MemoryMirrorQA: First-Person Video Question Answering for Aging Population

Mengye Ren, Assistant Professor of Computer Science & Data Science, NYU
John-Ross Rizzo, Associate Professor of Rehabilitation Medicine & Neurology, NYU

Keywords: Visual memory assistance, question answering, large language models

Overview. In our aging society, dementia and visual impairment are rising as common health issues among the elderly, causing dependency and disability. Advances in AI, particularly in computer vision and natural language understanding, are making personalized AI assistants a reality. MemoryMirror QA is an innovative AI-enabled solution for first-person video question answering (QA) that provides practical visual memory assistance. Users can inquire about past visual experiences and retrieve relevant moments. It utilizes AI technologies like large language models (LLMs) and video caption models and integrates seamlessly with augmented reality (AR) hardware, offering an intuitive user experience. Unlike existing video QA systems, our solution leverages general-purpose LLMs for higher domain-general reasoning capability. MemoryMirror QA has the potential to enhance the quality of life for the elderly and benefit the broader public.

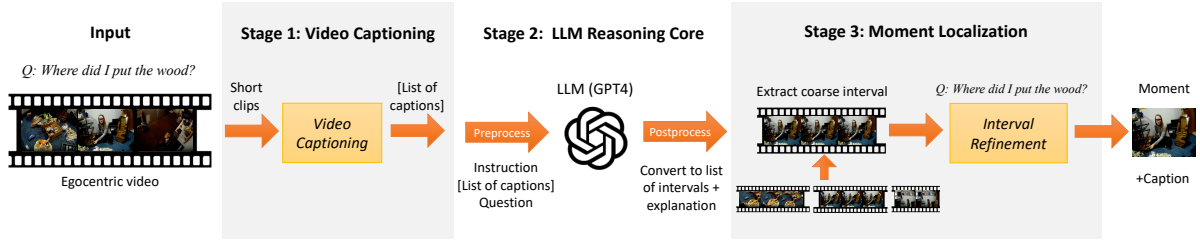

Figure 1: A technical overview of our proposed system *MemoryMirrorQA*

Scientific & AI Approach. Our research aims to explore **whether we can effectively leverage the state-of-the-art AI technology available today to develop a usable prototype for first-person video question answering, specifically designed for episodic memory retrieval**. The proposed system combines the strengths of video caption models [4] and large language models (LLM) [3], as depicted in Figure 1. We introduce our system, named *MemoryMirrorQA*, which consists of three stages. Initially, we capture episodic videos from AR devices or home cameras, segmenting them into shorter clips lasting around 3-5 seconds. In the first stage, we process these clips using video captioning to convert them into a list of text descriptions. Next, we input the descriptions, along with the user's question and a predefined instruction, into an LLM. The LLM outputs a list of intervals that potentially hold the answer to the user's question, accompanied by an explanation. Finally, we extract the most relevant video intervals and refine them to precisely locate the desired moment. The user is then presented with a concise video segment accompanied with an explanation produced by the LLM. This response can be projected onto the display of AR devices, or can be captioned and announced in speech format.

Our Team & Pilot Study. Ren lab is conducting a pilot study to evaluate the performance of our proposed system using a publicly available egocentric video dataset with human annotations (Ego4D) [2]. Remarkably, our preliminary version of the system achieved third place in the 2023 Ego4D challenge at CVPR, surpassing the previous year's winner. Compared to other winning solutions, our system is the first to leverage a general-purpose LLM

with advanced reasoning capabilities. These preliminary findings demonstrate the potential and feasibility of our approach for video question answering. We anticipate significant improvements in our system's ability to comprehend longer videos in the next stage.

Rizzo lab has developed the VIS4ION (Visually Impaired Smart Service System for Spatial Intelligence and Onboard Navigation) [1], an advanced wearable technology designed to capture images from the user's perspective. The technology utilizes object detection to identify and describe objects. Rizzo lab has well-established patient population comprising individuals with blindness and low vision as well as healthy elderly population for potential user testing. Both labs will collaborate on building object detection to enhance captions for more structured information retrieval of past events, and jointly deploy and test the device with existing patient population with Rizzo lab.

Relevance to Healthy Aging and AD/ADRD. The proposed project holds direct benefits for elderly individuals experiencing dementia, AD/ADRD, and visual impairment. Our system can provide an intuitive natural language interface to access visual content of past events, and improve users' overall quality of life. The potential benefits of this system extend beyond specific conditions, positively impacting the general aging population.

Aims and Expected Outcomes. In the next stage of our research and development, our aims are: 1) enhance video understanding capability through stronger video representations and captioning models (Ren); 2) improve long video processing by developing hierarchical video content storage and retrieval (Ren); 3) use real-time object detection to enrich caption information (Rizzo+Ren); 4) deploy and test the system on AR hardware (Ren+Rizzo). Successful deployment may require additional fundamental research: 1) lower power on-device LLM computing for reduced latency and enhanced privacy; 2) selection of key frames for efficient storage and retrieval. We expect further improvements in system accuracy for daily use. We plan for a real-time demo, and plan to push towards a user testing. We will open-source method details, release experiment data, deliver a project report, and submit papers to top-tier conferences on machine learning, computer vision, and human-computer interaction (e.g. NeurIPS, CVPR, CHI, etc.).

Potential for Commercialization of Project Deliverables. Our proposed system holds significant potential for commercialization. We envision deploying the video question answering capability on mobile computing devices. This could take the form of a software application designed for AR platforms, such as Apple's Vision Pro or Meta's Oculus Quest 3. It can also be implemented as a home camera accompanied by a phone app for connectivity and display. The product has the capacity to greatly benefit individuals with dementia and/or visual impairment, with the potential to face towards the general public.

[1] M. Beheshti, T. Naeimi, ..., and J.-R. Rizzo. A smart service system for spatial intelligence and onboard navigation for individuals with visual impairment (vis4ion thailand): Study protocol of a randomized controlled trial of visually impaired students at the ratchasuda college, thailand. *Trials*, 24(1):1–17, 2023.

[2] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *CVPR*, 2022.

[3] OpenAI. Gpt-4 technical report, 2023.

[4] Y. Zhao, I. Misra, P. Krähenbühl, and R. Girdhar. Learning video representations from large language models. *CoRR*, abs/2212.04501, 2022.