Encode-Store-Retrieve: An Accurate Memory Augmentation System via Language-Encoded Egocentric Perception

Anonymous ACL submission

Abstract

We depend on our own memory to **encode**, **store**, and **retrieve** experiences, but face the issue of memory lapses. Addressing this, life logging through wearable devices capturing egocentric videos is a promising memory augmentation method. However, it deals with challenges in efficiently managing large video data, computational intensity for retrieval, and privacy concerns. Our proposed system uses natural language encoding for video data, stored in a vector database, leveraging large vision language models for encoding and large language models for querying. In the QA-Ego4D dataset, our system achieved a BLEU score of 8.3, surpassing conventional models scoring between 3.4 and 5.8. A user study revealed our system significantly improves episodic memory task performance compared to human participants.

1 Introduction

Human memory, crucial for various cognitive functions, often relies on external aids like photographs and reminders for enhancing recollection of past events (Klein, 2015; Intons-Peterson and Newsome III, 1992; Intons-Peterson, 2014). However, these aids are limited in scope, making internal memory essential for encoding, storing, and retrieving experiences. Memory lapses, especially in aging, are common. Life logging, capturing images, videos, and personal data, is proposed as a method for memory augmentation (Harvey et al., 2016; Dingler et al., 2021; Chen and Jones, 2010; Hayes et al., 2004; Hodges et al., 2006; Gurrin et al., 2014b). The emergence of smart glasses and AR headsets introduces new opportunities for life logging, though challenges in data encoding, storage, and privacy compliance persist (Gurrin et al., 2014a).

This paper presents a *language-encoded* episodic memory system to tackle these issues. It employs language encoding of egocentric videos using a large vision language model, stores language embeddings in a vector database for efficient retrieval, and uses a large language model for open-ended question answering in episodic memory tasks. We fine-tune the Large Language and Vision Assistant (LLaVA) (Liu et al., 2023) for egocentric data and integrate OpenAI GPT-4 (OpenAI, 2023) with Chroma (Core, 2023) for memory storage and retrieval.

We utilize the QA-Ego4D dataset (Bärmann and Waibel, 2022; Grauman et al., 2022), designed for the Episodic Memory Question Answering (EMQA) task with a constant-size memory constraint. Our method achieves a BLEU score of 8.3, outperforming conventional models (Bärmann and Waibel, 2022).

We implemented our system on the HoloLens 2 device and conducted a week-long study. Our system outperformed human participants in episodic memory tasks, achieving a mean response score of 4.13/5 compared to the participants' score of 2.46/5. A post-study questionnaire revealed reduced privacy concerns, demonstrating the system's effectiveness in memory augmentation.

Hence, our research paper presents three contributions as outlined below:

- We introduce, for the first time to our knowledge, a memory augmentation system that combines egocentric vision language encoding with episodic memory question answering (EMQA) tasks, utilizing a vector database for efficient storage and retrieval.
- We present a large-scale quantitative evaluation of our system using the EMQA benchmark, specifically the QA-Ego4D dataset, demonstrating its effectiveness in episodic memory tasks.
- We conduct a user study to examine our system's potential applications, revealing its benefits in various scenarios and its superior performance on episodic memory tasks compared to humans.

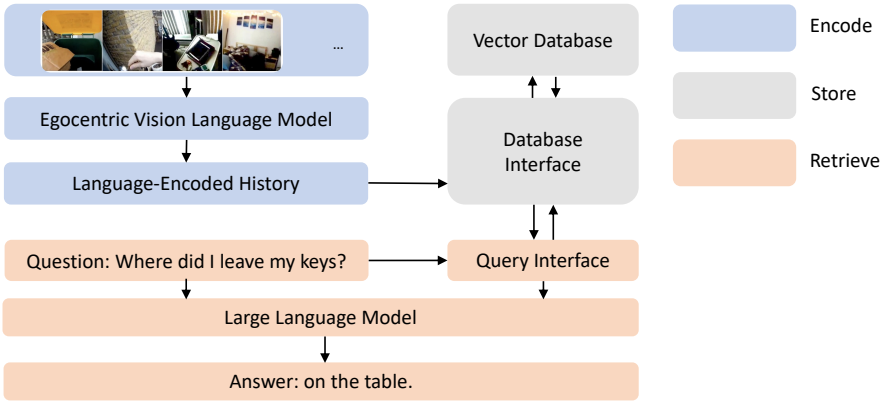

Figure 1: The memory augmentation system operates by first encoding egocentric videos into linguistic representations using a bespoke egocentric vision language model. These language-encoded outputs are stored in a buffer, segmented, and transformed into embeddings for storage in a vector database. Upon receiving a user query, like "Where did I leave my keys?", the system generates an embedding of the query. This embedding is used to search the database for relevant chunks using vector similarity algorithms. These chunks are combined with the query to create a prompt for a large language model, which then generates the final response for the user.

2 Memory Augmentation through Life Logging

Lifelogging, present for over 30 years, has evolved with technology (Harvey et al., 2016; Gurrin et al., 2013). Initially, it involved bulky equipment like helmets and battery packs (Wolf et al., 2014), but has since progressed to wearable devices like glasses (Harvey et al., 2016). A key development was Microsoft's SenseCam in 2006, a notable lifelogging device (Microsoft Research, 2004; Doherty et al., 2012). Lifelogging now includes data from GPS, audio, heart rate, emails, calendar events, and social media.

Significant progress has been made in memory augmentation through lifelogging (Harvey et al., 2016). Le et al. (2016) focused on video summaries for memory recall but didn't address data selection challenges. Davies et al. (2015) highlighted privacy and security concerns in pervasive computing but lacked comprehensive solutions. Byrne et al. (2010) developed a method for content relevance in visual lifelogs but it was limited to everyday concepts. Our work uniquely implements a memory augmentation system enabling open-ended episodic memory queries within a wearable headset.

3 Problem Formulation

Natural Language Video Localization (NLVL) and Video Question Answering (VideoQA) are distinct yet related tasks in video content analysis. NLVL focuses on finding a video segment matching a natural language query, requiring the model to understand both video and query context (Krishna et al., 2017; Gao et al., 2017; Regneri et al., 2013; Grauman et al., 2022). VideoQA, on the other hand, involves answering questions based on video content, demanding a deep understanding of the video and the ability to provide precise answers (Lei et al., 2020; Mun et al., 2020; Sun et al., 2021; Miyanishi and Kawanabe, 2021; Le et al., 2020b).

Episodic Memory Question Answering (EMQA) is a specific subtask of VideoQA, introduced by Bärmann and Waibel (2022). It differs in its memory constraints, shifting from offline analysis (VideoQA) to an online algorithm and setting a maximum limit on memory usage for computation, thus suitable for long-term or life-long use.

This paper focuses on EMQA due to its advantages over NLVL and traditional VideoQA. While NLVL produces non-textual output and VideoQA has scalability issues, EMQA offers textual outputs and a constant-size memory constraint, enhancing efficiency for long videos.

4 Methodology

Human memory involves encoding, storing, and retrieving information. Encoding is key for converting information into a format suitable for memory storage. Storage maintains this information until needed, and retrieval accesses and reinstates it into consciousness. Our system mimics these biological memory processes (Zlotnik and Vansintjan, 2019).

Initially, each video frame v is transformed into a language encoding l using the encoding function E,

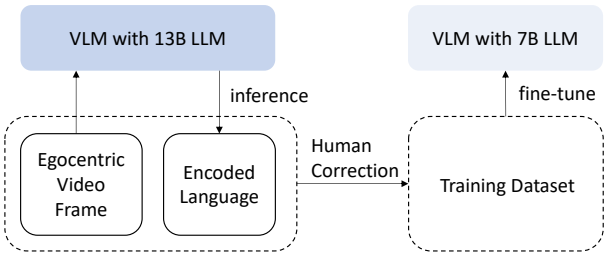

Figure 2: The Egocentric Vision-Language Model is developed through a process called fine-tuning. This process involves extracting knowledge from a large model and transferring it to smaller models, resulting in improved accuracy and faster inference times. The Egocentric Vision-Language Model combines the power of vision and language to effectively process and understand egocentric video data. 13B and 7B refer to large language models with 13 billion and 7 billion parameters.

denoted as $l = E(v)$. These language encodings are accumulated over time, forming a cumulative history L_{history}.

An embedding model acts as a transformation function T to convert chunks C of L_{history} into vector representation g, expressed as $g = T(C)$. The vector g is stored in a vector database via a storage function S, where $S(g) \rightarrow$ Database.

For retrieval, the system uses the same transformation function T to convert a natural language query q into a query vector q_v: $q_v = T(q)$. The retrieval function R then fetches the most relevant language encodings c as context based on q_v, formulated as $R(q_v, \text{Database}) \rightarrow c$. This context c and the query q are concatenated and processed by large language models to generate an answer based on the context.

4.1 Encode

We employ language as a means to encode egocentric visual perceptions. Specifically, our focus lies on adopting a frame-based approach rather than encoding clips using a sliding window. This decision stems from the fact that encoding clips using a sliding window can result in excessively long inference times, rendering it impractical for real-time usage. Furthermore, the field of video captioning is still in its early stages, and even state-of-the-art models are unable to provide accurate and detailed encodings (Xu et al., 2023; Zhang et al., 2023; Xu et al., 2016; Wang et al., 2022). Consequently, we opt to encode videos by their individual frames.

We present a novel model, Ego-LLaVA, for egocentric video encoding. This model is fine-tuned from LLaVA (Large Language and Vision Assistant) (Liu et al., 2023) on egocentric data, which captures first-person experiences in a 3D environment. This fine-tuning procedure leads to better performance in understanding first-person data which involves interpreting human-object interactions and complex social behaviors.

To tackle this issue, we curated our own egocentric video frame description dataset from Ego4D (Grauman et al., 2022) and fine-tuned the LLaVA model to learn egocentric features.

The fine-tuning process is described below:

- **Training Data**: We begin by employing LLaVA using a descriptor prompt P[1], to generate detailed descriptions for a randomly sampled set of 3,000 images. Subsequently, we engaged three research assistants from our institution to correct the descriptions in scenarios where objects were inaccurately identified or significant objects within the frames were missed. This process results in a collection of 3,000 image/video frame-text/description pairs. It has been observed by Zhu et al. (2023) that a set of 3,000 training pairs is adequate. Zhu et al. (2023) successfully fine-tuned a visual-language model using only 3,500 image-text pairs, which yielded exceptional performance in tasks such as image question answering.
  The practice of training language models using responses generated by larger language models has become increasingly common due to the robustness of these models. Vicuna-13B (Chiang et al., 2023) is an example of a model trained by fine-tuning the LLaMA-13B (Zhang et al., 2023) base model with approximately 70,000 user-shared conversations gathered from ShareGPT (ShareGPT, 2023), a website that collects conversational data from OpenAI ChatGPT. Similarly, MiniGPT-4 (Zhu et al., 2023) and LLaVA (Liu et al., 2023) are trained using large language model-generated content, achieving state-of-the-art results and saving significant time on human labeling.
- **Fine-Tuning**: In our experiment, as shown in Figure 2, we use Vicuna-13B (Chiang et al., 2023) as the 13B language model and MPT-7B (MosaicML, 2022) as the 7B language model. MPT (MosaicML Pretrained Transformer) is optimized for efficient training and fast inference, utiliz-

[1]See the full prompt in Appendix A.1

ing FlashAttention (Dao et al., 2022) and FasterTransformer (NVIDIA, 2023) techniques.

More specifically, Ego-LLaVA is fine-tuned on image-text pairs where the descriptor question P prompts a description of the video frame v, and the ground truth prediction answer l is the original detailed description. During training, the weights of both the visual encoder and LLM are kept constant, and the probability $p(l|v, P)$ of the target answers l is maximized by only training the parameters of the linear projection layer between the visual encoder and the LLM. This process allows for the alignment of the video frame features H_v with the pre-trained LLM word embedding.

4.2 Store

A vector database stores data as vectors $\mathbf{g}$ with each element g representing a data attribute (Core, 2023; Pinecone, 2022). This can enhance Large Language Models (LLMs) by storing and retrieving vector representations $\mathbf{g}$ for long-term memory retention and contextually relevant responses (Stata et al., 2000; Chase, 2022).

Our approach involves:

1. **Chunking:** Break the language-encoded history into smaller, manageable chunks C. Our approach employs fixed-size chunking. We set the target size of each text chunk in tokens to $T = 1024$, and the overlap chuck size to $O = 256$. The minimum size of each text chunk is set to $M = 350$. We discard chunks C shorter than five characters.

2. **Associate Metadata:** We associate metadata M, including frame number F and time t, with the chunks C. The metadata M can be used by Vector Database for more advanced searches.

3. **Vector Embeddings:** Create vector embeddings $\mathbf{g}$ for the segmented chunks C, capturing their semantic meaning. We utilize OpenAI model, *text-embedding-ada-002*, for the extraction of vector embeddings $g = T(c), c \in C$.

4. **Storage:** Store the vector embeddings $\mathbf{g}$ in a vector databas. We use Chroma (Core, 2023), which serves as a vector database solution, providing the capability to store, search, and access vector data on a large scale. The storage function is expressed as $S(\mathbf{g}) \rightarrow$ Database, where S is the storage function.

4.3 Retrieve

The retrieval system functions as follows: It starts by taking the user's input question q and employing the OpenAI model, *text-embedding-ada-002*, to generate an embedding for this question. This process is expressed as $q_v = T(q)$. This resultant embedding vector q_v is subsequently used to make a query to the Database Interface, aiming to find relevant chunks l related to the question. The retrieval function is formulated as $R(q_v, \text{Database}) \rightarrow l$. The query vector q_v does not necessarily have to exactly align with the vectors in the database, as the database engine from Chroma (Core, 2023) is capable of efficiently extracting data indexed by vectors that bear close resemblance.

First, chunks of context data c from a vector database that are semantically related to the query q are obtained. These context chunks c, along with the query q, are then inserted into a prompt template: *You are my AI assistant to support memory augmentation. Use the following pieces of context to answer the question at the end. {context} Question: {question}*. The placeholders *{context}* and *{question}* represent the acquired chunks c and the user's query q, respectively. Next, a chain of thought prompt is constructed, incorporating the prompt template, and presented to the OpenAI GPT-4.

5 Evaluation of the Memory Augmentation System on QA-Ego4D

To study the proposed memory augmentation system's performance we carry out a large-scale quantitative evaluation using the public dataset QA-Ego4D. The evaluation focuses on the EMQA task, which is detailed in Section 3.

5.1 Dataset - QA-Ego4D

The QA-Ego4D dataset, an extension of the Ego4D dataset's Natural Language Query (NLQ) subtask, features egocentric videos paired with natural language questions, answers, and annotations for answer-relevant video segments (Bärmann and Waibel, 2022; Grauman et al., 2022). Each video averages eight minutes in length. The dataset includes 19.2K queries from 227 hours of video across 34 scenarios from ten universities. Queries average 8.3 words, with response windows averaging 9.3 seconds, presenting a search challenge. The dataset omits "When?" questions due to undefined natural language answers.

It's divided into training, validation, and test sets, with 997 training videos, 162 for validation, and 166 for testing, comprising 10,746, 1,913, and

Model	BLEU	METEOR	ROUGE
DNC (Graves et al., 2016)	3.4 ± 0.27	17.9 ± 2.15	27.0 ± 3.24
STM (Le et al., 2020a)	5.8 ± 0.81	17.6 ± 1.32	26.2 ± 3.93
LT-CT (Rae et al., 2019)	5.3 ± 0.53	18.5 ± 1.85	27.5 ± 3.30
RM (Zhang et al., 2021)	4.5 ± 0.63	17.7 ± 2.66	26.6 ± 3.99
Language-Encoded QA (with EVLP (Lin et al., 2022))	4.3 ± 0.60	17.2 ± 1.72	27.0 ± 3.51
Language-Encoded QA (with Video-LLaMA (Zhang et al., 2023))	5.8 ± 1.11	19.3 ± 2.32	30.7 ± 4.60
Language-Encoded QA (with LLaVA (Liu et al., 2023))	7.4 ± 1.25	36.1 ± 5.42	50.7 ± 7.60
Language-Encoded QA (With Ego-LLaVA)	**8.3** ± 0.86	**42.3** ± 5.35	**54.7** ± 6.21

Table 1: EMQA results on the QA-Ego4D test set with standard deviations.

Category	Template	BLEU
Objects	Where is object X before / after event Y?	8.7 ± 0.87
	Where is object X?	8.9 ± 1.12
	What did I put in X?	7.6 ± 0.76
	How many X's? (quantity question)	7.4 ± 1.11
	What X did I Y?	8.3 ± 0.83
	In what location did I see object X?	8.5 ± 1.28
	What X is Y	8.0 ± 0.80
	State of an object	7.8 ± 1.17
	Where is my object X?	9.0 ± 1.35
Place	Where did I put X?	8.2 ± 0.98
People	Who did I interact with when I did activity X?	8.1 ± 1.22
	Who did I talk to in location X?	8.4 ± 1.26

Table 2: The templates span a wide range of inquiries that individuals can make use of to enhance their memory, and retrieve information about various objects, locations, and individuals they encounter in their daily lives. We also show the average BLEU score with standard deviations for the proposed memory augmentation system for each template.

1,854 question-answer pairs for each set respectively. The test data uses half of the validation set's canonical videos, as Ego4D's test data is unpublished.

5.2 Baseline Models

In our comparison, we include models from the QA-Ego4D paper (Bärmann and Waibel, 2022): Differentiable Neural Computer (DNC) (Graves et al., 2016), Self-attentive-Associative-Memory-based Two-memory Model (STM) (Le et al., 2020a), Long-Term Comprehensive Transformer (LT-CT) (Rae et al., 2019), and Rehearsal Memory (RM) (Zhang et al., 2021).

We also employ alterations to the encoding methods:

- Language-Encoded QA (with EgoVLP) Lin et al. (2022): The EVLP (Egocentric Video-Language Pretraining) model is a dual-encoder system for egocentric video-language pretraining, using separate video and text encoders. This model is optimized for tasks involving egocentric videos and their associated text.

- Language-Encoded QA (with Video-LLaMA (Zhang et al., 2023)): We employ a sliding window approach with a width and stride of 6 seconds each to encode video clips into language. Video-LLaMA is a state-of-the-art video QA model which is suitable for video captioning.

- Language-Encoded QA (with LLaVA (Liu et al., 2023)): We use the original LLaVA as the encoding method.

We prompt both the above two models using the same prompt as for Ego-LLaVA. These contrasting models provide a comprehensive comparison for the model proposed in this paper.

5.3 Evaluation Metrics

We report standard Natural Language Processing metrics for EMQA tasks, including BLEU-4 (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and ROUGE-L (f-score) (Lin, 2004).

5.4 Results

Our System Achieved State-of-the-Art in QA-Ego4D dataset: Table 1 demonstrates that our language-encoded method surpasses not only conventional vision-based machine learning models on the QA-Ego4D test set, but also exceeds the performance of other pre-trained vision-language models utilizing the language encoding technique. Notably, Ego-LLaVA exceeds the original LLaVA

in encoding effectiveness, likely due to LLaVA's fine-tuning with egocentric data. Despite a small fine-tuning dataset of 3,000 image-text pairs, the model aligns image and text features well. Nevertheless, employing Video LLaMA as an encoder yields subpar performance due to problems with hallucination. Similarly, the backbone model used by EVLP is not adequately equipped to describe images in fine-grained detail, resulting in lackluster performance when using the language encoding technique.

Varied System Performance Across Question Templates: Table 2 shows the system's performance on various question templates. We observe that simpler questions, such as "Where is object X?" or "Where did I put X?", generally yield better results due to their straightforward nature, demanding less complex reasoning from the system. Conversely, questions involving more intricate reasoning or understanding of dynamic elements, such as "Where is object X before / after event Y?" or "What X did I Y?", may not perform as well. This is attributed to the current encoding method's limitations in capturing temporal correlations, which are crucial for comprehending dynamic activities. Quantity-based questions, such as "How many X's?", pose a challenge due to the encoding model's resolution limitations, making accurate object counting difficult.

Questions about the state of an object could also be challenging if the state involves fine details, or dynamic elements that change over time. Without the ability to apply attention to the data, the system might not capture these dynamic subtleties, leading to a significant loss of crucial information. In essence, the system's performance on different question templates largely depends on the question's complexity, the required level of detail, and the system's ability to understand dynamic elements and temporal correlations. Future improvements in these areas could potentially enhance the system's performance on more complex question templates.

6 Human User Study of the Memory Augmentation System

Having established quantitative performance benefits of the memory augmentation system in theory it is natural to ask whether the system is usable in practice. To this end, we carried out a user study with two objectives. The first objective is to evaluate and contrast the performance of human participants with that of the memory augmentation system in answering a set of episodic memory questions. The second objective is to explore the framework's capability in handling open-ended questions, which could potentially demand strong reasoning power and access to an external knowledge base. We did not incorporate another memory augmentation system as a comparison because we are the first to propose such a memory augmentation framework and hence no such baseline exist. The conventional machine learning models have very limited usability as suggested by the results of the large-scale evaluation we described previously.

6.1 Study Protocol

Participants We recruited a total of 12 participants using opportunity sampling to take part in the study (average age = 26.7, sd = 5.2; 7 males and 5 females). A G*Power's analysis (Faul et al., 2007) based on a t-test suggested a sample size of 12 as being adequate for the study based on an effect size of 0.81 (calculated from the collected results), an error probability of 0.05, and a power of 0.8. Among the 12 participants in the user study, four were students, five were employed, and three were self-employed. Each participant is compensated with Amazon vouchers worth 10 GBP.

Materials The study used a HoloLens 2 device which has an inbuilt front camera to stream egocentric videos. The encoding, storage, and retrieval tasks were performed by calling APIs hosted on our server.

Procedure The study had two stages. In the first, participants used a HoloLens 2 for tasks in a house setting: viewing a painting, switching TV channels, cooking eggs, reading a book, and selecting a movie. After 5-7 days, reflecting the "forgetting curve" concept by Rivera-Lares et al. (2022), they entered the second stage, answering questions related to these tasks and others not directly linked but relevant to the scenarios.

Participants also asked the memory augmentation system five open-ended questions and rated both the system's and their own responses on a 1-5 Likert scale. They completed post-study surveys on their experience, assessing the system's value, accuracy, and creativity, and their willingness to use always-on camera for memory augmentation.

Additionally, they are also asked open-ended

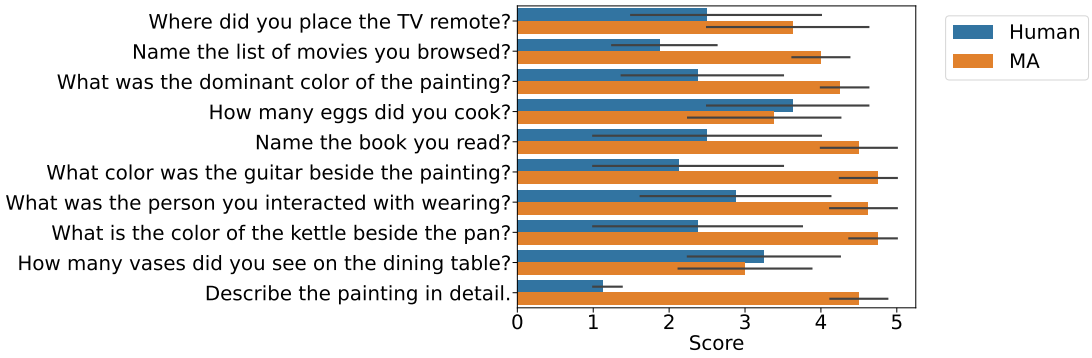

Figure 3: Comparative analysis of scores for the memory augmentation (MA) system and Human responses across various questions. Each question has multiple pairs of AI and human scores represented by the bars. The x-axis enumerates different questions, while the y-axis shows the scores ranging from 1 to 5. The bars are color-coded, with one color representing AI and another representing human scores. The legend on the top-right corner outside the plot area distinguishes between AI and human bars.

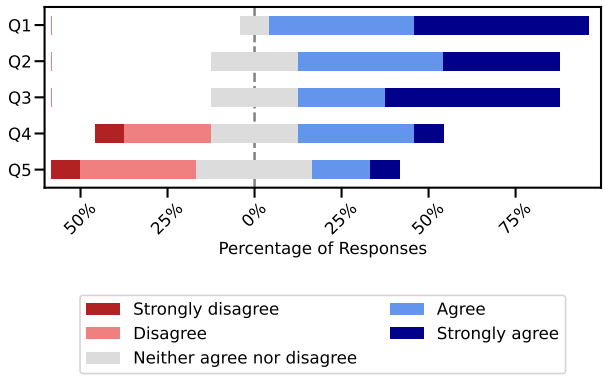

Figure 4: The five-point Likert responses to the post-study questionnaire. Q1. The memory augmentation capability is valuable; Q2. The information provided by the memory augmentation system is accurate; Q3. The response to my open-ended question by the memory augmentation system is creative; Q4. I am willing to wear an always-on camera for memory augmentation through language encoding; Q5. I am willing for others in my close vicinity to wear an always-on camera for memory augmentation through language encoding.

questions about using the feature, concerns, suggested improvements, and other feedback.[2]

6.2 Results

Episodic Memory Questions Figure 3 shows that the memory augmentation system generally outperforms human memory in episodic memory questions. The scoring correlates with response accuracy; for example, a precise color or quantity answer scores higher. The system excels in detailed descriptions, like of paintings, but struggles in tasks like counting or differentiating actions, such as determining the placement of objects.

Humans perform well in specific tasks but often forget details not directly related to their main focus, like the color of a kettle or a list of movies. The memory augmentation system is particularly useful in memory-intensive tasks, detailed descriptions, and overlooked details.

Statistically, the system's mean response score of 4.13/5 surpasses human responses at 2.46/5, with a Friedman's test indicating a significant difference ($\chi^2 = 37.928$, p = 0.0009). The system shows increased stability and consistency in responses with a smaller standard deviation compared to human memory.

Open-Ended Questions Participants gave the memory augmentation system's responses to open-ended questions an average rating of 3.97/5 with a standard deviation of 0.604. This reflects the benefit of integrating OpenAI GPT-4 as a conversational interface with an external knowledge base, enhancing its effectiveness in context-aware responses. Common queries included movie or book recommendations, with responses like suggesting "The Godfather" or "The Dark Knight" typically receiving high ratings of 5/5. Questions about tasks, like improving egg cooking, also received positive feedback, with the system providing detailed steps and showing awareness of the user's specific circumstances, often rated between 4/5 and 5/5.

Post-Study Survey Figure 4 shows participant responses to the memory augmentation system, categorized as Strongly Disagree to Strongly Agree, covering aspects like capability, accuracy, creativity, and willingness to use. The system is generally

[2]See Appendix B.1 for additional details.

seen as accurate, creative, and valuable, particularly for typical EMQA questions and open-ended responses.

Participants in open-ended surveys recognized the system's value in various settings like biomedical experiments and conferences. However, concerns were raised about social awkwardness, ethical implications of over-reliance on the system for memory, and privacy issues, particularly regarding recording others' actions. Some worries subsided as participants understood the language encoding function, but apprehensions remained for a few.

To address these issues, participants suggested using indicators for the system's operation, limiting its usage to specific scenarios, and making the camera device more discreet to enhance social acceptance.[3]

7 Discussion

In this paper we have demonstrated a novel memory augmentation system and demonstrated its performance. Besides being lightweight, as it is reliant on language-encoding as opposed to vision-based, we discuss three additional advantages of this approach: performance, privacy, and device agnosticism.

Ego-LLaVA Paried with Language Encoding Achieved State-of-the-Art in QA- Ego4D The system outperformed traditional models in the QA-Ego4D dataset evaluation with a BLEU score of 8.3. This superior performance in the EMQA task is due to addressing memory constraints that require compression of information into a fixed-size representation, specialized modeling beyond the capabilities of simpler models like STM, and effective relevance selection within the limited memory space, challenges that baseline models struggle with.

Language Encoding is Lightweight In this paper, the Language Encoding Approach and the Vision-based Approach are compared. Language Encoding stores textual data from video, requiring around 0.517 TB/year uncompressed, reduced to 0.246-0.345 TB with compression, while the Vision-based Approach needs 5.74 TB/year for low bitrate 720p video.

Language Encoding is Private Privacy concerns in life logging and memory augmentation systems, as highlighted in the post-study surveys in Section 6.2, are critical. Video data, rich in detail, is challenging to sanitize without compromising content. Language-encoded systems, conversely, can more easily anonymize or remove private information, maintaining information quality. Such systems can inherently prioritize privacy by excluding sensitive details during encoding, preserving both privacy and utility. Furthermore, hardware design considerations for privacy, as discussed in Section 6.2, can enhance user trust and consent.

Language Encoding is Device Agnostic For question and answering, the language encoding approach introduced in this paper gives rise device agnosticism due to its design. This contrasts with vision-based QA models which may exhibit diminished accuracy, or necessitate fine-tuning, when transitioning across different devices. Moreover, the device-agnostic nature is carried through in our language encoding model. The novel egocentric vision language model we introduced in this paper is cultivated using a diverse array of devices including GoPro, Vuzix Blade, Pupil Labs, ZShades, ORDRO EP6, iVue Rincon 1080, and Weeview, to capture egocentric videos. This breadth of training sources fortifies our framework's compatibility with any device capable of delivering egocentric video streaming. While we use the HoloLens 2 as the AR headset, its usage is solely as a conduit for streaming egocentric videos, further illustrating the adaptability of the model.

8 Conclusion

In conclusion, our research presents a novel memory augmentation system that employs a fine-tuned vision language model, Ego-LLaVA, on egocentric vision data for accurate language encoding. This system, combined with a vector database, enables efficient data storage and retrieval. Demonstrating superior performance in the EMQA benchmark, specifically the QA-Ego4D dataset, with a BLEU score of 8.3, it outperforms conventional vision-based models, and other pre-trained vision language models paired with language encoding techniques. A user study further confirmed its effectiveness in episodic memory tasks, surpassing human participants. These findings underscore the potential of our system for real-world applications in enhancing human memory.

[3] See Appendix B.2 for additional details.

Limitation

We introduce an encoding method that operates on an individual frame basis. However, this method struggles to capture temporal correlations, which are essential for understanding dynamic features like activities. Figure 3 shows the lower performance on the two questions related to dynamic elements, including "place" and "cook". Such movements or scene changes, are better understood when temporal correlations between frames are considered. Without this, the encoding method may miss these dynamic subtleties, leading to a significant loss of crucial information. Activities usually occur over a series of frames. Ignoring temporal correlations can make it challenging to fully understand these activities. For example, the action of a person picking up an object involves a sequence of movements across several frames. Despite these shortcomings, this encoding method excels in capturing static features, as each frame is encoded separately.

Ethics Statement

This work reports on using language-encoded egocentric perception to build an accurate memory augmentation system. There are no perceived ethical risks associated with this work.

References

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Leonard Bärmann and Alex Waibel. 2022. 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*, pages 1560–1568.

Daragh Byrne, Aiden R Doherty, Cees GM Snoek, Gareth JF Jones, and Alan F Smeaton. 2010. Everyday concept detection in visual lifelogs: validation, relationships and trends. *Multimedia Tools and Applications*, 49:119–144.

Harrison Chase. 2022. Langchain. [Online; accessed 25-May-2023].

Yi Chen and Gareth JF Jones. 2010. Augmenting human memory using personal lifelogs. In *Proceedings of the 1st augmented human international conference*, pages 1–9.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Chroma Core. 2023. Chroma-core/chroma: A vector database. Online. Accessed: 2023-06-01.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359.

Nigel Davies, Adrian Friday, Sarah Clinch, Corina Sas, Marc Langheinrich, Geoff Ward, and Albrecht Schmidt. 2015. Security and privacy implications of pervasive memory augmentation. *IEEE Pervasive Computing*, 14:44–53.

DeepSpeed Team. 2021. Deepspeed zero-3 offload. `https://www.deepspeed.ai/2021/03/07/zero3-offload.html`. Accessed: 2023-12-15.

Tilman Dingler, Passant El Agroudy, Rufat Rzayev, Lars Lischke, Tonja Machulla, and Albrecht Schmidt. 2021. Memory augmentation through lifelogging: opportunities and challenges. *Technology-Augmented Perception and Cognition*, pages 47–69.

Aiden R Doherty, Katalin Pauly-Takacs, Niamh Caprani, Cathal Gurrin, Chris JA Moulin, Noel E O'Connor, and Alan F Smeaton. 2012. Experiences of aiding autobiographical memory using the sensecam. *Human–Computer Interaction*, 27(1-2):151–174.

Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. *G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences*.

J. Gao, C. Sun, Z. Yang, and R. Nevatia. 2017. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012.

Alex Graves, Greg Wayne, Malcolm Reynolds, Tim Harley, Ivo Danihelka, Agnieszka Grabska-Barwińska, Sergio Gómez Colmenarejo, Edward Grefenstette, Tiago Ramalho, John Agapiou, et al. 2016. Hybrid computing using a neural network with dynamic external memory. *Nature*, 538(7626):471–476.

Cathal Gurrin, Rami Albatal, Hideo Joho, and Kaori Ishii. 2014a. A privacy by design approach to lifelogging. In *Digital enlightenment yearbook 2014*, pages 49–73. IOS Press.

Cathal Gurrin, Alan F Smeaton, Aiden R Doherty, et al. 2014b. Lifelogging: Personal big data. *Foundations and Trends® in information retrieval*, 8(1):1–125.

Cathal Gurrin, Alan F Smeaton, Zhengwei Qiu, and Aiden R Doherty. 2013. Exploring the technical challenges of large-scale lifelogging. *ACM Transactions on Intelligent Systems and Technology*.

Morgan Harvey, Marc Langheinrich, and Geoff Ward. 2016. Remembering through lifelogging: A survey of human memory augmentation. *Pervasive and Mobile Computing*, 27:14–26.

Gillian R Hayes, Shwetak N Patel, Khai N Truong, Giovanni Iachello, Julie A Kientz, Rob Farmer, and Gregory D Abowd. 2004. The personal audio loop: Designing a ubiquitous audio-based memory aid. In *Mobile Human-Computer Interaction-MobileHCI 2004: 6th International Symposium, MobileHCI, Glasgow, UK, September 13-16, 2004. Proceedings 6*, pages 168–179. Springer.

Steve Hodges, Lyndsay Williams, Emma Berry, Shahram Izadi, James Srinivasan, Alex Butler, Gavin Smyth, Narinder Kapur, and Ken Wood. 2006. Sensecam: A retrospective memory aid. In *UbiComp 2006: Ubiquitous Computing: 8th International Conference, UbiComp 2006 Orange County, CA, USA, September 17-21, 2006 Proceedings 8*, pages 177–193. Springer.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Margaret Jean Intons-Peterson. 2014. External memory aids and their relation to memory. In *Cognitive psychology applied*, pages 145–168. Psychology Press.

Margaret Jean Intons-Peterson and George L Newsome III. 1992. External memory aids: Effects and effectiveness. In *Memory improvement: Implications for memory theory*, pages 101–121. Springer.

Stanley B Klein. 2015. What memory is. *Wiley Interdisciplinary Reviews: Cognitive Science*, 6(1):1–38.

R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. Carlos Niebles. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

H. V. Le, S. Clinch, C. Sas, T. Dingler, N. Henze, and N. Davies. 2016. Impact of video summary viewing on episodic memory recall – design guidelines for video summarizations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 4793–4805. ACM.

Hung Le, Truyen Tran, and Svetha Venkatesh. 2020a. Self-attentive associative memory. In *International Conference on Machine Learning*, pages 5682–5691. PMLR.

T. Le, H. Nguyen, L. Duan, D. Q. Pham, and M. L. Shyu. 2020b. Hierarchical conditional relation networks for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

J. Lei, L. Yu, T. Berg, and M. Bansal. 2020. Tvqa+: Spatio-temporal grounding for video question answering. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 8211–8225.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81. Association for Computational Linguistics.

Kevin Qinghong Lin, Alex Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Zhongcong Xu, Difei Gao, Rongcheng Tu, Wenzhe Zhao, Weijie Kong, et al. 2022. Egocentric video-language pretraining. *arXiv preprint arXiv:2206.01670*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Microsoft Research. 2004. Sensecam. `https://www.microsoft.com/en-us/research/project/sensecam/`. (Accessed: Day Month Year).

T. Miyanishi and M. Kawanabe. 2021. Watch, listen, and answer: Open-ended videoqa with modulated multi-stream 3d convnets. In *European Signal Processing Conference (EUSIPCO)*, pages 706–710.

MosaicML. 2022. Mpt: Model parallel transformers from 100 million to 7 billion parameters.

J. Mun, P. H. Seo, I. Jung, and B. Han. 2020. Just ask: An interactive learning framework for vision and language navigation. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

NVIDIA. 2023. Fastertransformer.

OpenAI. 2023. Gpt-4: The fourth generation of generative pre-trained transformers. *OpenAI Blog*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Pinecone. 2022. Pinecone: A vector database for similarity search and recommendation. *Pinecone Documentation*.

Jack W Rae, Anna Potapenko, Siddhant M Jayakumar, and Timothy P Lillicrap. 2019. Compressive transformers for long-range sequence modelling. *arXiv preprint arXiv:1911.05507*.

M. Regneri, M. Rohrbach, D. Wetzel, S. Thater, B. Schiele, and M. Pinkal. 2013. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1:25–36.

Karim Rivera-Lares, Robert Logie, Alan Baddeley, et al. 2022. Rate of forgetting is independent of initial degree of learning. *Memory & Cognition*, 50:1706–1718.

ShareGPT. 2023. Sharegpt.

Raymie Stata, Krishna Bharat, and Farzin Maghoul. 2000. The term vector database: fast access to indexing terms for web pages. *Computer Networks*, 33(1-6):247–255.

G. Sun, L. Liang, T. Li, B. Yu, M. Wu, and B. Zhang. 2021. Video question answering: A survey of models and datasets. *Mobile Networks and Applications*, 26(5):1904–1937.

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*.

Katrin Wolf, Albrecht Schmidt, Agon Bexheti, and Marc Langheinrich. 2014. Lifelogging: You're wearing a camera? *IEEE Pervasive Computing*, 13(3):8–12.

Haiyang Xu, Qinghao Ye, Ming Yan, Yaya Shi, Jiabo Ye, Yuanhong Xu, Chenliang Li, Bin Bi, Qi Qian, Wei Wang, et al. 2023. mplug-2: A modularized multi-modal foundation model across text, image and video. *arXiv preprint arXiv:2302.00402*.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-llama: An instruction-finetuned visual language model for video understanding.

Zhu Zhang, Chang Zhou, Jianxin Ma, Zhijie Lin, Jingren Zhou, Hongxia Yang, and Zhou Zhao. 2021. Learning to rehearse in long sequence memorization. pages 12663–12673. PMLR.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Gregorio Zlotnik and Aaron Vansintjan. 2019. Memory: An extended definition. *Frontiers in psychology*, 10:2523.

A Training and Implementation Details

A.1 Prompt to Ego-LLaVa to Encode Video Frames

Describe the image in detail. Start with a high-level description: Begin by providing an overall description of the image, capturing its main subject or scene. Describe the visual elements: Break down the image into its key visual elements and describe them in detail. Provide context and additional details: Include any relevant context or details that enhance the understanding of the image. Use spatial relationships: Describe the arrangement of objects within the image, their relative sizes, and their positions. Consider proportions and scale: Specify the proportions and scale of various elements to ensure their accurate representation. This includes the size of objects, distances between them, and any other relevant measurements. Finally, avoid fabricating information.

A.2 Ego-LLaVA Fine-tuning Details

For the fine-tuning process, we employ a cluster of eight A100 GPUs. The whole training process takes around 10 hours. Our training protocol benefits from the use of DeepSpeed ZeRO-3 (DeepSpeed Team, 2021). Given the substantial computational demands, we incorporate LoRA (Hu et al., 2021), enabling the training process to fit within the constraints of eight A100-40G GPUs. The pretraining model is designed with a maximum text length of 1024 and an image size of 448. It incorporates gradient checkpointing and is tailored to operate with a chat template. The model parameters are loaded from `liuhaotian/LLaVA-Lightning-MPT-7B-preview` in Hugging Face, and the model is further augmented with LoRA, which is configured with a radius of 64 and an alpha of 16. The learning rate schedule adheres to the linear warmup cosine learning rate strategy. It starts with an initial learning rate of 1e-5, a minimum learning rate of 8e-5, and a warmup learning rate of 1e-6. The weight decay is set at 0.05, and the training process is designed to run for up to 50 epochs. Our training setup comprises six workers, with each epoch consisting of 1000 iterations and a warmup step of 1000. To ensure reproducibility, the training process is seeded at 66. Furthermore, the training process is configured to utilize mixed-precision training (amp) for optimized performance.

A.3 Memory Augmentation System Implementation Details

The process begun with the uploading of the video to the server. To optimize the speed of the encoding process, we extracted four frames per second from the video. To expedite this process and approach near real-time encoding, we used

multi-process threading. This technique allowsed multiple encoding tasks to be executed simultaneously, significantly reducing the overall time required. Once the frames were encoded, they were stored in a vector database. Vector databases, such as Pinecone (Pinecone, 2022) and Chroma (Core, 2023), are designed to handle high-dimensional vector data, making them ideal for this purpose. In our case, we decided to use Chroma as it is an open-source project. We used LangChain (Chase, 2022) as the implementation framework and OpenAI GPT-4 as the large language model to implement the conversational AI assistant that performs question-answering for memory augmentation.

B Human User Study

B.1 Study Protocol

Firstly, we have issued risk disclaimers to the participants. The study consisted of two stages. During the first stage, participants were equipped with a HoloLens 2 device and were instructed to perform a series of tasks. These tasks were divided into five different scenarios: (1) looking at a painting in a living room; (2) switching TV channels in a living room; (3) cooking eggs in a kitchen; (4) reading a book in a study room; and (5) selecting a movie on a laptop. These tasks took place in an actual house equipped with a variety of furniture and items. Participants were encouraged to freely engage in the tasks to simulate a normal daily life experiences.

Between five and seven days later, participants proceeded to the second stage of the study. This delay was used based on the concept proposed by Rivera-Lares et al. (2022). They suggest that after a week, the amount of retained information may have diminished to a level referred to as the "floor", making it challenging to detect or observe any additional instances of forgetting. This can be represented by the forgetting curve, which hypothesizes a decline in memory retention over time in the absence of deliberate attempts to retain information.

During the second stage of the study the participants were presented with a set of questions modeled after Table 2, which were derived from the tasks they performed. Standard questions included "Where did you place the TV remote?", "Name the list of movies you browsed on the laptop?", "What was the dominant color of the painting you observed?", "How many eggs did you cook?", and "Name the book you read?". In addition to these task-related questions, there were other queries that were not directly linked to the tasks but remained relevant to the scenarios, including "What color was the guitar beside the painting?", "What was the person you interacted with (study facilitator) wearing?", "What is the color of the kettle beside the pan" and "How many vases did you see on the dining table?". An example of a description type question for the third category of queries is the following: "Describe the painting in detail". We asked each participant these ten questions.

In addition, participants were encouraged to ask the memory augmentation system five open-ended questions through a interactive conversational interface. Examples of questions were "What movie would you recommend for next time?", "Based on what you know, do you think I eat healthily, and if not, what suggestions do you have for my diet?", and "What are the steps to better cook an egg?".

Thereafter the participants were asked to rate the responses generated by the system as well as their own answers using a scale ranging from 1 (very bad) to 5 (very good). Note that the order of queries and the order of which responses to score were randomized. Participants were also requested to score the system's responses to open-ended questions using the same scale. We then used post-study surveys to gather feedback from the participants regarding their subjective opinions on the overall experience with the system.

The participants were asked to respond to the following Likert scale questions: (1) The memory augmentation capability is valuable; (2) The information provided by the memory augmentation system is accurate; (3) The response to my open-ended questions by the memory augmentation system is creative; (4) I am willing to wear an always-on camera for language-encoded memory augmentation; and (5) I am willing for others in my close vicinity to wear an always-on camera for language-encoded memory augmentation.

Finally, we asked four open-ended questions: (1) Under what circumstances would you use this memory augmentation feature?; (2) Do you have any concerns regarding the memory augmentation capability?; (3) What improvements would you suggest for the memory augmentation system? and (4) Do you have any other feedback or suggestions regarding the memory augmentation feature?

B.2 Post-Study Survey

In the open-ended surveys, participants emphasized the importance and value of having a memory aug-

mentation system, highlighting various applicable scenarios such as biomedical experiments, conferences, attending lectures/meetings, and exploring new places *(P1, P3, P4, P6, P7, P10, P11, P12)*.

However, participants expressed concerns about the system being socially awkward to wear *(p1, p2, p6)*. Ethical concerns were also raised, such as the potential degradation of people's memorization capabilities if they rely solely on the system *(P4, P10)*. Additionally, some participants highlight privacy issues concerning individuals donning it and recording their actions. For example, *P3* noted that "the system's powerful and accurate capabilities could pose safety risks if breached", while *P2* expressed worries about "discomfort with others wearing the system and recording their activities". However, as the understanding of the system's function through language encoding grew, the majority of concerns diminished *(P1, P3, P5, P6, P7, P8, P12)*, although a few participants still felt uneasy *(P1, P2, P9, P11)*.

Participants proposed specific improvements to address these issues, including incorporating indicators to make people aware of the system's operation and limiting its use to specific scenarios such as teaching and conferences rather than everyday life *(P1, P3, P4, P12)*. Additionally, participants suggested making the always-on camera device as lightweight and inconspicuous as possible to minimize social awkwardness and increase social acceptance *(P2, P5, P6, P8, P9, P11)*.