Looking at Your Photo, What Comes to Mind? Personalized Memory Internalization for Dementia Reminiscence

Shunjie Wen, Kyung-Hwan Lee, Gimoon Lee, Seongsoo Heo, Jaeyeon Lee, Sangyeob Shin, Gyuwon Moon, Jiwoong Kim, Dong-Wan Choi*
Big Data Lab., Department of Computer Science and Engineering, Inha University, South Korea
{wenshunjie, lkh0224, nhtgb021030, woo555813, dlwodus159, ssyppsspp, gyuwon822, wldnd7145}@inha.edu, dchoi@inha.ac.kr

Abstract

We present PhoMi, an interactive recall assistant that supports dementia reminiscence by engaging users with personal photographs captured via a live camera interface. Given a photograph, PhoMi delivers spoken questions and receives spoken responses, creating an accessible reminiscence setting with reduced reliance of human therapists. Over repeated sessions, user responses are incorporated into lightweight adapters of a vision–language model, enabling progressively personalized question generation without reprocessing prior interaction logs. PhoMi serves as a prototype toward scalable, lifelong AI companions for dementia reminiscence.

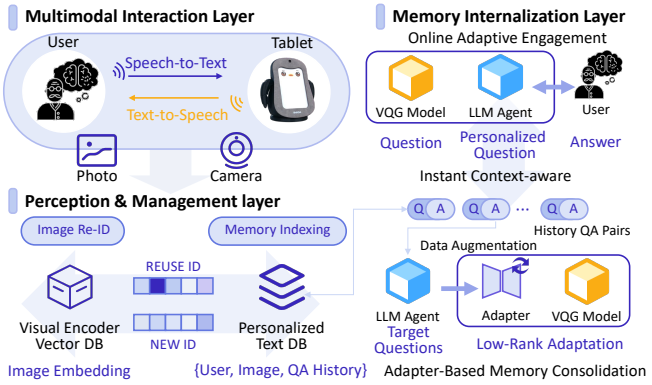

Figure 1: Schematic overview of PhoMi.

1 Introduction

Reminiscence therapy leverages sensory cues to stimulate autobiographical recall in individuals with dementia. In clinical settings, personal photographs often serve as anchors that elicit verbal expression of past experiences [Woods *et al.*, 2018]. Effective reminiscence sessions, however, typically rely on human therapists who guide interactions through questioning and empathetic engagement, making sustained and scalable support difficult to provide in everyday settings.

Early attempts at automated reminiscence therapy [Carós *et al.*, 2020; Gamborino *et al.*, 2021] introduced memory assistance systems based on deep neural networks. However, these systems generally treat interactions as isolated sessions, without longitudinal personalization across repeated encounters. Furthermore, prior implementations have relied on chatbot-style interfaces, limiting embodied interaction through live visual capture and spoken communication.

Recent advances in multimodal large language models (LLMs) enable more flexible visual-linguistic reasoning. Yet, when applied to dementia reminiscence settings, existing approaches still lack mechanisms for incremental, user-specific adaptation over time. In particular, there remains a gap in *internalizing* fragmented dialogue into model parameters without relying on ever-growing context windows.

In this paper, we present PhoMi, an interactive recall assistant for sustained and personalized dementia reminiscence. PhoMi integrates live visual capture with spoken question–answer interaction to support multimodal engagement. Given a personal photograph, the system anchors visual stimuli through a real-time camera interface and delivers spoken prompts to elicit autobiographical responses. To maintain continuity across sessions, PhoMi estimates whether the presented photo is familiar to the user and adjusts its prompting strategy to encourage deeper recall over time [Duan *et al.*, 2025; Hsu *et al.*, 2025]. Beyond interaction logging, PhoMi performs *memory internalization* by distilling accumulated question–answer pairs into parameter-efficient updates of lightweight adapters attached to a vision–language model (VLM). This adapter-based internalization provides a warm-start in future sessions, enabling the system to function as an increasingly personalized recall assistant.

2 The PhoMi Recall Assistant

The schematic overview of PhoMi is shown in Figure 1. The system is organized into three collaborative layers that jointly support personalized reminiscence: (i) a multimodal interaction layer, (ii) a perception and management layer, and (iii) a memory internalization layer. We describe each component in detail below.

*Corresponding author.

2.1 Multimodal Interaction

The multimodal interaction layer enables intuitive visual and spoken communication tailored to individuals with early-stage dementia.

Live Camera Unlike standalone image-upload models, PhoMi incorporates a live camera interface to capture photos in real time. This design makes "looking at a photo" the primary trigger for initiating each reminiscence session.

Speech-Driven Dialogue To support hands-free interaction, the system integrates speech-to-text (STT) and text-to-speech (TTS) modules instead of relying on on-screen text. This allows patients to verbally express memories and receive natural spoken prompts in return.

2.2 Perception and Management

This layer serves as the bridge between raw visual perception and internal memory indexing.

Re-Identification. When a photo is captured, the system employs a visual encoder to extract feature embeddings, which are stored in a vector database (Vector DB). Using these stored embeddings, PhoMi performs a similarity search to determine whether the captured image matches a previously encountered instance. If the similarity exceeds a predefined threshold, the corresponding `ImageID` is reused; otherwise, a new identifier is created for the unseen image.

Dual-Database Each `ImageID` serves as a relational key linking the visual embedding stored in the Vector DB and the associated conversational history maintained in a text database (Text DB). This coordinated design enables consistent referencing of image-specific memory across sessions.

2.3 Memory Internalization

A central design principle of PhoMi is shifting personalization from transient log retrieval to parameter-level adaptation. Instead of repeatedly injecting conversational history into the context window, PhoMi incrementally consolidates accumulated interactions into the model's parameters via Low-Rank Adaptation (LoRA) [Hu *et al.*, 2022]. This internalization is performed as a background process, allowing personalization to evolve without interrupting real-time interaction.

Data Augmentation To transform fragmented dialogue into learnable supervision, we employ an LLM (`LLaMA-v3-70B`) to filter and structure raw Q&A pairs stored in the Text DB across multiple intent categories (e.g., identity, emotion, routine, relationship, meaning, social memory, temporal context, caregiving, and preference). The LLM synthesizes semantically grounded fine-tuning targets with controlled variations, converting conversational traces into relevant questions.

Memory Consolidation These target questions are used to train image-specific LoRA adapters (rank 4), applied to the $W_{Q,K,V,O}$ projection matrices and attached to a VLM (`LLaVA-1.5-7B`) [Liu *et al.*, 2024]. Each adapter is also indexed by a unique `ImageID`, enabling image-level parameter specialization across sessions. Upon re-identification, the agent activates the corresponding adapter, allowing the VLM to immediately condition on previously internalized personal information without retrieval of historical logs.

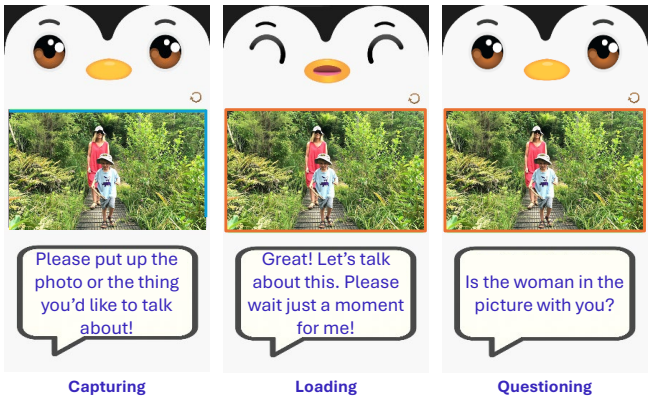

Figure 2: User interface of PhoMi.

3 Demonstration

3.1 Implementation Details

Hardware and Frontend PhoMi is implemented as a native Android application and deployed on an 11-inch tablet device (Galaxy Tab S11). A penguin-shaped enclosure and animated facial expressions, adapted from the MoMo robot[1], are incorporated in our system to provide a friendly interaction setting. As illustrated in Figure 2, when a photo is positioned under the integrated camera, the UI displays a dynamic loading frame to indicate visual processing. Once identified, the system delivers a spoken prompt, capturing and transcribing user responses in real time via the built-in microphone.

Backend and Software Stack The backend uses an asynchronous microservice architecture. `Gunicorn` and `Flask-SocketIO` serve as the API gateway, while task orchestration is managed through `Celery` with `Redis` as the message broker. This enables sequential execution across workers dedicated to visual encoding (`DINO`), visual question generation, dialogue generation, speech recognition (`Whisper-Medium`) [Radford *et al.*, 2023], and speech synthesis (`MeloTTS`) [Zhao *et al.*, 2023]. For knowledge management, we utilize a `FAISS`-based vector DB [Johnson *et al.*, 2019] for visual features and an `SQLite`-based text DB[2] for conversation logs.

Resources and Training PhoMi runs on a workstation with four NVIDIA A6000 GPUs: two for the LLM and two for the VLM, STT, TTS, and visual encoder modules. During memory internalization, the LLM generates 8 core questions from accumulated Q&A history, which are expanded into 40 variants and used to train the image-specific LoRA adapter for 30 epochs.

3.2 Use Cases

We evaluate the practical effectiveness of PhoMi through several representative use cases, illustrating the interaction flow

[1]https://www.whydots.com/en/robot-momo
[2]https://sqlite.org/

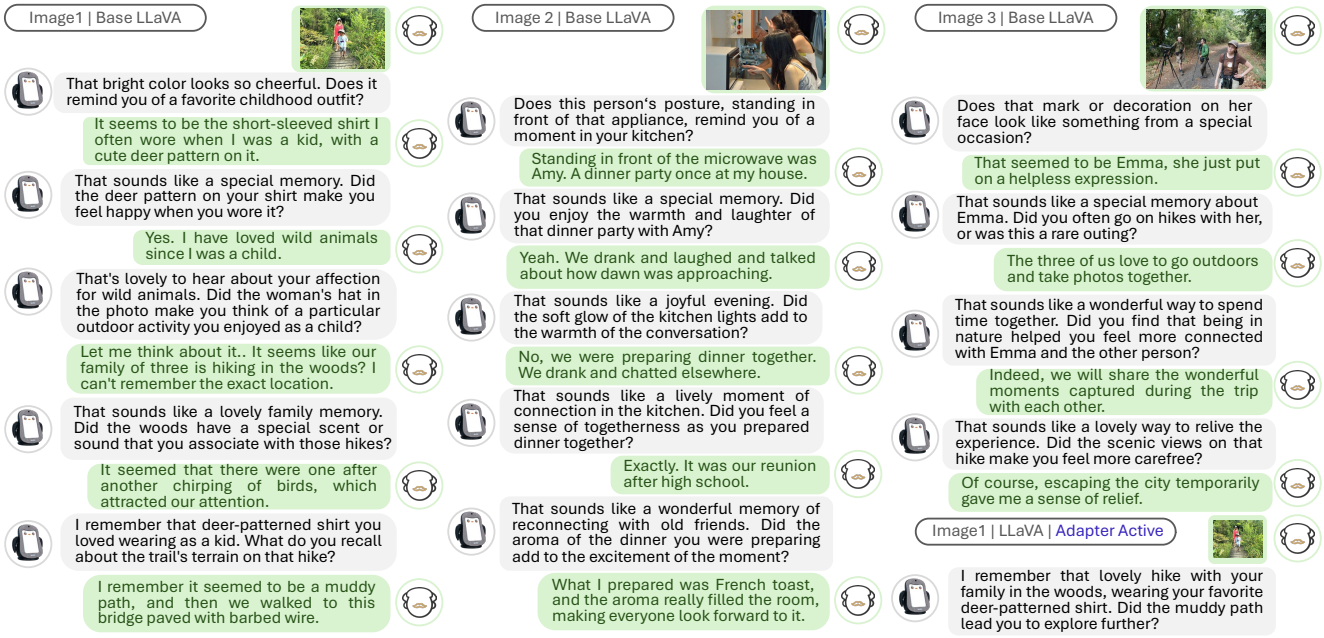

Figure 3: Conversation flow of PhoMi. The actual interaction was conducted via speech.

Raw Questions from Base LLaVA

- What is the purpose of the wooden bridge in the image?
- What's the name of the little boy in the blue shirt?
- What is the relationship between the woman and the child in the image?
- What is the little boy holding?
- What is the color of the woman's shirt?
- What is the boy wearing?
- What is the little boy wearing on his head?

Raw Questions from LLaVA with an Adapter

- Do you remember wearing a shirt with a cute deer design?
- Did you hike on a muddy path?
- Was this when you went on a hike with birds singing?
- Do you remember walking on a muddy path?
- Do you remember wearing a deer-patterned shirt when you were young?
- Did this outdoor activity remind you of a specific hat?
- Do you remember a hike when birds were chirping loudly?

Figure 4: Generated questions before (top) and after (bottom) memory internalization.

and the impact of memory internalization on personalized question generation. A full demo video showing the multimodal sensing and real-time interaction process is available at: https://youtu.be/m6AoiS7m5e0.

Interaction Flow Figure 3 illustrates PhoMi interacting with a user across three distinct sessions, each initiated by a novel photograph captured via the live camera. As the perception layer recognizes the photograph as unseen, the VLM starts with generating image-oriented questions without personalized context. For example, it queries noticeable attributes such as a brightly colored outfit, a person in a kitchen, or distinctive facial features. These generic prompts serve as entry points for reminiscence, encouraging the user to share verbal memories, which are transcribed in real time to form the initial dialogue history.

Memory Internalization Memory internalization marks the transition of PhoMi from a generic visual assistant to a persistent personal companion. As underlined in Figure 4, questions generated with the image-specific adapter are enriched with personalized keywords, such as previously mentioned participants or locations. Through this consolidation process, personalized information is embedded into the model's parametric space, enabling personalized warm-starts in subsequent sessions without requiring access to prior textual logs.

Personalized Warm-Start When a previously seen photograph is reloaded through the live camera interface (bottom-right of Figure 3), PhoMi re-identifies the corresponding `ImageID` and activates the image-specific adapter within the VLM. As a result, the system resumes the interaction from prior context, incorporating confirmed personal facts (e.g., family hiking or a deer-patterned shirt) into the opening question.

4 Conclusion

In this work, we presented PhoMi, a multimodal reminiscence assistant that internalizes personal memories into parametric space, enabling personalized warm-start interactions without relying on textual logs. By tightly integrating live perception, speech-driven dialogue, and lightweight adapter-based memory consolidation, PhoMi transcends static assistive tools and evolves into a self-growing companion. We envision this paradigm as a foundation for AI systems that do not merely recall the past, but accumulate it, scaling toward *lifelong, privacy-preserving personalized memory support*.

Acknowledgments

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grants funded by the Korea government (MSIT) (No.2022-0-00448/RS-2022-II220448, RS-2026-25544527, RS-2026-25529503).

References

[Carós *et al.*, 2020] Mariona Carós, Maite Garolera, Petia Radeva, and Xavier Giro-i Nieto. Automatic reminiscence therapy for dementia. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*, pages 383–387, 2020.

[Duan *et al.*, 2025] Jinhao Duan, Xinyu Zhao, Zhuoxuan Zhang, Eunhye Grace Ko, Lily Boddy, Chenan Wang, Tianhao Li, Alexander Rasgon, Junyuan Hong, Min Kyung Lee, Chenxi Yuan, Qi Long, Ying Ding, Tianlong Chen, and Kaidi Xu. GuideLLM: Exploring LLM-guided conversation with applications in autobiography interviewing. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5558–5588. Association for Computational Linguistics, 2025.

[Gamborino *et al.*, 2021] Edwinn Gamborino, Alberto Herrera Ruiz, Jing-Fen Wang, Tsung-Yuan Tseng, Su-Ling Yeh, and Li-Chen Fu. Towards effective robot-assisted photo reminiscence: Personalizing interactions through visual understanding and inferring. In Pei-Luen Patrick Rau, editor, *Cross-Cultural Design. Applications in Cultural Heritage, Tourism, Autonomous Vehicles, and Intelligent Agents*, pages 335–349. Springer, 2021.

[Hsu *et al.*, 2025] Long-Jing Hsu, Manasi Swaminathan, Weslie Khoo, Kyrie Jig Amon, Hiroki Sato, Sathvika Dobbala, Kate Tsui, David Crandall, and Selma Sabanovic. Bittersweet snapshots of life: Designing to address complex emotions in a reminiscence interaction between older adults and a robot. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2025.

[Hu *et al.*, 2022] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022*, 2022.

[Johnson *et al.*, 2019] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.

[Liu *et al.*, 2024] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024*, pages 26286–26296. IEEE, 2024.

[Radford *et al.*, 2023] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning, ICML 2023*, pages 28492–28518. PMLR, 2023.

[Woods *et al.*, 2018] Bob Woods, Laura O'Philbin, Emma M Farrell, Aimee E Spector, and Martin Orrell. Reminiscence therapy for dementia. *Cochrane database of systematic reviews*, (3), 2018.

[Zhao *et al.*, 2023] Wenliang Zhao, Xumin Yu, and Zengyi Qin. Melotts: High-quality multi-lingual multi-accent text-to-speech, 2023.