MemMod4CVQA: Using Memory Models to Infer Causal Relationships for Visual Question Answering on Life-Log Data

Sohil Khan
Department of Electronics and Electrical Engineering,
BITS Pilani K.K. Birla Goa Campus
India

Rakshit Mishra
Department of Electronics and Electrical Engineering,
BITS Pilani K.K. Birla Goa Campus
India

Mehlam Songerwala
Department of Electronics and Electrical Engineering,
BITS Pilani K.K. Birla Goa Campus
India

Meera Radhakrishnan
Data Science Institute,
University of Technology Sydney
Australia

Dulanga Weerakoon
Singapore-MIT Alliance for Research and Technology Centre
Singapore

Vigneshwaran Subbaraju
IHPC, Agency for Science, Technology and Research
Singapore

Sougata Sen
Department of Computer Science and Information Systems,
BITS Pilani K.K. Birla Goa Campus
India

Abstract

Egocentric cameras are increasingly adopted as an effective tool that aids in self-monitoring by individuals. These cameras automatically record event details into a video which are reviewed by the users later, to identify items of interest. However, this process is time-consuming and labor intensive. With the availability of LLMs, we posit that the task of identifying the items of interest can be automated into a natural language Question-Answering format. Furthermore, these LLMs, in conjunction with memory graph models can help identify causal relationship between events. To this end, in this paper, we propose MemMod4CVQA, a framework that enables users to pose questions that the framework will answer using the egocentric camera details. The MemMod4CVQA framework uses a memory model comprising of semantics memory, episodic memory, and causal memory to answer these questions. Through a small simulation-based study we observed that it is possible to realize the framework and obtain an overall F1 score of 44.80% on predicting causal relationships, which is 22.10% higher than standard baseline approaches. We believe that the framework architecture can lead to significant improvements in causal visual question answering in the future.

CCS Concepts

• **Human-centered computing → Ubiquitous and mobile computing systems and tools**; • **Computing methodologies → Knowledge representation and reasoning**.

Keywords

Ubiquitous Sensing; Lifelogging; Visual Question Answering; Causal Reasoning; Large Language Models; Vision Language Models

ACM Reference Format:
Sohil Khan, Rakshit Mishra, Mehlam Songerwala, Meera Radhakrishnan, Dulanga Weerakoon, Vigneshwaran Subbaraju, and Sougata Sen. 2025. MemMod4CVQA: Using Memory Models to Infer Causal Relationships for Visual Question Answering on Life-Log Data. In *Companion of the 2025 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp Companion '25), October 12–16, 2025, Espoo, Finland.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3714394.3756290

1 Introduction

The proliferation of wearable cameras and life logging devices has made capturing continuous streams of daily human activity increasingly feasible. The growing "quantified self" movement has further fueled this interest and encourages individuals to record their everyday experiences by donning a wearable camera [19]. Individuals subsequently review the captured life-logs to identify specific events, which is tedious and time-consuming. Furthermore, one might need to review disjointed life-log streams to determine the causal reasoning for an activity, a cognitively challenging task.

With the advancement in Machine Learning (ML) and the availability of Large Language Models (LLMs), automating the review process to quantify moments quickly is gradually becoming possible. Although researchers have put substantial effort into obtaining causal relationships between events for Visual Question Answering (VQA), several research gaps remain. Current VQA models primarily rely on static frame-level features or short temporal windows. They often do not effectively model causal and episodic information over continuous video feeds acquired over long time-horizons (24/7) [2]. With the rise of egocentric video datasets (e.g., CASTLE dataset [27]), we have an opportunity to extract rich, continuous real-world contexts.

UbiComp Companion '25, October 12–16, 2025, Espoo, Finland

ACM ISBN 979-8-4007-1477-1/2025/10
https://doi.org/10.1145/3714394.3756290

We envision that, in the future, LLMs can leverage memory-based knowledge graphs to infer and reason about causal relationships between events. These memory-based models can improve the performance of personal assistants that employ VQA techniques on *personal big data*, such as visual life logs (videos), by exploiting causal relations between events. In this paper, we introduce *MemMod4CVQA*, a framework that builds on memory models to seamlessly infer causal relationships in life log data to answer VQAs. Such a framework would help generate explainable, context-aware, and reasoned answers to questions about egocentric videos.

Our Vision: The MemMod4CVQA, our framework for causally grounded insight generation from raw egocentric videos, leverages:

- Event Segmentation: Using activity-based temporal clustering to chunk continuous videos into meaningful, discrete events.
- Memory Model: Construct dual memory representations: (a) *Episodic memory*: Raw segmented events with time, location, and context, (b) *Semantic memory*: Abstracted summaries and concepts extracted from episodes.
- Causal graph extraction: Learning or inferring directed causal relationships between events and actions.
- Explainable question answering: Reasoning over structured event representations and their causal relationships.

Our proposed approach is particularly well-suited to ubiquitous sensing environments, where sensor-data streams are heterogeneous, unstructured, and temporally extended. Several categories of application will benefit from MemMod4CVQA in the future. For example, grounding LLMs on causal graphs for a structured understanding of events over time, especially when processing visual or multi-modal data. Human-like understanding requires models to generate possible responses and reason about causes, intentions, and outcomes.

While generative AI models have demonstrated their ability to encode knowledge from vast amounts of data and retrieve information as a response to natural language queries, the relevance of the responses provided by them hinges heavily on the soundness of the prompts/queries provided to them. This has led to the development of several techniques for prompt-engineering, reasoning techniques etc. Understanding the prompt or input query is therefore a crucial step for Gen-AI in producing relevant outputs. This aspect becomes more acute in retrieval and reasoning tasks involving *personal big data* such as visual life logs. A user who wants to query their own personal big data, generally generates queries based on what they know/remember and need answers to things that they may not know/remember (e.g., “Where did I leave my car keys?"). Therefore, a model that provides an estimate of what the user remembers/knows from their past provides crucial context to understand their query, by providing an approximate common-ground (e.g., the user knows that she left the keys somewhere in the house because the car is at the house, but they want to know exactly *where* in the house). Establishing this understanding is crucial to provide relevant and accurate answers as well as optimize computational resources by reducing the search space (e.g., don't look for the key in the office). To facilitate such a mechanism, we envision two important steps– (1) organizing personal data into chunks/segments similar to how the human mind organizes information, and (2) performing causal reasoning to establish spatial and temporal connections between crucial information pieces extracted from chunks of personal data.

2 Related Work

Visual Question Answering (VQA) is a multimodal task that involves generating a natural language answer to a question based on an input image or video. It remains a challenging problem due to the need for joint reasoning across both visual and linguistic modalities, often requiring large, resource-intensive models. Early approaches treated VQA as a multi-stage pipeline. For instance, Multimodal Compact Bilinear (MCB) Pooling employs Convolutional Neural Networks (CNNs) to encode visual inputs and Recurrent Neural Networks (RNNs) for textual feature extraction, combining the two modalities by computing their outer product in a compact form [10, 17, 31]. Over time, the CNNs were replaced by Vision Transformers (ViTs) [8] and RNNs were then replaced with BERT [7, 21] based models in VideoQA [9, 35], to obtain improved representations and benefit from self-supervised cross-modal pretraining (Eg. CLIP-ViT [25]). Recent VQA models have integrated frozen LLMs (e.g., LLaMA [29]) with instruction tuning of projection and adaptation modules. However, despite their superior accuracy, these models tend to be highly resource-intensive, posing challenges in terms of computational cost and scalability as well as lacking the explainability of neuro-symbolic approaches [30]. Consequently, recent LLM based approaches have explored causal and temporal reasoning in VideoQA [18, 36]. However, such efforts have been overwhelmingly focused on videos acquired from a third-person view point and only recently, there has been attention on the ego-centric view point [13].

Visual question answering on long-form ego-centric videos is an important step toward achieving everyday memory support (e.g., answering “Where did I leave my keys") [2]. It is also considered crucial for the personalization of gamified cognitive interventions [14, 33, 34] and memory support [12, 24]. While past work has focused on vision-language modelling, the inclusion of human memory models to provide context is considered important in practical applications that help end-users [32]. For example, it is known that humans tend to have better recall of images acquired from the beginning/end of an event, when compared to the middle [11, 16]. Similarly, the presence of a human face is known to influence the recall of an event. Understanding how humans segment and remember events is hence crucial as it provides context to their queries (which are grounded in what they remember and need answers to what they don't) as well as the answers generated by the systems. This also motivates the need for causal reasoning to retrieve the frames that are relevant to the given query. Another compelling reason to include such memory mechanisms is to reduce the processing overheads involved in answering queries over longer time horizons. Earlier work has shown that clustering the video feed into semantically relevant event-segments has been found to be useful in retrieval tasks [6, 20]. A proper event-segmentation based on human memory models could provide better organization of the video-feed that helps to improve accuracy of retrieval as well as minimize the overheads.

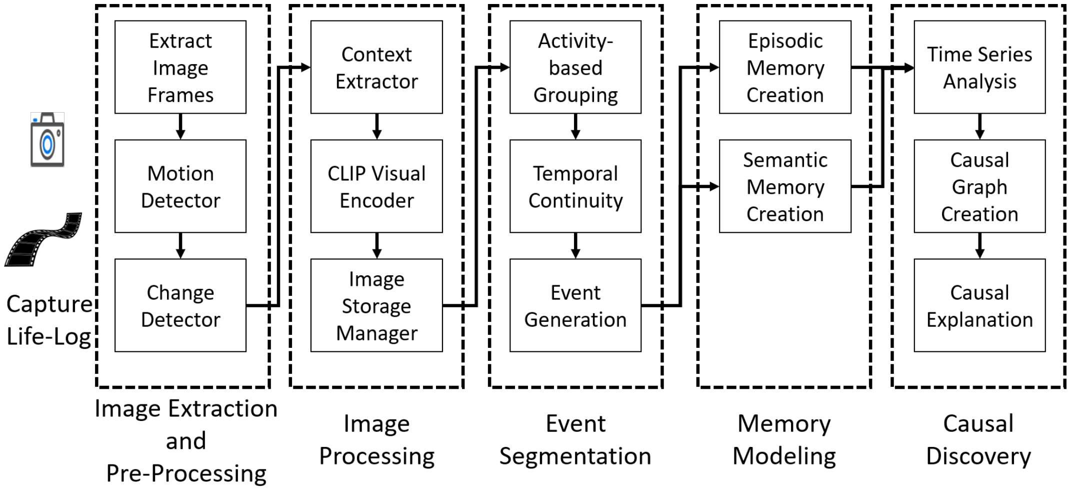

Figure 1: The MemMod4CVQA framework for inferring causal relationships from life-log data for Visual Question Answering.

3 The MemMod4CVQA Framework

We next describe our envisioned MemMod4CVQA framework that can help extract causal relationships in life log data. Figure 1 provides a high level overview of MemMod4CVQA. The MemMod4CVQA framework consists of five primary blocks– image extraction and pre-processing, image processing, event segmentation, memory modeling, and causal discovery. We next describe each block.

Image extraction and pre-processing: Once the wearable camera captures continuous egocentric video, the system extracts the corresponding image frames. The system then employs an adaptive background modeling approach to account for gradual environmental changes while remaining sensitive to sudden movements. This prevents unnecessary processing during non-active situations. The implementation follows the two key steps:

- *Calibration Phase*: At system initialization, 30 frames of a static environment are captured to build a baseline background model (β). This baseline model provides a reference for motion detection and helps compute the Mean Absolute Deviation (MAD) threshold used for adaptive motion sensitivity.
- *Background Update*: The system continuously updates the background model (β) over time using a weighted average controlled by a *LearningRate* parameter ($\lambda = 90$). The background model is a pixel-by-pixel representation of the "normal" static environment, initially established during calibration but continuously evolved over time. It is essentially the same as the calibration baseline but dynamically updated to adapt to gradual environmental changes. The background modeling is done using the following formula, where i represents the pixel index in the frame arrays.

$$\beta[i] = (\beta[i] \times \lambda + currentFrame[i] \times (100 - \lambda))/100 \tag{1}$$

Using the background model computed above, the change detection pipeline involves the following steps: (a) *Absolute Difference Calculation*: Each pixel in the current frame is compared to the corresponding pixel in the background model, (b) *Morphological Processing*: A combination of erosion and dilation operations clean up the difference map, eliminating noise while preserving significant motion regions, (c) *Adaptive Thresholding*: The system employs a MAD-based dynamic thresholding (threshold factor = 2.5), and (d) *Temporal Smoothing*: A 5-frame sliding window is used to stabilize detection and suppress transient changes.

Image Processing: The You Only Look Once (YOLO) model [26] performs classification on the images, identifying up to 5 top objects in each image. A dual-embedding approach combines visual and semantic features using the Contrastive Language-Image Pretraining (CLIP) model: (a) Visual embedding from the image itself, and (b) Textual embedding from the detected object labels. This information is stored along with the captured timestamp.

Event Segmentation: Activity-based temporal clustering implements change-point analysis of multimodal feature distributions, operating on piecewise stationary assumptions where statistical properties remain stable within events but shift at boundaries. The method combines statistical transition detection with LLM semantic validation to ensure that boundaries represent meaningful activity changes rather than noise. Hierarchical decomposition enables the capturing of coarse activity transitions and fine sub-activity boundaries while maintaining semantic coherence of extracted events.

Memory Modelling: Although there are various memory modeling techniques, we focus on using the Self-Memory System (SMS) as described by Conway and Pleydell-Pearce [5]. SMS posits autobiographical memory as transitory mental constructions within a goal-driven control system. The model distinguishes between episodic memory (experience-specific records) and semantic memory (abstracted conceptual knowledge) [4]. Episodic memory is hierarchical in nature consisting of lifetime periods, general events, and specific knowledge. It is built based on objects, location, emotion, and activities observed in a time period. Semantic memory, on the other hand, is based on the abstractive consolidation process, as described in the SMS framework.

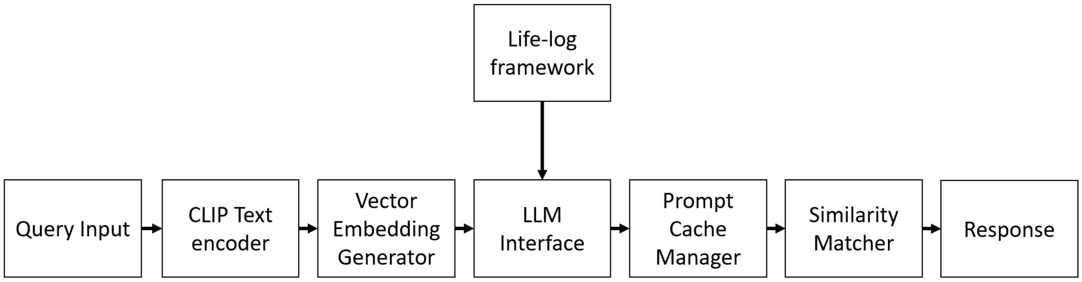

Figure 2: Using the Life-Log Framework to answer queries provided by users

Each segmented life log event instantiates an **episodic memory** object containing: core event data (temporal boundaries, activities, objects, emotions), contextual metadata (frame indices, timestamps), retrieval cues (union of detected concepts), emotional valence, and confidence-weighted vividness scores.

While **semantic memory** emerges through statistical consolidation where concept frequencies and co-occurrences across episodic memories generate association strengths and prototypical feature vectors. This consolidation process creates conceptual knowledge from experiential patterns, enabling temporal and associative memory retrieval pathways essential for causal relationship discovery.

Causal Discovery: Causal Discovery encompasses computational methods for inferring causal relationships from observational data, typically through causal graph learning. Our approach derives these causal relationships by systematically extracting and analyzing the rich relational information embedded within both episodic and semantic memory representations constructed from life log data. In MemMod4CVQA, we adopt a multi-modal structure learning approach tailored to the heterogeneous nature of life logging data. The causal discovery process operates directly on the structured memory representations, where episodic memories provide temporally-ordered, context-rich event sequences, while semantic memory contributes abstracted conceptual associations and co-occurrence patterns learned from multiple episodic instances. To enable robust causal inference, we integrate visual, behavioral, and physiological modalities via a specialized preprocessing pipeline. This pipeline constructs cross-modal interaction features that capture meaningful relationships across modalities, and validates statistical properties to ensure reliable causal inference. Structure learning operates hierarchically, first identifying intra-modal relationships within each data type, then discovering inter-modal causal pathways using constraint-based (PC) and score-based (Hill Climbing with BIC/AIC) algorithms.

4 Visual Question Answering

The system takes in a user query in a text form. For example, the user can ask "*Where did I leave my car key?*". The query is first converted into vector embeddings, which are then compared against stored memory embeddings using cosine similarity. The top-k most relevant memory entries are retrieved based on this similarity and passed to the LLM interface. The MemMod4CVQA framework integrates with this interface to provide contextual grounding for the LLM, ensuring that the response is informed by relevant episodic and semantic memories rather than passing the entire memory model directly. The overall steps that system follows is listed in Figure 2. Two important components in this system are the prompt cache manager and the similarity matcher.

```
"prompts": [
    {
    "prompt": "What was happening at 3 PM yesterday?",
    "embedding": [0.021, -0.085, 0.143, ...],
    "response": "At 3 PM yesterday, the camera detected...",
    "timestamp": "2025-04-03T15:00:00"
    },
    ...
]
```

Figure 3: Example of a JSON for caching prompts.

Prompt Cache Manager: The system incorporates an intelligent prompt caching mechanism to improve response quality and reduce computational overhead for similar queries. This caching mechanism stores each prompt and its corresponding generated output are stored in a structured JSON file as shown in Figure 3.

Similarity Matcher: Upon receiving a user prompt, the system generates a vector embedding for the current prompt. It then computes cosine similarity against three memory banks:

(1) Episodic memory: the time-stamped sequences with the associated metadata and image frames
(2) Semantic memory: abstracted knowledge patterns extracted from the episodic memories
(3) Prompt cache: previous prompts paired with their corresponding event representation.

Any entry in these repositories whose similarity score exceeds 0.85 is deemed relevant, and if there are more than k such entries, then the top-k entries are selected. In our processor, we are using top 5 matching cache entries, and top 10 similar epsiodic, semantic and causal relations. For the top relevant matches, the following information is assembled and concatenated into the LLM's context window prior to response generation:

- The current user prompt
- The most similar past prompt
- The response previously generated for that prompt
- Any linked event metadata and image frames

To balance recency with a bounded memory footprint, the system maintains a prompt-event cache that stores newly generated

Table 1: Comparison of query responses for different modes: Causal-only (C), Episodic-only (E), Semantic-only (S), Causal-Episodic (CE), Causal-Semantic (CS), Episodic-Semantic (ES), and all three combined (All).

Query Examples	Causal	Episodic	Semantic	CE	CS	ES	All
Query Type: Temporal *List the instances when I wrote using a pen and paper*	Writing with pen and paper in living area, afternoon; focused, table present; pen leads to paper	Writing at 08:36 in living area; focused; chair, chocolate, consent form, cup	Writing in living area, afternoon; focused; table usually present; laptop and flowers sometimes	Writing at 08:36-08:39 living area; pen, chocolate, consent form, cup	Writing at table in living area, afternoon; pen and paper present; alone	Writing at 08:36-08:39 in living area; pen, paper, laptop, flowers, cup, mug	Writing at 08:36-08:39 in living area; focused; chair, chocolate, paper, pen, mug present
Query Type: Causal *Why am I walking around in the kitchen?*	Walking in kitchen primarily to prepare food; busy state; kitchen leads to content; alone common	Walking in kitchen to prepare food; gathering ingredients, stove, oven; sometimes with people	Walking in kitchen to prepare food in afternoon; busy state; sink, cabinets, appliances	Walking in kitchen to prepare food; busy; neutral, focused, happy emotions	Walking in kitchen for food prep; content; sometimes alone; relaxed leads to kitchen	Walking kitchen to prepare food; afternoon; gathering ingredients; emotions neutral, focused, happy; people present sometime	Walking in kitchen for food prep; stands precedes kitchen; objects like counter, bottles; emotions content, relaxed
Query Type: Descriptive *What was I cooking?*	Preparing food in kitchen, afternoon, focused, alone, standing, food linked to focus and short duration	Cooking and food prep in kitchen, focused, with person_2 and person_3, fruit, pineapples, oven, stove, cabinets	Preparing food, afternoon, focused, kitchen, oven and appliances present	Cooking and preparing food in kitchen various times, focused, sometimes neutral or happy	Preparing food, afternoon, focused, kitchen, content, alone, standing	Cooking, preparing food in kitchen, afternoon, fruit, pineapple, stove, oven, happy and focused sometimes	Cooking and preparing food in kitchen, fruit, pineapple, oven, countertop, bottles, cookware, f ocused, happy, content, relaxed
Query Type: Event-based *Where do I usually do my writing activities?*	Living area, afternoon, table and paper, cognitive activity, very short duration	Living area, focused, chair, chocolate, consent form, cup	Living area, afternoon, table, paper, laptop, focused	Living area, focused, chair, chocolate, consent form, cup, possibly specific spot	Living area, afternoon, table, paper, laptop, pen, food, alone	Living area, afternoon, table, chair, paper, pen, consent form, cup, mug, flowers, laptop, TV, shelves, decorations	Living area, focused, chocolate, chair, consent form, cup, pen, mug, paper, paper leads to pen

prompt–event–response tuples. Each new entry is inserted into the cache with a time-decay weighting, which gradually reduces the similarity contribution of older entries. To prevent unbounded growth, the system employs a Least Recently Used (LRU) eviction policy that removes the least recently accessed items once a predefined capacity is reached. This unified retrieval and context augmentation strategy allows the LLM to ground its responses in both up-to-date episodic details and distilled semantic insights, while ensuring the memory remains manageable and efficient.

5 Preliminary Evaluation & Results

We conduct a preliminary evaluation of the proposed MemMod4CVQA framework on both real-world egocentric video data from the CASTLE dataset and a simulated lifelogging dataset.

5.1 Egocentric Dataset Evaluation

We utilize two hours of egocentric video data from the CASTLE dataset [27], focusing on its rich multi-context sequences of daily human activity. We performed a qualitative evaluation of the correctness, causal soundness, and explainability of generated answers to structured queries over segmented events and causal graphs.

We constructed a questions benchmark comprising four query types: (a) Causal (*"why"*), (b) Temporal (*"when"*, "what happened before/after"), (c) Descriptive (*"what"*), (d) Event-based (*"where"*). Each category included four curated queries with manually annotated ground truth answers.

To support this evaluation, we generated episodic-semantic memory representations and causal relationships from the video data using our MemMod4CVQA framework. The *episodic* and *semantic* memory modules were extracted and populated using scene and activity representations extracted from the videos.

Two annotators evaluated and assessed the output responses based on (a) correctness of answers in alignment with ground truth responses, (b) logical plausibility with known causal links, (c) clarity and relevance of the reasoning provided.

We evaluate the robustness of our framework under ablated memory conditions, where all possible combinations of episodic, semantic, and causal memory modules were used. Table 1 lists an example from each query-type for the different modes we tested.

Our evaluation reveals that relying on a single memory mode limits the scope of responses–*episodic memory* provides detailed, context-rich answers, *semantic memory* offers generalized patterns, and *causal reasoning* enables logical interpretation. In contrast, blended modes (e.g., causal-episodic, causal-semantic, or all combined) significantly enhance the depth, coherence, and explanatory quality of responses. Notably, the "**All**" mode, which integrates episodic, semantic, and causal information, provided the most semantically coherent, causally grounded, interpretable answers.

These findings support our hypothesis that a memory-augmented, causally grounded architecture can significantly improve insight generation from life logging data.

5.2 Simulated Dataset Evaluation

We also generate a synthetic non-egocentric life logging dataset that simulates comprehensive real-world human behavior and environmental interactions over time. The dataset simulates 1500 hours and covers five distinct modalities, each designed to reflect key aspects of everyday life.

1. Environmental modality: Temperature and weather conditions.
2. Contextual modality: Time of the week (weekend or a work hour) and location settings (e.g., at home, in the work place or in a social setting)
3. Visual modality: Scene brightness, visual complexity, color diversity and number of objects in the scene.
4. Behavioral modality: Physical activity level, social interaction frequency, digital device usage hours and daily step count.
5. Physiological modality: Heart rate and stress level, influenced by behavioral activity levels and contextual factors such as time of day or workload intensity.

We emulate the above conditions and modalities using standard behavioral patterns derived from prior literature [3, 15, 23, 37]. Simulated dataset exhibits the following causal relationships:

1. **Cross-Modal Dependencies:** Environmental and contextual factors (such as weather and location) influence behavioral patterns. In turn, these behavioral patterns affect physiological responses, which subsequently impact psychological or semantic outcomes. Visual characteristics of scenes are also shaped by both environmental and contextual factors.
2. **Temporal Dependencies:** The dataset models temporal dynamics, including lag effects in psychological states (e.g., mood persistence), carry-over effects in physiological states (such as

Table 2: Performance comparison of different causal discovery algorithms on the synthetic dataset. PC(Peter-Clark) uses conditional independence testing (constraint-based), HC(Hill-Climbing) uses graph search with BIC/AIC scoring (score-based), Enhanced versions include optimizations for mixed datatype support and cross modal discovery, and Correlation Baseline provides simple pairwise correlation reference.

Algorithm	F1 Score	Cross-Modal F1 Score	Precision	Recall	Predicted Edges
Correlation_Baseline	0.202	0.171	0.134	0.406	97
Standard_HC	0.227	0.239	0.179	0.312	56
Standard_PC	0.215	0.226	0.212	0.219	33
Our_Enhanced_HC_BIC	0.400	0.353	0.321	0.531	53
Our_Enhanced_HC_AIC	0.340	0.312	0.258	0.500	62
Our_Enhanced_PC	**0.448**	**0.458**	**0.500**	**0.406**	**26**

stress and fatigue), and behavioral momentum (such as sustained physical activity across time).

(3) **Realistic Constraints:** The simulation also incorporates realistic constraints such as circadian rhythm effects on activity and physiology, distinct behavior patterns between weekdays and weekends, activity changes based on weather conditions, and stress variations influenced by work schedules.

With this comprehensive set of scenarios, our synthetic dataset has 1500 timestamped observations collected over 62.5 days of hourly data with 26 distinct feature variables and 32 known causal relationships.

To evaluate the performance of predicting causal relationships, we use the standard evaluation metrics: Precision, Recall, and F1 score, applied to the synthetically generated dataset. Table 2 presents the performance comparison for different algorithms on the synthetic dataset. From the table we observe that MemMod4CVQA's conditional independence testing produces the highest F1 score of 44.8%. This is 24.6% higher than the correlational baseline which is performed on a pairwise correlational reference.

6 Discussion and Next Steps

Working with Egocentric life log videos We have currently performed initial evaluation of the MemMod4CVQA framework on a simulated dataset. Although the results are promising, substantial effort in all modules of MemMod4CVQA is necessary to improve the system's performance. Furthermore, the simulation currently does not consider egocentric images. As a next step, we will use the CASTLE dataset to detect causal relationship among events [27]. The dataset already poses several challenges which we believe can be addressed using the MemMod4CVQA framework.

Application and Use cases: Smartglasses have been of interest to the wearable sensing community for several years. Several smartglasses have been proposed over the years, with Meta's AI Glasses gaining traction recently [22]. These smartglasses currently are capable of capturing various egocentric life log videos. In future, they can benefit from MemMod4CVQA enabling causal visual question answering to answer questions like 'Why is Tom looking for something?', when Tom is searching for his keys. Furthermore, this framework opens up a lot of potential applications in the Human Robot Interaction space in a collaborative environment.

Optimization possibilities: One big challenge in the CVQA field is handling large amounts of data to process. The pre-processing steps that we used for MemMod4CVQA are an initial step towards reducing computational burdens. From a system point, the models described should run with minimal footprint to reduce energy consumption, a goal of MemMod4CVQA. In future, we will work towards detecting human attention so that cognitive models can be improved further.

Ethical challenges: Working with egocentric life logging videos pose serious privacy concerns. Indeed, researchers have noted various concerns with egocentric videos [1]. To mitigate these concerns, researchers have focused on obfuscating unnecessary information in human activity recognition tasks [28]. However, balancing between privacy and usability of information from image is a major challenge. We will devise approaches to balance privacy and usability concerns.

7 Concluding Remarks

In this work, we proposed MemMod4CVQA, a vision for causally grounded insight generation and question answering from lifelogging videos. By integrating event segmentation, dual memory modeling based on the Cohen-Conway framework, and causal graph extraction, our approach aims to bridge the gap between raw multimodal data and explainable, structured reasoning. We believe this work would open new pathways for integrating causal reasoning into ubiquitous sensing, and sets the stage for future systems that not observe but *understand* human behavior in context.

Acknowledgment

Sougata Sen's work is partially supported by a gift from the TensorFlow team to develop educational resources, and partially by Department of Science & Technology, Government of India's grant number DST/INT/USA/NSF-DST/Sougata/P-12/2024 (G). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of the funding organizations.

References

[1] Rawan Alharbi, Tammy Stump, Nilofar Vafaie, Angela Pfammatter, Bonnie Spring, and Nabil Alshurafa. 2018. I Can't Be Myself: Effects of Wearable Cameras on the Capture of Authentic Behavior in the Wild. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 3, Article 90 (Sept. 2018), 40 pages. https://doi.org/10.1145/3264900

[2] 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*. 1560–1568.

[3] Martina Cinquini, Fosca Giannotti, and Riccardo Guidotti. 2021. Boosting Synthetic Data Generation with Effective Nonlinear Causal Discovery. In *2021 IEEE Third International Conference on Cognitive Machine Intelligence (CogMI)*. IEEE, 54–63. https://doi.org/10.1109/cogmi52975.2021.00016

[4] Gillian Cohen and Martin A Conway. 2007. *Memory in the real world*. Psychology press.

[5] Martin A Conway and Christopher W Pleydell-Pearce. 2000. The construction of autobiographical memories in the self-memory system. *Psychological review* 107, 2 (2000), 261.

[6] Ana Garcia del Molino, Bappaditya Mandal, Vigneshwaran Subbaraju, and Vijay Chandrasekhar. [n. d.]. VC-I2R@ ImageCLEF2017: Ensemble of Deep Learned Features for Lifelog Video Summarization.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*. 4171–4186.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[9] Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. 2023. An empirical study of end-to-end video-language transformers with masked visual modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22898–22909.

[10] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847* (2016).

[11] David A Gold, Jeffrey M Zacks, and Shaney Flores. 2017. Effects of cues to event segmentation on subsequent memory. *Cognitive research: principles and implications* 2 (2017), 1–15.

[12] Gabriele Goletto, Tushar Nagarajan, Giuseppe Averta, and Dima Damen. 2024. Amego: Active memory from long egocentric videos. In *European Conference on Computer Vision*. Springer, 92–110.

[13] 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*. 18995–19012.

[14] New Fei Ho, Jordon Xin Jie Tng, Mingyuan Wang, Guoyang Chen, Vigneshwaran Subbaraju, Suhailah Shukor, Desiree Si Xian Ng, Bhing-Leet Tan, Shu Juan Puang, Sok-Hong Kho, et al. 2020. Plasticity of DNA methylation, functional brain connectivity and efficiency in cognitive remediation for schizophrenia. *Journal of Psychiatric Research* 126 (2020), 122–133.

[15] Jeremy Howick, Paul Kelly, and Michael Kelly. 2019. Establishing a causal link between social relationships and health using the Bradford Hill Guidelines. *SSM - Population Health* 8 (May 2019), 100402. https://doi.org/10.1016/j.ssmph.2019.100402 Erratum in: SSM Popul Health. 2020 Dec 10;12:100711. doi: 10.1016/j.ssmph.2020.100711.

[16] R Reed Hunt and James B Worthen. 2006. *Distinctiveness and memory*. Oxford University Press.

[17] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2017. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2758–2766.

[18] Dohwan Ko, Ji Soo Lee, Wooyoung Kang, Byungseok Roh, and Hyunwoo J Kim. 2023. Large language models are temporal and causal reasoners for video question answering. *arXiv preprint arXiv:2310.15747* (2023).

[19] Victor R Lee. 2014. What's happening in the" Quantified Self" movement? *ICLS 2014 proceedings* (2014), 1032.

[20] Jie Lin, Ana Garcia del Molino, Qianli Xu, Fen Fang, and Vigneshwaran Subbaraju. [n. d.]. VCI2R at the NTCIR-13 Lifelog-2 Lifelog Semantic Access Task.

[21] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[22] Meta. 2025. Meta AI-glasses. https://www.meta.com/ai-glasses/. Accessed:2025-7-7.

[23] Wei Ning, Jiahui Yin, Qiang Chen, and Xiaogang Sun. 2023. Effects of brief exposure to campus environment on students' physiological and psychological health. *Frontiers in Public Health* Volume 11 - 2023 (2023). https://doi.org/10.3389/fpubh.2023.1051864

[24] Alessandro Ortis, Giovanni M Farinella, Valeria D'Amico, Luca Addesso, Giovanni Torrisi, and Sebastiano Battiato. 2017. Organizing egocentric videos of daily living activities. *Pattern Recognition* 72 (2017), 207–218.

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PmLR, 8748–8763.

[26] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 779–788.

[27] Luca Rossetto, Werner Bailer, Duc-Tien Dang-Nguyen, Graham Healy, Björn Þór Jónsson, Onanong Kongmeesub, Hoang-Bao Le, Stevan Rudinac, Klaus Schöffmann, Florian Spiess, et al. 2025. The CASTLE 2024 Dataset: Advancing the Art of Multimodal Understanding. *arXiv preprint arXiv:2503.17116* (2025).

[28] Soroush Shahi, Rawan Alharbi, Yang Gao, Sougata Sen, Aggelos K Katsaggelos, Josiah Hester, and Nabil Alshurafa. 2022. Impacts of image obfuscation on fine-grained activity recognition in egocentric video. In *2022 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*. IEEE, 341–346.

[29] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

[30] Dulanga Weerakoon, Vigneshwaran Subbaraju, Nipuni Karumpulli, Tuan Tran, Qianli Xu, U-Xuan Tan, Joo Hwee Lim, and Archan Misra. 2020. Gesture enhanced comprehension of ambiguous human-to-robot instructions. In *Proceedings of the 2020 International Conference on Multimodal Interaction*. 251–259.

[31] Junbin Xiao, Angela Yao, Zhiyuan Liu, Yicong Li, Wei Ji, and Tat-Seng Chua. 2022. Video as conditional graph hierarchy for multi-granular question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 2804–2812.

[32] Qianli Xu, Fen Fang, Ana Molino, Vigneshwaran Subbaraju, and Joo-Hwee Lim. 2021. Predicting event memorability from contextual visual semantics. *Advances in Neural Information Processing Systems* 34 (2021), 22431–22442.

[33] Qianli Xu, Vigneshwaran Subbaraju, Chee How Cheong, Aijing Wang, Kathleen Kang, Munirah Bashir, Yanhong Dong, Liyuan Li, and Joo-Hwee Lim. 2018. Personalized serious games for cognitive intervention with lifelog visual analytics. In *Proceedings of the 26th ACM international conference on Multimedia*. 328–336.

[34] Qianli Xu, Jiayi Zhang, Joanes Grandjean, Cheston Tan, Vigneshwaran Subbaraju, Liyuan Li, Kuan Jin Lee, Po-Jang Hsieh, and Joo-Hwee Lim. 2020. Neural correlates of retrieval-based enhancement of autobiographical memory in older adults. *Scientific Reports* 10, 1 (2020), 1447.

[35] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. 2021. Just ask: Learning to answer questions from millions of narrated videos. In *Proceedings of the IEEE/CVF international conference on computer vision*. 1686–1697.

[36] Chuanqi Zang, Hanqing Wang, Mingtao Pei, and Wei Liang. 2023. Discovering the Real Association: Multimodal Causal Reasoning in Video Question Answering. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 19027–19036. https://doi.org/10.1109/CVPR52729.2023.01824

[37] Radoslava Švihrová, Alvise Dei Rossi, Davide Marzorati, Athina Tzovara, and Francesca Dalia Faraci. 2025. Designing digital health interventions with causal inference and multi-armed bandits: a review. *Frontiers in Digital Health* Volume 7 - 2025 (2025). https://doi.org/10.3389/fdgth.2025.1435917