OmniQuery: Enabling Question Answering on Personal Memory by Augmenting Multimodal Album Data

Jiahao Nick Li
UCLA
Los Angeles, USA
ljhnick@ucla.edu

Zhuohao (Jerry) Zhang
University of Washington
Seattle, USA
zhuohao@uw.edu

Jiaju Ma
Stanford University
Palo Alto, USA
jiajuma@stanford.edu

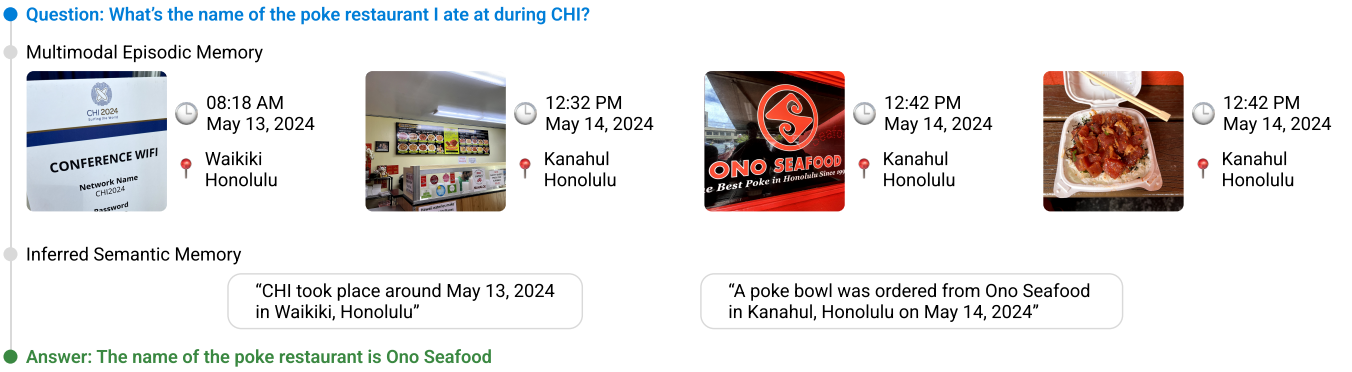

Figure 1: OmniQuery enables natural language QA on users' multimodal memory. This is achieved by augmenting the episodic memories in photo albums with inferred semantic memory.

ABSTRACT

We present OmniQuery, an interactive system that augments users' personal photo albums and enables free-form question answering on users' past memories. OmniQuery processes multimodal media data in personal albums, aggregates them into related episodic memory databases in different levels, and infers semantic knowledge including personal facts like social relationships, preferences, and experiences. OmniQuery then allows users to interact with their database using natural language, giving media that directly matches the query or an exact answer supported by related media as a result.

CCS CONCEPTS

• **Human-centered computing** → **User studies**; **Interactive systems and tools**; *Interaction techniques.*

KEYWORDS

personal AI assistant, semantic memory, episodic memory, large language models, retrieval augmented generation

ACM Reference Format:
Jiahao Nick Li, Zhuohao (Jerry) Zhang, and Jiaju Ma. 2024. OmniQuery: Enabling Question Answering on Personal Memory by Augmenting Multimodal Album Data. In *The 37th Annual ACM Symposium on User Interface Software and Technology (UIST Adjunct '24), October 13–16, 2024, Pittsburgh, PA, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3672539.3686313

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
UIST Adjunct '24, October 13–16, 2024, Pittsburgh, PA, USA
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0718-6/24/10.
https://doi.org/10.1145/3672539.3686313

1 INTRODUCTION

People often take pictures and screenshots to record and share memorable moments and important information in their lives [4], making their digital photo album a personal dataset of episodic memories. Such a dataset includes a diverse range of data structures from textual (e.g., newly added contacts, chat history among user groups, etc.) to graphical information (e.g., conference event confirmations, scenic photos, etc.), which has the huge potential to be used for personal memory retrieving and enhancing our interaction with past experiences.

With the rapid advancement in AI, numerous work and products were developed to utilize personal history data for interacting with past episodic memories, including intelligent search of text and objects within pictures actively taken by users [1, 2, 6] and passively captured by the system (*e.g.*, Microsoft's Recall[1] and pervasive AR systems [5]). However, existing tools usually treat graphic and textual information as individual pieces to be matched and retrieved. They do not provide a comprehensive semantic understanding on the memory context. For example, current systems might be able to locate vacation photos using metadata or image content but cannot synthesize this information to understand the user's travel plan (e.g., multi-city visits, activities, and travel companions). In this paper, we present OmniQuery, a framework that augments users' personal media database and enabling free-form question answering on users' past memory.

[1] https://learn.microsoft.com/en-us/windows/ai/apis/recall

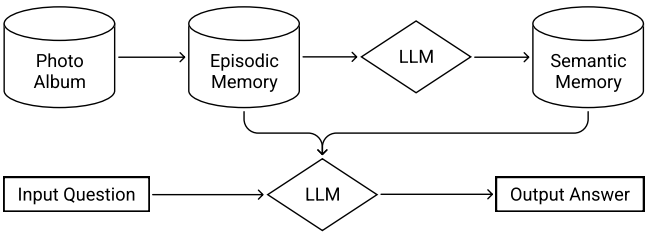

Figure 2: The system diagram of OmniQuery

OmniQuery takes multimodal media such as photos and videos, along with metadata including the time and location of the captured media, to construct a personalized semantic knowledge base. Such multimodal media represents the episodic memories of the user and OmniQuery processes them into structured data leveraging multimodal AI models. Then it performs different levels of semantic analysis on episodic memories to infer semantic memory, such as events occurring within a single day or over extended periods.

Upon detecting these events, OmniQuery re-examines the raw media data to further infer personal knowledge and detailed information about the events, functioning as an intelligent assistant that automatically comprehends past experiences. Example inferred knowledge includes personal facts like social relationships, preferences like hobbies and favorite food, and experiences like travels and events. Ultimately, OmniQuery enables natural language querying and answering, allowing users to interact with their indexed episodic memories. As a result, OmniQuery gives either the exact media that matches the query or an answer powered by supporting media.

2 SYSTEM DESIGN

OmniQuery consists of two major components: *(i)* the structure of a personalized semantic knowledge base from augmenting multimodal data in personal albums, and *(ii)* an interactive system that enables natural language QA on the semantic knowledge base.

2.1 Memory augmentation

Every picture and video in the personal photo album is referred to as *episodic memory* of the user as it stores detailed and context-specific information about what the user saw and heard on that particular day and at that specific location.

Oppositely, such episodic memories lack a general understanding and knowledge of the user's history. For example, they do not identify patterns in the user's hobbies (e.g., regularly going to the gym) or detail activities during specific events (e.g., dining in a good restaurant during a conference). Such semantic knowledge could enable more general question answering capabilities on the personal photo albums, however, it requires processing and understanding multiple semantically, temporally and spatially related episodic memories simultaneously.

We discuss the workflow of building such *personalized semantic knowledge* as follows.

2.1.1 Pre-processing. OmniQuery processes pictures (visuals) and videos (visuals and audio) in the photo albums, each containing metadata including the capture time and location. As the first step, OmniQuery processes the data into structured text data (*e.g.*, `<scene caption> <visible text> <object> <speech> <sound description>`) leveraging multimodal AI models.

2.1.2 Event and semantic knowledge inference. Media files taken on the same date are grouped together and the corresponding structured text are sent to an LLM to detect event-related information (e.g., pictures of conference registration indicating a conference happening on that day). Then all the event information within a week is grouped and OmniQuery further leverages an LLM to distill more significant events (e.g., conference that spans multiple days).

Once events are detected, the raw episodic memories are revisited, and together with the event information, LLMs are used to infer semantic knowledge (e.g., I stayed at a hotel with marine view during the conference).

2.1.3 Indexing. All episodic memories (image embeddings) and the parsed semantic knowledge (text embeddings) are indexed for efficient retrieval and enhanced question-answering capabilities.

2.2 End-to-end interactive system

As shown in Figure 2, OmniQuery adopts a RAG-based system architecture [3] to enable users interact with the augmented multimodal database using natural language. Specifically, OmniQuery maintains three major vector database – multimodal episodic memories, detected events, and semantic knowledge. Given a natural language query, OmniQuery first computes its text embeddings and calculates the similarity between it and each vector database. Then, OmniQuery retrieves the top K most similar elements from each vector database and leverages an LLM to generate the answer. For answers that can be supported by specific episodic memories (e.g., a photo of the restaurant when querying the name of the restaurant), the corresponding episodic memories are also provided as context.

3 PRIVACY CONCERN

Note that media in photo albums may contain a lot of private information, including personally identifiable information (PII). To mitigate privacy concerns, locally deployed models or on-device models should be utilized. However, as a proof-of-concept, we demonstrate the potential of our system by leveraging state-of-the-art models (such as OpenAI APIs, which also claim not to save user data on their servers for privacy reasons) in this project.

REFERENCES

[1] Peter Fornaro and Vera Chiquet. 2020. Artificial Intelligence for Content and Context Metadata Retrieval in Photographs and Image Groups. *Archiving Conference* 17, 1 (April 2020), 79–82. https://doi.org/10.2352/issn.2168-3204.2020.1.0.79

[2] Google. 2024-07-05. Google Photos: Edit, Organize, Search, and Backup Your Photos. https://www.google.com/photos/about/

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

[4] Jiahao Nick Li, Yan Xu, Tovi Grossman, Stephanie Santosa, and Michelle Li. 2024. OmniActions: Predicting Digital Actions in Response to Real-World Multimodal Sensory Inputs with LLMs. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 8, 22 pages. https://doi.org/10.1145/3613904.3642068

[5] Xingyu Bruce Liu, Jiahao Nick Li, David Kim, Xiang 'Anthony' Chen, and Ruofei Du. 2024. Human I/O: Towards a Unified Approach to Detecting Situational

Impairments. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 965, 18 pages. https://doi.org/10.1145/3613904.3642065

[6] Dmytro Nikulin and Olena Buchko. 2020. Automated Approach for the Importing the New Photo Set to Private Photo Album to Make it More Searchable. *NaUKMA Research Papers. Computer Science* 3, 0 (Dec. 2020), 141–148. https://doi.org/10.18523/2617-3808.2020.3.141-148

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009