nature reduces its realism and limits its applicability to real-world use cases and its value for research. The third dataset is MemoriQA [24], an open-ended lifelog QA dataset combining manual and synthetic QA generation. This dataset is built on 61 days of lifelog data and contains 3,644 QA pairs. This is a relatively small dataset, but it is built on a practical approach. Across all three datasets, several challenges remain from heavy dependence on synthetic data to small-sized data due to the high cost of manual annotation. This challenge restricts the diversity and practicality of the QA pairs.

From the previous datasets, we observe a key drawback in the lack of practicality and diversity in lifelog QA. In our current work, we have constructed a broad, diverse lifelog QA dataset based on an existing large lifelog dataset from Lifelog Search Challenge [9] to ensure diversity and practicality with open-ended Q&A on different types of questions. In addition, with the rise of large language models (LLMs) in recent years, generative and open-ended QA can help to increase the human-likeness of answers. We use them as a parallel annotator to human annotators to increase the efficiency of Q&A generation.

3 Dataset Construction

In this section, we provide information about the process of constructing our dataset. We start with the source of lifelog data and then the annotation process. Finally, we describe the quality control process to measure the quality of Q&A generation.

3.1 Data Source

We use the lifelog dataset from Lifelog Search Challenge 2024 (LSC'24) [9] as the starting point to create the QA dataset. It contains 722,606 PoV images from a single anonymous lifelogger, ranging over a period of 18 months from

January 2019 to June 2020. The images are captured from a PoV camera[1] worn on the chest of the lifelogger. The images have a resolution of $1{,}024 \times 768$ pixels. All faces and readable text have been removed, and certain scenes and activities have been manually filtered out to respect local privacy requirements. In addition to the images, we also use the metadata of the lifelog collection. This includes the capture time of each image, the capture location of each image, and every minute if there is no image captured at that time, and related biometrics from a smartwatch. From this data source, we perform the QA generation described in the next section.

3.2 QA Generation Process

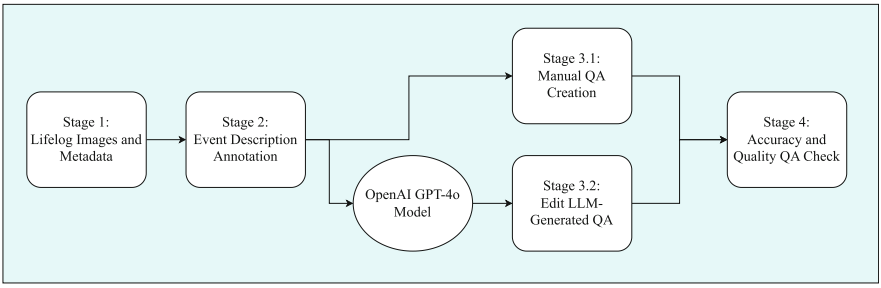


Fig. 1. Annotation Process.

From the lifelog data source, and in line with prior lifelog QA datasets, we first generated *event descriptions* to serve as the basis for questions and answers. To support this process, we recruited 10 undergraduate volunteers majoring in Data Science and Information Technology. Their English proficiency ranged from good to proficient (TOEIC $\geq$ 650), and they possessed basic knowledge of QA tasks. The volunteers participated in two training sessions covering lifelog concepts, lifelog QA, and the annotation workflow. In addition, a comprehensive guideline was provided to describe the overall procedure and address common annotation issues.

We performed QA generation month by month, beginning with January 2019. Each volunteer was assigned four days of lifelog data per month, with one day overlapping among all volunteers to assess the consistency of event description generation. The process consisted of four stages, as illustrated in Fig. 1. An annotation platform was developed to support these tasks.

– **Stage 1: Event Review.** Volunteers reviewed all lifelog images and associated metadata for a given day to obtain an overview of the lifelogger's activities.

[1] https://getnarrative.com/.

Table 1. Statistics on three sets of OpenLifelogQA.

Set	# atomic question	# temporal question	# aggregation question	# context events	Avg questionlength	Avg answer length
Training	6578	946	3653	1.64	12.74	6.67
Validation	895	131	476	1.75	11.99	5.74
Testing	882	156	470	1.75	12.23	5.73

- **Stage 2: Event Description.** Volunteers segmented the data into events and wrote first-person descriptions for each. An event was defined as a sequence of images representing a single activity (e.g., eating, watching TV, or working on a laptop). Each event was associated with a start time and a representative image, while the end time was defined by the start of the next event. For example: *"I am working on my laptop"* at 2019-01-01 12:23:42, with ImageID `20190101_122342_000`.
- **Stage 3: QA Generation.** Both volunteers and the GPT-4o model [10] generated QA pairs from event descriptions. Each volunteer created 10 QA pairs per day, while GPT-4o generated 20 QA pairs, balancing quality, scalability, and human effort. Every QA pair was linked to one or more events from Stage 2, with corresponding event IDs (the ImageID of the event's start time) recorded as ground truth for retrieval tasks. The QA format was deliberately unconstrained to encourage diversity and practicality. We particularly emphasised complex aggregation questions requiring reasoning or calculation (e.g., counting or measuring duration). For instance: *"How long did I drive from my house to the church on the morning of January 21, 2020?"*.
- **Stage 4: Quality Control.** Volunteers conducted a final review to verify the accuracy of QA pairs and event descriptions. One day per month was annotated by all volunteers to measure consistency in event description creation. We measure the BERT Score [27] between event descriptions generated by annotators and ensure a threshold of 0.85 BERT Score. This score indicates a high semantic relevance and agreement of annotators on event descriptions. Since every annotator is freely generating QA, so we cannot measure the inner annotator agreement. Instead, volunteers cross-reviewed and corrected GPT-4o outputs, resolving errors, duplicates, and inconsistencies before finalizing them as official QA. There are about 200 QAs are corrected after the cross-check.

4 Dataset Information

In this section, we provide information about the OpenLifelogQA dataset after the QA generation process. We describe the information and statistics of the dataset, as well as some analysis of questions and answers. Finally, we discuss licensing and ethical & privacy considerations arising from this, and any similar lifelog dataset.

Table 2. Lifelog QA datasets comparison.

Dataset	# QA	Lifelog data	Question format	# Event description	Image	Time & Location	Open-sourced
LLQA [20]	15,065	85 days	Multiple-choice	None	✓	✓	✓
Timeline QA [19]	600,000	Synthetic	Open-ended	128,023,476		✓	✓
MemoriQA [24]	3,644	61 days	Open-ended	1,925	✓	✓	
OpenLifelogQA	14,187	514 days	Open-ended	27,705	✓	✓	✓

Table 3. Some QA pairs and ground-truth context examples

Question	Answer	Context	Type
How long did I spend at the store on the evening of June 26, 2020?	I spent 23 min at the store.	20200626_203417_000: I am at a store, 20200626_205727_000: I am driving away from the store	Aggregation
How many times did I watch TV on June 1, 2020?	I watched TV 3 times	20200601_115751_000: I am watching TV. 20200601_165938_000: After dinner, I started watching TV. 20200601_182032_000: I stopped working and started watching TV again	Aggregation
When did I go to the kitchen to make a smoothie on June 21, 2020?	I went at 11:02 AM	20200621_110203_000: I go to the kitchen to make a smoothie	Atomic
What did I do before preparing to leave the house on June 21, 2020?	I watched TV	20200621_061246_000: I continue to sit and watch TV 20200621_062502_000: I have left the house and am moving by car	Temporal

4.1 Descriptions

After the QA generation process, we obtained a lifelog QA dataset with 27,705 event descriptions and 14,187 QA pairs. This data spans 514 days of lifelog data, which means an average of 54 events and 28 questions per day. The events and QA pairs are independent, as a QA pair can involve between 1 and several events, and some events can relate to no QA. The average number of events for each QA pair is 1.66 events. Table 3 shows some examples of QA pairs and groundtruth contexts of the OpenLifelogQA dataset. We also compare this dataset with previous datasets in Table 2. Our proposed OpenLifelogQA dataset offers a more comprehensive resource for lifelog QA research. This dataset includes a richer set of events and more diverse lifelog data than previous datasets, except for the synthetic data in Timeline QA. It also incorporates open-ended questions, as well as images, time, and location metadata.

We split the dataset into three sets for model development and evaluation, including training, validation and test sets. The splits were determined by the month of lifelog data, April and May 2020 is used for validation, March and June 2020 for testing, and the rest for training. This splitting strategy aims to avoid data leakage in model development and evaluation. Statistics of the three datasets are shown in Table 1. The training set contains 11,117 QA pairs, while the validation and testing sets each include approximately 1,500 QA pairs. An average question has about 12 words, and answers 6 words. This length suggests concise, direct questions and answers in real-world applications. We also categorise the questions into several types, which are described in the following section.

4.2 Analysis

The questions are divided into three categories: atomic, temporal, and aggregation, following classifications used in prior studies [19].

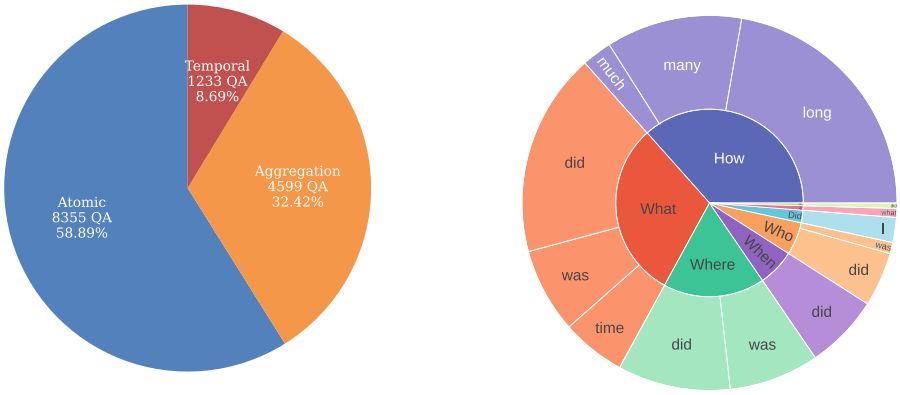
- **Atomic questions** can be answered using a single context, such as "*What time did I have lunch yesterday?*".
- **Temporal questions** involve reasoning about the sequence of events in lifelog data, asking about what happened before or after a known event, such as "*What did I do after having lunch yesterday?*".
- **Aggregation questions** require synthesizing information across multiple events to derive an answer, such as calculating a duration, count, or total. An example is: "*How long did I have lunch yesterday?*".

The distribution of these types of questions is illustrated in Fig. 2a. We further analyse the distribution of question words in Fig. 2b. The distribution of question words shows that the dataset reflects realistic lifelog use cases, with a strong focus on "what", "where", "when" and "how" questions, and less attention to more exploratory or uncommon queries.

4.3 Licensing and Access

For the purpose of fostering research in lifelog QA, we release the dataset for research purposes only. The dataset is available under a Creative Commons Attribution-NonCommercial 4.0 International License. Anyone interested in this dataset can contact the authors of this paper to complete a user agreement form to obtain the event descriptions and QA pairs of the OpenLifelogQA dataset. The images and other metadata from the LSC lifelog dataset are provided separately at the main page of Lifelog Search Challenge[2]. This is to ensure the privacy and ethical issues in lifelog data. For the reviewers of this dataset paper, we provide a link to the dataset to review the content of the dataset as follows: Link.

[2] http://lifelogsearch.org/lsc/lsc_data/.



(a) Question type distribution (b) Question word distribution

Fig. 2. Analysis on questions.

4.4 Ethical Considerations and Privacy

As lifelog data is private and prone to many privacy problems due to the highly sensitive nature of personal data, which can include health details, social interactions, and location history, the use of this dataset should be considered in an ethical and privacy context. The data source of LSC'24 has already removed all faces and readable text in the lifelog images to ensure privacy for bystanders. Moreover, any analysis or model training on lifelog data should avoid generating intrusive inferences that could compromise individual autonomy or dignity. Ethical oversight and compliance with data protection regulations such as GDPR are essential to maintain public trust and ensure responsible research practices.

5 Baseline Experiment

In this section, we run a baseline experiment on the OpenLifelogQA dataset with a well-known and easy-to-reproduce multi-modal LLM, LLaVA-NeXT-Interleave. The experimental setting and results, as well as some analysis, are discussed in the following sections.

5.1 Settings

From the test set of OpenLifelogQA, we use the LLaVA-NeXT-Interleave model [15] to predict the answers to questions without finetuning. This is to measure the performance of a multi-modal LLM on lifelog data. The LLaVA-NeXT-Interleave model is a multi-modal LLM with the Qwen [1] as the base LLM model and Siglip [26] as the base vision tower. This model is fine-tuned with visual instructions on multi-image, video, 3D, and multi-patch of single images. The model shows leading results in various benchmarks as well as several emerging capabilities.

With a strong performance on various tasks, we use this model as the baseline to run on the OpenLifelogQA dataset. We use two versions of the model, which are Qwen 7B and 0.5B, to see the effect of model size on the performance on a test set of OpenLifelogQA. We do not fine-tune the model on our dataset due to the limited computing resources, as well as to measure the performance of transferring tasks from multi-modal data to lifelog data of LLaVA-NeXT-Interleave.

Table 4. Predicted answers versus ground-truth answers and LLM Score.

Question	Ground-truth Answer	Predicted Answer	LLM Score
How long did I spend at the store on the evening of June 26, 2020?	I spent 23 min at the store	I spent 17 min at the store on the evening of June 26, 2020	3
How many times did I watch TV on June 1, 2020?	I watched TV 3 times	I watched TV three times on June 1, 2020	5
When did I go to the kitchen to make a smoothie on June 21, 2020?	I went at 11:02 AM	I went to the kitchen to make a smoothie on June 21, 2020 at 11:02 AM	5
What did I do before preparing to leave the house on June 21, 2020?	I watched TV	I was watching TV at home	5

We provide the question, groundtruth contexts, and the images of these contexts, as some examples in Fig. 3, to the model and ask for an answer. We form a prompt with a simple instruction: "Please act as a life assistant to answer questions based on the given context. You are given a question about life and a context describing time, location, and a description of an event involving a person. Your goal is to provide an answer to that question based on deriving information from the context. Only provide the answer from the given context and derive the answer carefully. The time in the context is the start time of the event, and the end time of the event is the start time of the next event. Thinking step by step to derive the answer. Provide a short answer with no explanation. Context: <image1> <context1>, ... <imageN> <contextN>. Question: <question>. Answer: ". We have configured some hyperparameters, such as temperature = 0, top_p = 0. All the experiments are run on a single A100 GPU. This experiment is just a baseline to evaluate the effectiveness of multi-modal LLM models for lifelog QA. In real-world applications, the contexts are not always available, and there is a need to retrieve the correct contexts for the QA model.

We use three metrics of three evaluation approaches for Open-ended QA to evaluate the performance of the model, including lexical matching, semantic matching, and LLM-based evaluation. Lexical matching focuses on word-level overlap between predicted and ground truth answers, using metrics such as Exact

Match, F1 score, and ROUGE [16], which measures n-gram and sequence similarity. Although effective, lexical metrics may overlook semantically correct answers with different phrasing. To address this, semantic matching evaluates the underlying meaning using metrics like BERT Score [27], which compares BERT-based embeddings [4] of predicted and reference answers. LLM-based evaluation [3] involves using large language models like GPT-4o [10] to assign a score from 1 to 5 based on answer correctness. In this work, we adopt ROUGE, BERT Score, and LLM Score evaluation to comprehensively evaluate model performance.

5.2 Results

Table 5 shows the experimental results of the LLaVa-NeXT-Interleave model on the test set. We can see that the 7B model gives a higher LLM Score at 3.9665 than the 0.5B at 3.3873. However, the 0.5B provides a better performance at BERT Score and ROUGE-L at 90.63% and 26.13%, respectively. This can be explained since the 0.5B model provides predicted answers that match lexically and semantically with the groundtruth answers, but does not truly give the most accurate answer. For example, the predicted answer: "I drive to school in 25 min" and the groundtruth answer "I drive to school in 5 min" have a very high BERT Score and ROUGE-L due to the match of words and meaning, but the main important information 25 min and 5 min are wrong so LLM Score is low. The overall performance of the two models shows that the multi-modal LLM can solve this task with high performance when the correct contexts are provided. However, the words of predicted and groundtruth answers can be different, resulting in a low ROUGE-L score, but this can be improved using instruction tuning on the training set.

Table 5. Experimental results from the LLaVa-NeXT-Interleave models on the test set.

# Parameters	BERT Score	ROUGE-L	LLM Score
0.5B	0.9063	0.2613	3.3873
7B	0.8970	0.2587	3.9665

5.3 Analysis

Table 4 shows some qualitative examples of predicted answers from the 7B model versus groundtruth answers and their score. We can see that the model mostly generates correct answers, but using different words and with more information than needed in the groundtruth answers. Only the first predicted answer is wrong due to the time calculation of the model, also there is an event of buying food in the store that confuses the model about the time of leaving the store.

Table 6. Performance of 7B model on different question types

Question type	BERT Score	ROUGE-L	LLM Score
Atomic	0.8938	0.2572	4.2574
Temporal	0.8935	0.2627	4.2115
Aggregation	0.9042	0.26	3.3362

We also analyse the performance of LLaVa-NeXT-Interleave 7B model on different types of questions. Results are shown in Table 6. Atomic is the easiest question to solve, with an LLM Score of 4.2574, but it has the lowest ROUGE-L due to the diversity in words in the answer. Aggregation is the most difficult question because it requires calculation or counting in the context, so the LLM Score is only 3.3362. However, it has the highest BERT Score and the second highest ROUGE-L because its answers mainly contain the questions with a number. This analysis highlights the challenge of aggregation questions in the lifelog QA problem.

6 Conclusion

In this paper, we introduced **OpenLifelogQA**, a large-scale open-ended QA dataset for lifelog data. The dataset contains 14,187 QA pairs and over 27K event descriptions collected from 18 months of lifelogging with more than 725K images and metadata. Unlike prior resources, OpenLifelogQA combines *scale, diversity, and realism*, offering a strong foundation for practical lifelog applications. Our dataset captures a wide range of high-level daily life questions, supporting tasks such as lifestyle analysis, memory exploration, and personal data interaction.

We also conducted baseline experiments with the LLaVA-NeXT-Interleave model, which achieved 25.87% ROUGE-L and a 3.97 LLM Score without fine-tuning. These results underscore both the promise and the difficulty of lifelog QA, particularly for reasoning-intensive aggregation questions.

Looking ahead, we plan to extend our work by exploring temporal reasoning and chain-of-thought methods to better address complex queries. More broadly, we envision OpenLifelogQA enabling future **personal lifelog assistants** capable of context-aware memory aids, healthcare monitoring, and lifestyle coaching. By releasing this dataset, we aim to foster research that brings lifelog technologies closer to practical, human-centered applications.

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] and [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. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., et al.: Qwen technical report (2023). https://arxiv.org/abs/2309.16609
2. Bärmann, L., Waibel, A.: Where did i leave my keys?-Episodic-memory-based question answering on egocentric videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1560–1568 (2022)
3. Desmond, M., Ashktorab, Z., Pan, Q., Dugan, C., Johnson, J.M.: EvaluLLM: LLM assisted evaluation of generative outputs. In: Companion Proceedings of the 29th International Conference on Intelligent User Interfaces, IUI '24 Companion, pp. 30–32. Association for Computing Machinery, New York (2024). https://doi.org/10.1145/3640544.3645216
4. Devlin, J.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint: arXiv:1810.04805 (2018)
5. Fan, C.: EgoVQA-an egocentric video question answering benchmark dataset. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, pp. 0–0 (2019)
6. Grauman, K., et al.: Ego4D: around the world in 3,000 hours of egocentric video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18995–19012 (2022)
7. Gurrin, C., et al.: Introduction to the third annual lifelog search challenge (LSC'20). In: Proceedings of the 2020 International Conference on Multimedia Retrieval, pp. 584–585 (2020)
8. Gurrin, C., Smeaton, A.F., Doherty, A.R.: LifeLogging: Personal Big Data. Found. Trends Inf. Retrieval **8**(11), 1–125 (2014). https://doi.org/10.1561/1500000033
9. Gurrin, C., et al.: Introduction to the seventh annual lifelog search challenge, LSC'24. In: Proceedings of the 2024 International Conference on Multimedia Retrieval, ICMR '24, pp. 1334–1335. Association for Computing Machinery, New York (2024). https://doi.org/10.1145/3652583.3658891
10. Hurst, A., et al.: GPT-4o system card. arXiv preprint: arXiv:2410.21276 (2024)
11. Jia, B., Lei, T., Zhu, S.C., Huang, S.: EgoTaskQA: understanding human tasks in egocentric videos. In: Advances in Neural Information Processing Systems, vol. 35, pp. 3343–3360 (2022)
12. Kikhia, B., Hallberg, J., Bengtsson, J.E., Savenstedt, S., Synnes, K.: Building digital life stories for memory support. Int. J. Comput. Healthc. **1**(2), 161–176 (2010)
13. Kim, S., Yeom, S., Kwon, O.J., Shin, D., Shin, D.: Ubiquitous healthcare system for analysis of chronic patients' biological and lifelog data. IEEE Access **6**, 8909–8915 (2018). https://doi.org/10.1109/ACCESS.2018.2805304
14. Kumar, G., Jerbi, H., Gurrin, C., O'Mahony, M.P.: Towards activity recommendation from lifelogs. In: Proceedings of the 16th International Conference on Information Integration and Web-based Applications & Services, pp. 87–96 (2014)
15. Li, F., et al.: LLaVA-next-interleave: tackling multi-image, video, and 3D in large multimodal models (2024). https://arxiv.org/abs/2407.07895
16. Lin, C.Y.: ROUGE: a package for automatic evaluation of summaries. In: Text Summarization Branches Out, pp. 74–81. Association for Computational Linguistics, Barcelona (2004). https://aclanthology.org/W04-1013
17. Majumdar, A., et al.: OpenEQA: embodied question answering in the era of foundation models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16488–16498 (2024)

18. Rossetto, L., et al.: The castle 2024 dataset: advancing the art of multimodal understanding. arXiv preprint: arXiv:2503.17116 (2025)
19. Tan, W.C., et al.: TimelineQA: a benchmark for question answering over timelines. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Findings of the Association for Computational Linguistics: ACL 2023. pp. 77–91. Association for Computational Linguistics, Toronto (2023). https://doi.org/10.18653/v1/2023.findings-acl.6
20. Tran, L.D., Ho, T.C., Pham, L.A., Nguyen, B., Gurrin, C., Zhou, L.: LlQA-lifelog question answering dataset. In: International Conference on Multimedia Modeling, pp. 217–228. Springer (2022)
21. Tran, L.D., Zhou, L., Nguyen, B., Gurrin, C.: Interactive question answering for multimodal lifelog retrieval. In: International Conference on Multimedia Modeling, pp. 68–81. Springer (2024)
22. Tran, Q.L., Nguyen, B., Jones, G.J.F., Gurrin, C.: Memoriease 2.0: A conversational lifelog retrieve system for LSC'24. In: Proceedings of the 7th Annual ACM Workshop on the Lifelog Search Challenge, LSC '24, pp. 12–17. Association for Computing Machinery, New York, NY, USA (2024).https://doi.org/10.1145/3643489.3661114
23. Tran, Q.L., Nguyen, B., Jones, G.J.F., Gurrin, C.: Memoriease 3.0: a rag-enhanced conversational lifelog retrieval system at LSC'25. In: Proceedings of the 8th Annual ACM Workshop on the Lifelog Search Challenge, LSC '25, pp. 52–57. Association for Computing Machinery (2025). https://doi.org/10.1145/3729459.3748689
24. Tran, Q.L., Nguyen, B., Jones, G.J., Gurrin, C.: MemoriQA: a question-answering lifelog dataset. In: Proceedings of the 1st ACM Workshop on AI-Powered Q&A Systems for Multimedia, pp. 7–12 (2024)
25. Tran, Q.L., Tran, L.D., Nguyen, B., Gurrin, C.: MemoriEase: an interactive lifelog retrieval system for LSC'23. In: Proceedings of the 6th Annual ACM Lifelog Search Challenge, LSC '23, pp. 30–35. Association for Computing Machinery, New York (2023). https://doi.org/10.1145/3592573.3593101
26. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training (2023). https://arxiv.org/abs/2303.15343
27. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: BertScore: evaluating text generation with BERT (2020). https://arxiv.org/abs/1904.09675