Multi-modal Context Reranking for Lifelog Question Answering

Quang-Linh Tran*, Ly-Duyen Tran*, Binh Nguyen†, Gareth J. F. Jones*, Cathal Gurrin*
*ADAPT Centre, School of Computing, Dublin City University, Dublin, Ireland
linh.tran3@mail.dcu.ie, allie.tran@dcu.ie, gareth.jones@dcu.ie, cathal.gurrin@dcu.ie
†University of Science, Vietnam National University, Ho Chi Minh City, Vietnam
ngtbinh@hcmus.edu.vn

***Abstract*—Lifelog question-answering (QA) involves seeking answers to users' questions from within their personal lifelog. A lifelog consists of passively collected multimodal personal information from the owner's life experiences, including images, biometrics, geolocation data, and textual descriptions. Data in a lifelog can become vast, spanning the user's lifetime. QA for such large collections presents significant challenges: finding the most relevant lifelog events (contexts) that may contain the answer before generating a response. We propose a reranking model designed to improve retrieval accuracy by effectively ranking the relevant contexts. Our model integrates multimodal information from images and text, employing a combination of visual extractors and language models. We experiment with three visual extractor models, Vision Transformer, BLIP2, and CLIP, as well as three language models, namely BERT, MiniLM, and ModernBERT. Compared with a retrieval baseline using cosine similarity ranking from Stella-1.5B embeddings, our experimental results on two lifelog QA datasets demonstrate a substantial improvement. Recall@1 is observed to increase from 37.65% to 65.57% and Precision@1 from 52.65% to 85.88% when using the ModernBERT+ViT reranking model for the OpenLifelogQA task. These findings show the robustness of multimodal reranking in context selection for lifelog QA and provide a mechanism for accurate and efficient retrieval in lifelog applications.**

***Index Terms*—Reranking, Lifelog Question Answering, Cross-Encoders, Information Retrieval**

I. Introduction

Lifelogging is the process of automatically collecting information about a person's daily life by a so-called lifelogger from multiple sensors [1]. Lifelog data is taken from multiple sources, including, for example, Point-of-View (PoV) images/videos, biometrics, GPS, and internet footprints. This data provides input for applications such as health monitoring [2], lifestyle analytics [3], and memory enhancement [4], [5]. Lifelog retrieval is the process of locating content from a lifelog to support the current application. At this point, this is a well-investigated task, including lifelog retrieval benchmarks challenges, such as the Lifelog Search Challenge (LSC) [6] and the NTCIR Lifelog task [7]. Interest is now increasing in the more challenging task of question answering (QA) for lifelogs. A key component of lifelog QA is finding lifelog-relevant contexts from a lifelog to enable accurate and efficient answering of a question. Context retrieval itself is a challenging task since contexts for lifelog QA include not only action/activity information but also require contextualized information such as time and location. It is also important to note that only the question is available as the query to identify contexts that may contain the answer, which only gives limited information for the semantic matching between QA and contexts.

Lifelog QA was introduced as a task at LSC in 2022 [8]. This task requires finding the answers to questions posed to a supplied lifelog dataset. QA on a lifelog can help the lifelogger to learn about their lifestyle, find memorable events, and get insights into their habits in an ask-and-answer manner. Lifelog QA can be seen as an Open-Domain QA [9] problem because the data is contained in a vast lifelog, and the answer is hidden within the lifelog. To resolve this task, we first need to find all relevant contexts from within that lifelog that contain information related to the question, and then generate the answer from within these contexts. A context is a lifelog event (eating, working) with additional contextual data such as time and location. For example, with the event of eating sushi, we may have a context: "At 12:10 PM on 1st January 2020, I ate sushi in a Japanese restaurant in Tokyo, Japan". From the context, we can generate an answer to the question: "What did I eat for lunch on 1st January 2020?".

Finding the correct context to answer a question is challenging due to the vast and diverse nature of lifelog data. The multi-modal aspect of lifelog, encompassing images, metadata, and textual event descriptions, complicates context retrieval. A single-modal retrieval model often fails due to semantic mismatches and variations in how time, location, and visual semantics are represented in questions and contexts. Moreover, retrieved false positive context can lead to incorrect answers.

To address this problem of context retrieval, we propose a two-stage retrieval framework for Lifelog QA. Figure 1 depicts our framework. The first stage retrieves a set of candidate contexts using a fast retrieval model, while the second stage reranks these candidates using a more computationally expensive but precise model, to refine the results [10]. Reranking models can significantly improve retrieval accuracy by prioritizing the most relevant contexts while filtering out noise. It is a particularly effective approach in multi-modal datasets like lifelogging, where different data types need to be integrated for better context selection, which is a challenge for traditional retrieval models [11]. Lifelog

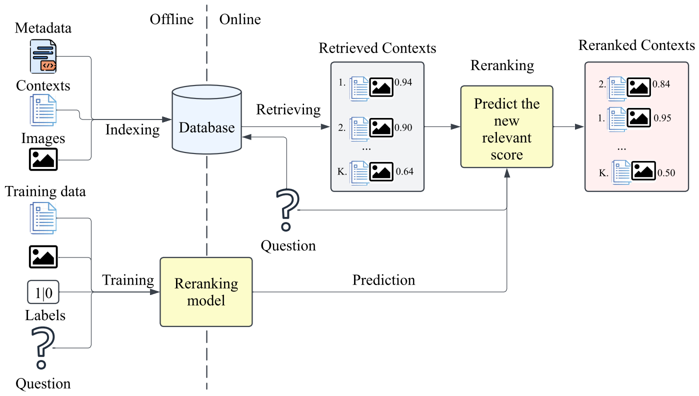

Fig. 1. A two-stage retrieval framework for Lifelog QA task. The retrieval stage is in the grey area, and the reranking stage is in the pink area.

data with a multi-modal nature also significantly differs from normal passage reranking, which largely relies on text [10], [12]. This paper proposes a cross-encoder reranking model that evaluates retrieved results and predicts their relevance to the question. By processing both lifelog images and textual descriptions, the model seeks to mitigate semantic mismatches and address the multimodal challenges inherent in lifelog data. There are two core contributions in this paper as follows:

- A novel multi-modal reranking model for the lifelog QA task, which integrates visual and textual information to improve context retrieval performance.
- A comprehensive experiment to examine the effectiveness of the model on two lifelog QA datasets, OpenLifelogQA and MemoriQA. Our results show substantial improvements in the retrieval accuracy of our reranking model.

II. Related Work

In this section, we provide an overview of related works in the fields of lifelog QA and reranking.

A. *Lifelog Question Answering*

Lifelog question-answering (QA) is a challenging task due to the multimodal, longitudinal, and diverse nature of lifelog data. Several datasets have been constructed to support research in this area. LLQA [13] was among the first to formally introduce the lifelog QA task, augmenting the Lifelog Search Challenge (LSC) dataset [8] with automatically generated question-answer pairs. MemoriQA [14] addresses some limitations of LLQA by combining manually crafted question-answer pairs with further pairs generated synthetically using GPT-4 [15], to achieve higher authenticity. In contrast, TimelineQA [16] employs a schema-based synthetic generation approach to produce a large-scale dataset containing 128 million lifelog entries in text format. Despite its scale, the synthetic nature of this dataset and its text-only format inherently limit the authenticity of the lifelog entries and questions, where performance may not generalize well to real-world scenarios.

Regarding methodological approaches, existing solutions have largely adapted NLP-based methods to lifelog QA, including retrieval-augmented generation (RAG) [17], extractive QA [18], and TableQA [19]. RAG retrieves relevant contexts to generate answers, while extractive methods identify spans within retrieved contexts as answers, typically using RoBERTa [20] as the base model. TableQA approaches, including models such as Tapex [21] and BART [22], are suitable for complex multi-hop and aggregate questions requiring named-entity extraction and arithmetic aggregation, which remain challenging due to their computational complexity.

Despite progress, these text-based approaches may not fully leverage lifelog data, given its inherently multimodal characteristics, including visual, biometric, geolocation, and textual components. Our work addresses this gap by proposing a multimodal reranking model specifically designed for lifelog QA, integrating visual and textual information to enhance the quality of retrieved contexts and thereby improve QA accuracy.

B. *Cross-Encoders for Reranking*

Cross-encoder as a reranking model is a well-investigated area of research in information retrieval for text [10], [12], [23]. This model aims to rerank the ranked list from the retrieval model to improve the accuracy of the rank of items in the list. Many reranking models adapt the transformer architecture as a cross-encoder model to rank passages for the query [10]. Nogueira et al. [24] proposed using BERT to rank passages by concatenating the query and passage into a single input and predicting the relevance through binary classification. Additionally, MacAvaney et al. [23] integrated a neural ranking hybrid approach with cross-encoders by combining BERT-based representations with traditional IR features, enhancing reranking effectiveness. We also use a transformer-based model for reranking, but in our work, we include image information in the reranking model, as images play a vital role in lifelogs.

From reranking models trained on small language models like BERT, reranking models have evolved into large language models with billions of parameters and increased the accuracy of reranking [25]–[27]. For multimodal reranking, Wen et al. [28] proposed a multimodal reranking model for

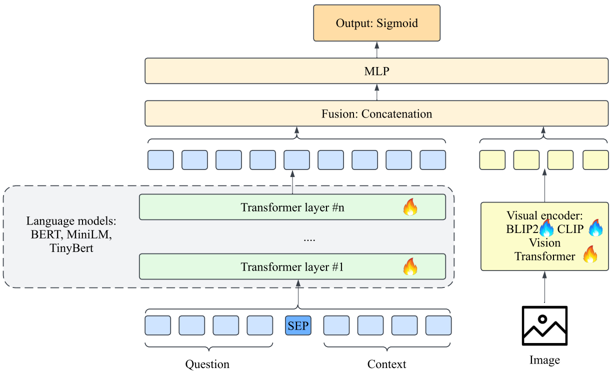

Fig. 2. Cross-encoder Multi-modal Reranking Model. Red flame indicates a fine-tuned model, while blue flame indicates a frozen model.

the knowledge-intensive visual QA task. This model combines the Vision Transformer and the Encoder-Decoder architecture to incorporate images and text for reranking problems. Their research inspires us to apply context reranking to the lifelog QA problem when QA is used in visual data and textual information. From the previous work, the transformer-based cross-encoders show their superior performance in textual passage reranking. However, lifelog data is textual descriptions of events, images, and metadata. We aim to develop a multi-modal reranking model for context retrieval in lifelog QA.

III. Methodology

This section provides information about our multimodal context reranking model. We first define the problem formulation of context reranking for lifelog QA, then describe the reranking model. The pre-trained language and vision models are also presented in this section.

A. Problem Formulation

Given the lifelog QA task of asking questions on a lifelog collection, a two-stage approach involves first retrieving (including reranking) contexts which may contain the answer to the question and then generating answers in the second stage. For a set of N questions $Q = \{q_1, q_2, ..., q_n\}$ and a set of M contexts $C = \{c_1, c_2, ..., c_m\}$, the goal of context retrieval and reranking is to retrieve the top relevant contexts to the questions.

In the retrieval stage, given the question $q_i \in$ Q, the goal is to retrieve a subset of contexts $C_{retrieved} \subseteq$ C. We use a cosine similarity score between the embedding from the Stella-1.5B model of questions and contexts to retrieve the ranked list of top-K contexts $C_{retrieved}$. We chose this model as it is ranked on top of the MTEB benchmark[1] [29] at the time of the experiment (February 2025). The cosine similarity between the question embedding $\mathbf{e}(q_i)$ and a context embedding $\mathbf{e}(c_j)$ is given by:

$$\text{cosine\_similarity}(q_i, c_j) = \frac{\mathbf{e}(q_i) \cdot \mathbf{e}(c_j)}{\|\mathbf{e}(q_i)\| \|\mathbf{e}(c_j)\|}$$

The top-K contexts $C_{retrieved}$ are retrieved by ranking all contexts based on their cosine similarity to q_i.

The context reranking problem can be formulated as learning a ranking function $f : (q_i, C_{\text{retrieved}}) \rightarrow \mathbb{R}^K$ that assigns a relevance score to each context $c_j \in C_{\text{retrieved}}$. Given a question q_i and the initially retrieved contexts $C_{\text{retrieved}} = \{c_1, c_2, \ldots, c_K\}$, the model M computes a new relevance score s_j for each context c_j as:

$$s_j = f_R(q_i, c_j)$$

where: f_R is the reranking model depicted in figure 2, and $s_j \in \mathbb{R}$ represents the relevance score of context c_j with respect to question q_i.

The final output is a reordered list of $C_{\text{retrieved}}$ based on descending s_j values.

B. Multimodal reranking model

Figure 2 illustrates the architecture of our proposed multi-modal reranking model. The model receives a concatenated question and context as the textual input and the image representing the context as the visual input. The textual input is tokenized and fed into a language model. The output of this stage is an embedding that represents information for the textual input. The image is encoded as an embedding by a visual encoder. The textual and visual embedding are concatenated in the fusion layer before going through several feed-forward layers. The output layer has a single logit and goes through a sigmoid activation function to determine the relevance of context to the question. Details about the language models and visual encoders are as follows:

We experiment with three language models, namely BERT, MiniLM, and ModernBERT. Details about the three models are as follows;

- BERT-base-uncased [30]: Bidirectional Encoder Representations from Transformers is a powerful language model designed to understand the meaning and context of words in a sentence by leveraging deep bidirectional encoding. BERT is highly effective in various tasks related to natural language understanding, including question-answering and sentiment analysis.
- MiniLM[2]: This model is a lightweight yet efficient cross-encoder model designed for passage ranking. This model is a distilled version of BERT with 6 layers of Transformers [31] and finetuned on the MS Marco dataset [32]. Unlike bi-encoder models that independently encode queries and documents, the cross-encoder processes both together, enabling deeper interaction and improved ranking accuracy.
- ModernBERT-base[3] [33]: is a powerful, modernized BERT-style encoder-only Transformer model designed

[1] https://huggingface.co/spaces/mteb/leaderboard

[2] https://huggingface.co/cross-encoder/ms-marco-MiniLM-L6-v2

[3] answerdotai/ModernBERT-base

TABLE I

EXAMPLES OF QAS AND GROUND-TRUTH CONTEXTS IN THE TWO DATASETS

Dataset	Question	Answer	Ground-truth ContextID	Context Description
OpenLifelogQA	How long did I have lunch on April 24, 2019?	I had lunch for 38 minutes	['20190424_111039_000', '20190424_114858_000']	{'20190424_111039_000': I begin to have lunch., '20190424_114858_000': I finish lunch.}
MemoriQA	What was my last activity on 27th January 2024?	I worked on laptop	['20240127_183733']	{'20240127_183733': I worked on laptop}

for long-context understanding and efficient processing. It integrates several advancements and is pre-trained on 2 trillion tokens of English and code, making it ideal for tasks involving long documents, such as retrieval, classification, and semantic search.

For the visual modality, we employ a visual encoder to extract information from images. We experiment with three visual encoders, including BLIP2, CLIP, and Vision Transformer. Details about these models are as follows:

- BLIP2 [34]: Bootstrapped Language-Image Pretraining is a vision-language model designed for efficient and powerful language-image pre-training. BLIP-2 employs a pre-trained vision encoder to extract visual features, which are then aligned with a language model through a lightweight Q-Former. The embedding of images from BLIP-2 provides a semantic embedding to represent the information in the image.
- CLIP [35]: Contrastive Language-Image Pretraining is a powerful vision-language model developed by OpenAI that encodes images and text into a shared embedding space. It learns by training on large-scale image-text pairs using a contrastive objective.
- Vision Transformer [36] is a deep learning model that applies transformer architecture to image processing. Unlike traditional convolutional neural networks (CNNs), ViT splits an image into fixed-size patches, linearly embeds them, and processes them through a self-attention-based transformer encoder. This model has the potential for extracting image embedding for the reranking model due to its performance in various tasks.

IV. EXPERIMENTAL INVESTIGATION

A. Datasets

We use two lifelog QA datasets for our investigation: OpenLifelogQA [37] and MemoriQA [14]. The OpenLifelogQA is an open-ended lifelog QA dataset with 14,322 QA and 27,688 contexts. Each QA has at least 1 ground-truth context and a maximum of 27 contexts. The average number of ground-truth contexts for each QA is 1.64. The dataset is split into training, validation, and testing sets by year and month of lifelog data, with the first 14 months for training and the last 4 months for validation and testing. There are 11,277 QA in the training set, 1,520 QA in the validation and 1,525 QA in the testing set. The reranking model is trained on the training set, and the hyperparameters are tuned in the validation set, while the result is reported on the test set.

The MemoriQA dataset is a smaller lifelog QA dataset with only 61 days of lifelog data and 3,644 QA. There are only 1,925 contexts in the dataset, and each QA has an average of 1.59 contexts. The dataset is also split by date. There are 2,823 QA for training, 326 QA for validation, and 282 QA for testing. Table I below provides some examples of the QA instances in the OpenLifelogQA and MemoriQA datasets.

B. Baseline

We use 4 retrieval models to serve as the baseline in the retrieval-only approach, including BM25 [38], Stella-1.5B [39], CLIP [35], and a hybrid of Stella+CLIP. While BM25 calculates the relevant score between questions and contexts using lexical matching, three other models calculate the cosine similarity of embeddings. Stella-1.5B is used to embed contexts and questions. CLIP embeds the images and questions into the embedding. A hybrid approach combines the cosine similarity of CLIP and Stella by weighted average to produce the final relevant score. BM25 serves as a strong lexical retrieval baseline, while Stella-1.5B and CLIP enable semantic matching in textual and multimodal contexts, respectively.

We also use MiniLM and Electra-base finetuned on MS Marco data with no fine-tuning on the lifelog dataset to rerank the retrieved contexts as baselines. After receiving a retrieved list of contexts, these two models perform reranking on the list of contexts, with the relevant score being the model's final logit. We cannot run the BERT model without fine-tuning because BERT is not designed for reranking, so the final layer has more than one logit. All the models' weights are from the Huggingface repository [4]. Using MiniLM and Electra-base as rerankers for baselines allows us to assess whether cross-encoder architectures can further improve retrieval quality in a zero-shot setting.

C. Experiment Setting

We use a hard negative sampling method to generate negative samples for the training and validation data by using BM25 to retrieve the top 10 most relevant contexts, excluding the ground-truth contexts. The maximum length for concatenated contexts and questions is set to 128. The batch size is 32, and the loss function is Binary Cross-Entropy loss. We use AdamW [40] optimizer with a 1×10^{-5} learning rate and weight decay 1×10^{-4}. We run an experiment in 20 epochs and apply early stopping for a patience of 2 epochs. All experiments are run on a single RTX 4090 GPU.

[4] https://huggingface.co/cross-encoder

TABLE II
EXPERIMENTAL RESULTS ON OPENLIFELOGQA TEST SET. BOLD NUMBER INDICATES THE HIGHEST RESULT, WHILE THE UNDERLINED NUMBER SHOWS THE SECOND HIGHEST.

Model	Type	Modality	R@1	R@5	R@10	R@20	R@50	P@1	P@5	P@10	P@20	P@50
BM25	Retrieval	Text	0.2299	0.6829	0.8171	0.9138	0.9859	0.3090	0.1999	0.1273	0.0795	0.0518
Stella-1.5B	Retrieval	Text	0.3765	0.7767	0.8801	0.9435	0.9918	0.5265	0.2479	0.1476	0.0854	0.0524
CLIP	Retrieval	Image	0.1369	0.3602	0.5305	0.7296	0.9506	0.2241	0.1296	0.0950	0.0684	0.0510
Hybrid (Stella+CLIP)	Retrieval	Text+Image	0.2949	0.6693	0.8029	0.9012	0.9800	0.4350	0.2245	0.1377	0.0822	0.0520
MiniLM	Reranking	Text	0.3811	0.7674	0.8843	0.9570	0.9919	0.5398	0.2500	0.1496	0.0868	0.0524
Electra-base	Reranking	Text	0.4012	0.8288	0.9208	0.9707	0.9971	0.5630	0.2613	0.1534	0.0877	0.0526
Fine-tuned BERT	Reranking	Text	0.6342	0.9125	0.9569	0.9801	0.9976	0.8263	0.2926	0.1609	0.0886	0.0526
Fine-tuned MiniLM	Reranking	Text	0.6408	0.9203	0.9629	0.9838	0.9978	0.8342	0.2954	0.1626	0.0890	**0.0527**
Fine-tuned ModernBERT	Reranking	Text	0.6541	0.9236	0.9638	0.9833	0.9969	0.8541	0.2968	0.1626	0.0891	0.0526
ModernBERT+CLIP	Ours	Text+Image	0.6330	0.9118	0.9606	0.9846	0.9968	0.8276	0.2875	0.1592	0.0881	0.0526
ModernBERT+BLIP2	Ours	Text+Image	0.6532	0.9234	**0.9679**	0.9846	**0.9979**	0.8548	0.2968	**0.1637**	0.0892	**0.0527**
ModernBERT+ViT	Ours	Text+Image	**0.6557**	**0.9304**	0.9676	**0.9866**	0.9963	**0.8588**	**0.2983**	0.1636	**0.0894**	0.0526

TABLE III
EXPERIMENTAL RESULTS ON MEMORIQA TEST SET. BOLD NUMBER INDICATES THE HIGHEST RESULT, WHILE THE UNDERLINED NUMBER SHOWS THE SECOND HIGHEST.

Model	Type	Modality	R@1	R@5	R@10	R@20	R@50	P@1	P@5	P@10	P@20	P@50
BM25	Retrieval	Text	0.1769	0.6712	0.8372	0.9544	0.9982	0.2660	0.2161	0.1452	0.0951	**0.0807**
Stella-1.5B	Retrieval	Text	0.4399	0.8546	0.9409	0.9926	1	0.6596	0.3007	0.1706	0.0976	0.0797
CLIP	Retrieval	Image	0.1810	0.5046	0.6885	0.9304	1	0.3156	0.1915	0.1298	0.0923	0.0797
Hybrid (Stella+CLIP)	Retrieval	Image+Text	0.4121	0.8116	0.9252	0.9858	1	0.6312	0.2858	0.1681	0.0972	0.0797
MiniLM	Reranking	Text	0.4174	0.8126	0.9024	0.9846	1	0.6135	0.2844	0.1631	0.0972	0.0797
Electra-base	Reranking	Text	0.4051	0.8100	0.9273	0.987	1	0.6099	0.2759	0.1656	0.0970	0.0797
Fine-tuned BERT	Reranking	Text	0.5192	0.8805	0.9704	0.9965	1	0.7482	0.3135	0.1770	**0.0983**	0.0797
Fine-tuned MiniLM	Reranking	Text	0.5396	0.9042	0.9710	0.9923	1	0.7766	0.3177	0.1770	0.0976	0.0797
Fine-tuned ModernBERT	Reranking	Text	0.5390	0.9019	0.9637	0.9932	1	0.7837	0.3170	0.1752	0.0978	0.0797
ModernBERT+CLIP	Ours	Image+Text	0.5437	**0.9179**	0.9743	0.992	1	0.7837	**0.3227**	**0.1780**	0.0976	0.0797
ModernBERT+BLIP2	Ours	Image+Text	**0.5484**	0.9170	**0.9778**	0.9956	1	**0.7872**	0.3213	0.1777	0.0981	0.0797
ModernBERT+ViT	Ours	Image+Text	0.5431	0.9160	0.9770	**0.9982**	1	0.7801	0.3206	**0.1780**	0.0981	0.0797

In the evaluation step, we first retrieve the top 100 relevant contexts to a question using cosine similarity between Stella 1.5B embedding of questions and contexts. A date filter is extracted from questions and applied to reduce the retrieval space. After that, we rerank the top 100 contexts using a trained reranking model and evaluate the metrics. We use two metrics, Precision@K (P@K) and Recall@K (R@K), to evaluate the accuracy of top K contexts. Because ranking between correct contexts is unimportant, metrics like NDCG and MRR are unsuitable.

D. Results

Table II and III provide the experimental results on the two datasets. These results show that baseline retrieval models, such as BM25 and CLIP, perform poorly across both the OpenLifelogQA and MemoriQA test sets. BM25 cannot capture semantic relationships because of the lexical matching method, while CLIP, which shows retrieval based on images, cannot provide enough information. Stella-1.5B model provides better performance with 88.01% R@10 and 14.76% P@10 in OpenLifelogQA, thanks to the semantics in embedding. The hybrid approach provides a poorer performance due to the combination of CLIP.

Using a reranking model without fine-tuning does not improve the performance. We can see in the performance of MiniLM and ModernBERT on both datasets that both models improve a small margin in the Stella-1.5B retrieval performance after reranking in OpenLifelogQA and even decrease the performance in the MemoriQA dataset.

Fine-tuning language models for reranking based on the textual questions and contexts significantly boosts the performance. To be specific, on the OpenLifelogQA dataset, the performance increases by approximately 28% in R@1, 8% in R@10, and 33% in P@1 compared to retrieval performance from the Stella-1.5B model in OpenLifelogQA. Fine-tuned ModernBERT achieves the highest performance among finetuned models. The increase in the MemoriQA dataset is smaller but also significant. In general, the results suggest that a fine-tuned reranking model significantly improves retrieval accuracy. With a 96.38% R@10, this provides a good list of context for the OpenLifelogQA dataset.

Although incorporating images into the reranking model does not enhance the ranking performance significantly, it also achieves the highest performance in all metrics. The multimodal reranking models leverage textual and visual features, allowing a more comprehensive representation of textual contexts and related images. While fine-tuning ModernBERT alone significantly improves results, integrating visual information provides further enhancements, particularly in Recall at higher ranks (K is 20 and 50). However, the improvement from visual fine-tuning is modest compared to the fine-tuned language model. This suggests that textual context remains the primary contribution to context information, with images providing a supplementary contribution.

TABLE IV
ERROR EXAMPLES

Question	Answer	Ground-truth Contexts	Top 5 Ranking Contexts
What did I do after cycling on March 31, 2020?	I prepared dinner with a woman	['From 05:27 PM to 06:06 PM on March 31, 2020, I am preparing dinner with a woman']	['From 04:30 PM to 05:25 PM on March 31, 2020, I start riding my bike', 'From 05:25 PM to 05:27 PM on March 31, 2020, I have returned home', 'From 06:53 PM to 06:54 PM on March 31, 2020, I am eating and watching', 'From 04:09 PM to 04:30 PM on March 31, 2020, I am in the kitchen', 'From 01:14 PM to 01:33 PM on March 31, 2020, I am having lunch and drinking water']
How did I get to DCU on March 25, 2020?	By bicycle	['From 11:35 AM to 12:16 PM on March 25, 2020, I left home and am biking to go to the workplace']	['From 05:32 PM to 06:09 PM on March 25, 2020, I have left DCU and started biking home', 'At 2020-03-25 18:09:11, I have returned home', 'From 05:15 PM to 05:18 PM on March 25, 2020, I have reached where I parked my bike at DCU', 'From 12:16 PM to 12:20 PM on March 25, 2020, I have arrived at Tesco by bike and am moving to eat', 'From 05:25 PM to 05:32 PM on March 25, 2020, I finished working, left my room with my bike']

TABLE V
QA PERFORMANCE ON DIFFERENT CONTEXT SETTINGS

Context Setting	EM	BERT Score	ROUGE-L	LLM Score
Top 20 Retrieved	0.0935	0.9095	0.3690	3.9748
Top 50 Retrieved	0.0862	0.9065	0.3714	3.9516
Top 20 Reranked	0.0948	0.9097	0.3782	4.0146
Top 50 Reranked	0.0962	0.9097	0.3778	4.0405

E. Evaluation on QA

To further evaluate the effectiveness of the reranking model in the QA task, we ran a QA experiment for different context settings and compared this to previous results from [41]. Given the question q_i and the list of N context $c = \{c_1, c_2, ..., c_n\}$:

$$\hat{a}_i = f_{LLM}(q_i, c)$$

where $\hat{a}_i$ is the predicted answer of model f_{LLM} on question q_i and the list of N context c.

We use LLama-3.1-8b-Instruct[5] [42] with several examples of QA in the prompt and ask it to generate answers for the test set of the OpenLifelogQA dataset. We use 4 context settings, including the top 20 and 50 retrieved contexts from the Stella-1.5B model, the top 20 and 50 reranked contexts from 100 retrieved contexts by ModernBERT+ViT. We evaluate the accuracy of predicted answers through BERT Score [43], ROUGE-L [44], and LLM Score [45]. Table V shows the performance. The reranked contexts bring better performance than the retrieved contexts, even when the length of the list of contexts is higher.

F. Error Analysis

We analyze some errors in the context reranking list and provide some examples in the table IV below. The reranking model struggles to find events for temporal questions that ask for information after or before a known event. For example, the question: “What did I do after cycling on March 31, 2020?” asks for an event after cycling, but the ranking contexts only rank cycling activities at the top of the list. To deal with this problem, we suggest incorporating more information from previous and next events into the context to help the model identify the correct contexts. Another problem is the ambiguity in the data. For example, the context for the question, "How did I get to DCU on March 25, 2020?" is leaving home and riding a bike to the workplace, but the model does not understand the workplace and DCU, so it fails to rank this context on top. We can address this problem by providing more information, such as that DCU is a workplace.

V. Conclusion

This paper proposes a multimodal context reranking model to improve retrieval accuracy for lifelog question-answering tasks. By integrating visual and textual data, our model effectively addresses the challenges of lifelog data’s vast, multimodal nature. Experimental results on OpenLifelogQA and MemoriQA demonstrate significant improvements in retrieval performance, with the ModernBERT+ViT model achieving the highest recall and precision scores. These findings show the effectiveness of cross-encoder reranking in refining retrieved contexts, which helps to provide more accurate and relevant contexts to lifelog QA. For future work, we plan to incorporate temporal and sequential modelling to capture event dependencies, allowing the system to provide more coherent and temporally aware answers.

Acknowledgment

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]. 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.

[5] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

References

[1] C. Gurrin, A. F. Smeaton, A. R. Doherty *et al.*, "Lifelogging: Personal big data," *Foundations and Trends® in information retrieval*, vol. 8, no. 1, pp. 1–125, 2014.

[2] J. Choi, C. Choi, H. Ko, and P. Kim, "Intelligent healthcare service using health lifelog analysis," *Journal of medical systems*, vol. 40, pp. 1–10, 2016.

[3] G. Kumar, H. Jerbi, C. Gurrin, and M. P. O'Mahony, "Towards activity recommendation from lifelogs," in *Proceedings of the 16th international conference on information integration and web-based applications & services*, 2014, pp. 87–96.

[4] T. Dingler, P. E. Agroudy, R. Rzayev, L. Lischke, T. Machulla, and A. Schmidt, "Memory augmentation through lifelogging: opportunities and challenges," *Technology-augmented perception and cognition*, pp. 47–69, 2021.

[5] Q.-L. Tran, B. Nguyen, G. J. F. Jones, and C. Gurrin, "Memoriease 2.0: A conversational lifelog retrieve system for lsc'24," in *Proceedings of the 7th Annual ACM Workshop on the Lifelog Search Challenge*, ser. LSC '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 12–17. [Online]. Available: https://doi.org/10.1145/3643489.3661114

[6] C. Gurrin, L. Zhou, G. Healy, W. Bailer, D.-T. Dang Nguyen, S. Hodges, B. Þ. Jónsson, J. Lokoč, L. Rossetto, M.-T. Tran *et al.*, "Introduction to the seventh annual lifelog search challenge, lsc'24," in *Proceedings of the 2024 International Conference on Multimedia Retrieval*, 2024, pp. 1334–1335.

[7] L. Zhou, C. Gurrin, D.-T. Dang-Nguyen, G. Healy, C. Lyu, T. Ji, L. Wang, J. Hideo, L.-D. Tran, and N. Alam, "Overview of the ntcir-17 lifelog-5 task," in *Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies. https://doi.org/10.20736/0002001329*, 2023.

[8] C. Gurrin, L. Zhou, G. Healy, B. Þór Jónsson, D.-T. Dang-Nguyen, J. Lokoć, M.-T. Tran, W. Hürst, L. Rossetto, and K. Schöffmann, "Introduction to the fifth annual lifelog search challenge, lsc'22," in *Proceedings of the 2022 International Conference on Multimedia Retrieval*, 2022, pp. 685–687.

[9] F. Zhu, W. Lei, C. Wang, J. Zheng, S. Poria, and T.-S. Chua, "Retrieving and reading: A comprehensive survey on open-domain question answering," *arXiv preprint arXiv:2101.00774*, 2021.

[10] A. Yates, R. Nogueira, and J. Lin, "Pretrained transformers for text ranking: Bert and beyond," in *Proceedings of the 14th ACM International Conference on web search and data mining*, 2021, pp. 1154–1156.

[11] L. Zhou and C. Gurrin, "Multimodal embedding for lifelog retrieval," in *International Conference on Multimedia Modeling*. Springer, 2022, pp. 416–427.

[12] Y. Mao, P. He, X. Liu, Y. Shen, J. Gao, J. Han, and W. Chen, "Rider: Reader-guided passage reranking for open-domain question answering," *arXiv preprint arXiv:2101.00294*, 2021.

[13] L.-D. Tran, T. C. Ho, L. A. Pham, B. Nguyen, C. Gurrin, and L. Zhou, "Llqa-lifelog question answering dataset," in *International Conference on Multimedia Modeling*. Springer, 2022, pp. 217–228.

[14] Q.-L. Tran, B. Nguyen, G. J. Jones, and C. Gurrin, "Memoriqa: A question-answering lifelog dataset," in *Proceedings of the 1st ACM Workshop on AI-Powered Q&A Systems for Multimedia*, 2024, pp. 7–12.

[15] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[16] W.-C. Tan, J. Dwivedi-Yu, Y. Li, L. Mathias, M. Saeidi, J. N. Yan, and A. Y. Halevy, "Timelineqa: A benchmark for question answering over timelines," *arXiv preprint arXiv:2306.01069*, 2023.

[17] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in neural information processing systems*, vol. 33, pp. 9459–9474, 2020.

[18] V. Karpukhin, B. Oguz, S. Min, P. S. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih, "Dense passage retrieval for open-domain question answering." in *EMNLP (1)*, 2020, pp. 6769–6781.

[19] G. Badaro, M. Saeed, and P. Papotti, "Transformers for tabular data representation: A survey of models and applications," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 227–249, 2023.

[20] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[21] Q. Liu, B. Chen, J. Guo, M. Ziyadi, Z. Lin, W. Chen, and J.-G. Lou, "Tapex: Table pre-training via learning a neural sql executor," *arXiv preprint arXiv:2107.07653*, 2021.

[22] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019.

[23] S. MacAvaney, A. Yates, A. Cohan, and N. Goharian, "Cedr: Contextualized embeddings for document ranking," in *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '19. ACM, Jul. 2019. [Online]. Available: http://dx.doi.org/10.1145/3331184.3331317

[24] R. Nogueira and K. Cho, "Passage re-ranking with bert," *arXiv preprint arXiv:1901.04085*, 2019.

[25] G. d. S. P. Moreira, R. Ak, B. Schifferer, M. Xu, R. Osmulski, and E. Oldridge, "Enhancing q&a text retrieval with ranking models: Benchmarking, fine-tuning and deploying rerankers for rag," *arXiv preprint arXiv:2409.07691*, 2024.

[26] A. Shakir, D. Koenig, J. Lipp, and S. Lee. (2024) Boost your search with the crispy mixedbread rerank models. [Online]. Available: https://www.mixedbread.ai/blog/mxbai-rerank-v1

[27] Y. Yu, W. Ping, Z. Liu, B. Wang, J. You, C. Zhang, M. Shoeybi, and B. Catanzaro, "Rankrag: Unifying context ranking with retrieval-augmented generation in llms," 2024. [Online]. Available: https://arxiv.org/abs/2407.02485

[28] H. Wen, H. Zhuang, H. Zamani, A. Hauptmann, and M. Bendersky, "Multimodal reranking for knowledge-intensive visual question answering," *arXiv preprint arXiv:2407.12277*, 2024.

[29] N. Muennighoff, N. Tazi, L. Magne, and N. Reimers, "Mteb: Massive text embedding benchmark," 2023. [Online]. Available: https://arxiv.org/abs/2210.07316

[30] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019. [Online]. Available: https://arxiv.org/abs/1810.04805

[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023. [Online]. Available: https://arxiv.org/abs/1706.03762

[32] P. Bajaj, D. Campos, N. Craswell, L. Deng, J. Gao, X. Liu, R. Majumder, A. McNamara, B. Mitra, T. Nguyen, M. Rosenberg, X. Song, A. Stoica, S. Tiwary, and T. Wang, "Ms marco: A human generated machine reading comprehension dataset," 2018. [Online]. Available: https://arxiv.org/abs/1611.09268

[33] B. Warner, A. Chaffin, B. Clavié, O. Weller, O. Hallström, S. Taghadouini, A. Gallagher, R. Biswas, F. Ladhak, T. Aarsen, N. Cooper, G. Adams, J. Howard, and I. Poli, "Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference," 2024. [Online]. Available: https://arxiv.org/abs/2412.13663

[34] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," 2023. [Online]. Available: https://arxiv.org/abs/2301.12597

[35] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," 2021. [Online]. Available: https://arxiv.org/abs/2103.00020

[36] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021. [Online]. Available: https://arxiv.org/abs/2010.11929

[37] Q.-L. Tran, B. Nguyen, G. J. F. Jones, and C. Gurrin, "Openlifelogqa: An open-ended multi-modal lifelog question-answering dataset," 2025. [Online]. Available: https://arxiv.org/abs/2508.03583

[38] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: Bm25 and beyond," *Found. Trends Inf. Retr.*, vol. 3, no. 4, p. 333–389, Apr. 2009. [Online]. Available: https://doi.org/10.1561/1500000019

[39] D. Zhang, J. Li, Z. Zeng, and F. Wang, "Jasper and stella: distillation of sota embedding models," 2025. [Online]. Available: https://arxiv.org/abs/2412.19048

[40] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2019. [Online]. Available: https://arxiv.org/abs/1711.05101

[41] Q.-L. Tran, D. N. Pham, T. Q. Truong, H. M. Nguyen, C. H. Le, K. D. Vu, T. V. M. Nguyen, K. V. Nguyen, L. L. P. N. Nguyen, T. Le, P. M. Dang, B. Nguyen, G. J. F. Jones, and C. Gurrin, "A rag approach

for multi-modal open-ended lifelog question-answering," in *Proceedings of the 48th International ACM ICMR Conference on Research and Development in Information Retrieval (ICMR '25)*. Chicago, IL, USA: ACM, 2025, p. 10 pages.

[42] A. G. et al., "The llama 3 herd of models," 2024. [Online]. Available: https://arxiv.org/abs/2407.21783

[43] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," 2020. [Online]. Available: https://arxiv.org/abs/1904.09675

[44] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. [Online]. Available: https://aclanthology.org/W04-1013/

[45] T. Kocmi and C. Federmann, "Large language models are state-of-the-art evaluators of translation quality," 2023. [Online]. Available: https://arxiv.org/abs/2302.14520