

EgoTrigger: Toward Audio-Driven Image Capture for Human Memory Enhancement in All-Day Energy-Efficient Smart Glasses

Akshay Paruchuri , Sinan Hersek , Lavisha Aggarwal , Qiao Yang , Xin Liu , Achin Kulshrestha , Andrea Colaco , Henry Fuchs , and Ishan Chatterjee 

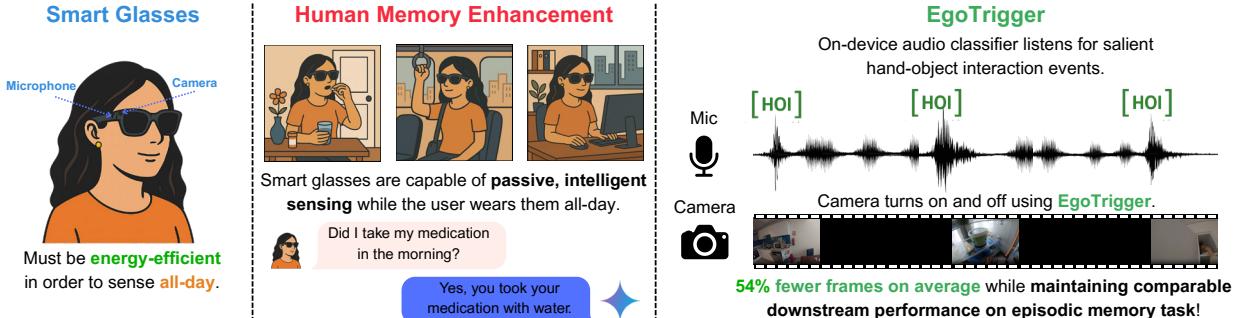


Fig. 1: We present *EgoTrigger*, an image capture approach designed for resource-constrained smart eyeglass systems. *EgoTrigger* uses a lightweight audio classification model and a custom classification head to trigger image capture from hand-object interaction (HOI) audio cues, such as the sound of a medication bottle being opened. Subsequently, image capture can be triggered off either based on a fixed duration (e.g., 1 second) or a hysteresis-based approach. *EgoTrigger* is capable of using 54% fewer frames on average while maintaining or exceeding performance on an episodic memory task, signifying potential for energy-efficient usage as a part of all-day smart glasses capable of human memory enhancement.

Abstract—All-day smart glasses are likely to emerge as platforms capable of continuous contextual sensing, uniquely positioning them for unprecedented assistance in our daily lives. Integrating the multi-modal AI agents required for human memory enhancement while performing continuous sensing, however, presents a major energy efficiency challenge for all-day usage. Achieving this balance requires intelligent, context-aware sensor management. Our approach, *EgoTrigger*, leverages audio cues from the microphone to selectively activate power-intensive cameras, enabling efficient sensing while preserving substantial utility for human memory enhancement. *EgoTrigger* uses a lightweight audio model (YAMNet) and a custom classification head to trigger image capture from hand-object interaction (HOI) audio cues, such as the sound of a drawer opening or a medication bottle being opened. In addition to evaluating on the QA-Ego4D dataset, we introduce and evaluate on the Human Memory Enhancement Question-Answer (HME-QA) dataset. Our dataset contains 340 human-annotated first-person QA pairs from full-length Ego4D videos that were curated to ensure that they contained audio, focusing on HOI moments critical for contextual understanding and memory. Our results show *EgoTrigger* can use 54% fewer frames on average, significantly saving energy in both power-hungry sensing components (e.g., cameras) and downstream operations (e.g., wireless transmission), while achieving comparable performance on datasets for an episodic memory task. We believe this context-aware triggering strategy represents a promising direction for enabling energy-efficient, functional smart glasses capable of all-day use — supporting applications like helping users recall where they placed their keys or information about their routine activities (e.g., taking medications).

Index Terms—Smart glasses, Microphones, Cameras, Multimodal sensing, Artificial intelligence, Energy efficient computing

1 INTRODUCTION

As an all-day wearable, smart glasses provide a unique platform to enable continuous contextual sensing from an egocentric perspective. This long-context sensing can be processed by powerful agentic multi-modal foundation models (MLLMs) to enable unprecedented assistance in our daily lives. For example, researchers have long envisioned enhancing users' memory of their daily actions via wearables and glasses [13, 23, 43]. Users may eventually be able to ask in natural language, "Where did I leave my keys earlier?" or "Did I turn the oven

off before leaving?", among other things.

However, realizing this vision of memory enhancement hinges on overcoming significant energy efficiency hurdles. As face-worn devices, smart glasses face strict form-factor limitations, constraining battery size, thermal dissipation, and onboard computation, and making all-day operation challenging [6]. Even the most optimized, highly capable MLLMs (e.g., Gemini Nano [48]) contain over a billion parameters, making local deployment intractable on smart glasses hardware. Rather, these MLLMs run on powerful processors in the cloud [1, 2, 48] with the glasses passing their sensor data over wireless link by way of a connected phone (e.g. Vuzix M400 [11], Ray-Ban Meta [42]) or separate compute puck (e.g., Meta's Orion Prototype [36]).

Despite offloading inference, a critical energy bottleneck for memory enhancement remains: continuous environmental sensing and wireless transmission to the agent, particularly from power-hungry and high-bandwidth cameras required for visual understanding [28]. To enable passive, intelligent sensing for applications like memory enhancement without constant battery drain, smarter sensing strategies are required.

We hypothesize that inexpensive sensing modalities (e.g., audio from

• Akshay Paruchuri and Henry Fuchs are with UNC Chapel Hill. Contact: akshay,fuchs@cs.unc.edu

• Ishan Chatterjee and the remaining co-authors are with Google. Contact: ishanc@google.com

Received 11 April 2025; revised 18 June 2025; accepted 22 July 2025. Date of publication 7 October 2025; date of current version 3 November 2025. This article has supplementary downloadable material available at <https://doi.org/10.1109/TVCG.2025.3616866> provided by the authors. Digital Object Identifier no. 10.1109/TVCG.2025.3616866

a microphone) will be readily available in future smart glasses and can be leveraged to effectively trigger capture of imagery from the camera. Prior research underscores the significance of hand-object interactions (HOIs) in shaping our memories, emphasizing that bodily engagement, particularly involving our hands, plays a crucial role in how we encode, recall, and structure key life experiences [24]. Audio cues, such as those indicating an HOI, can therefore serve as reliable indicators for moments when capturing visual information is most pertinent.

Building on this insight, we introduce *EgoTrigger*, an energy-efficient approach for episodic memory enhancement using audio-gated visual capture. *EgoTrigger* employs a lightweight audio classification model to detect salient audio events (e.g., HOI sounds like opening a medication bottle) and trigger brief camera captures. This selective capture strategy significantly reduces the need for continuous camera operation, maintaining performance on challenging human memory enhancement baselines. Our evaluations show *EgoTrigger* is capable of using 54% fewer frames on average while achieving comparable performance on downstream episodic memory question-answering tasks compared to continuous capture and fixed decimation baselines.

The primary contributions of this work are:

- We present *EgoTrigger*, a novel framework that uses audio cues from hand-object interactions (HOIs) to intelligently gate power-intensive cameras, such as those that would be a part of smart glasses. We validate its practical viability through an analysis of HOI audio event detection, its robustness to environmental noise, and a real-world assessment of its false positive rate on continuous video streams.
- We demonstrate that this audio-driven strategy reduces required visual frames by 54% on average while maintaining downstream episodic memory task performance within 2% of continuous capture, achieving a near Pareto-optimal balance between utility and energy savings.
- We introduce and release the Human Memory Enhancement Question-Answer (HME-QA) dataset, the first multimodal egocentric QA dataset curated for HOI-centric scenarios with guaranteed, paired audio, to facilitate future research.

Our code and access to the HME-QA dataset annotations can be found on our project page: <https://egotigger.github.io/>

2 RELATED WORK

Energy-Efficient Sensing for Wearables. Wearable, network connected devices have become increasingly prevalent in our daily lives and typically feature a battery that needs to be recharged [38]. Such devices, which most commonly include smart watches today, are expected to last all day akin to our smartphones [28] in order to be usable. Prior works on improving energy-efficiency in wearables include device-specific optimization [26], energy harvesting [4], and adaptive device management frameworks [30, 39, 46]. Frameworks like AdaSense [39] and SmartAPM [30] have demonstrated significant energy savings in wearables by dynamically adjusting sensor sampling rates and leveraging deep reinforcement learning for power management, achieving up to 51% and 36% energy reductions respectively while preserving user experience. Approaches specific to device capabilities such as imaging have also been explored - for example, event cameras offer compelling advantages such as low latency, high dynamic range, and reduced power consumption, but are currently hindered by challenges including high costs, limited availability, and the need for specialized algorithms [7].

Smart glasses in particular face significant energy constraints due to their compact form factors, limited battery capacities, and the continuous operation of power-intensive sensors like cameras [6]. These constraints continue to be limiting for present day devices, such as the Vuzix M400 [11] and Ray-Ban Meta [42], typically requiring some kind of "split compute" system across a separate phone or puck in order to enable meaningful user experiences with resource-constrained smart glasses. Researchers have optimized energy use in XR devices from the perspective that such devices will feature power-hungry displays that subsequently can be optimized using perceptually guided

techniques such as adaptive shading, adaptive frame rates, luminance clipping, and color foveation [9, 25]. A co-optimization framework that balances energy consumption, latency, and accuracy in specific tasks such as scene reconstruction [49] has also been proposed and shows promise for holistic system design in energy-constrained XR applications. Lastly, approaches such as the intent-driven arbitration framework by Gonzalez et al. [19] leverage pose and sensor data to enable energy-efficient XR interactions by seamlessly switching between devices—such as smartphones, tablets, and smartwatches—based on user gaze and proximity.

In contrast to prior works, *EgoTrigger* targets the usage of a less expensive, common sensing modality, audio from a microphone, in order to effectively gate image capture via pertinent audio cues (e.g., hand-object interaction audio cues). This enables energy savings specifically for vision-centric smart glasses capable of passive, intelligent sensing and human memory enhancement, without requiring continuous multimodal sensing or sacrificing too much downstream task performance.

Audio-based Context Recognition. Audio-based context recognition has become a pivotal component in wearable and egocentric computing, enabling devices to infer user activities and environmental cues with minimal energy consumption. Foundational tools like openSMILE [16], AudioSet [18], and YAMNet [22] have been instrumental in classifying a wide range of audio events, from environmental sounds to speech and emotion detection. Models inspired by such tools are particularly valuable for resource-constrained devices due to their lightweight architectures and real-time processing capabilities.

Recent advancements have focused on integrating audio with visual data to enhance context recognition, especially in egocentric settings. For instance, EPIC-Fusion [27] employs mid-level fusion of audio, RGB, and optical flow modalities to improve action recognition in first-person videos. Listen to Look [17], a framework that leverages audio as a preview mechanism to reduce visual redundancy in action recognition tasks, enables efficient recognition by distilling video features from a single frame and its accompanying audio. Furthermore, foundation models like Qwen-Audio [10] have demonstrated strong performance across tasks such as audio classification, speech recognition, and emotion recognition, indicating their potential for robust audio understanding in diverse applications.

Prior works underscore the importance of audio as a low-power, informative modality for context recognition in wearable devices. By leveraging audio cues, systems like our proposed *EgoTrigger* can efficiently manage sensor activation, reducing energy consumption while maintaining performance in applications such as human memory enhancement.

Human Memory Enhancement from Egocentric Data. Research on egocentric data, which most commonly involves visual data, has increasingly focused on memory augmentation, leveraging wearable devices to capture and retrieve personal experiences. Datasets like Ego4D [20] and an extension of its episodic memory benchmark QA-Ego4D [5] provide extensive egocentric video data paired with question-answer annotations, supporting models that can answer queries such as "Where did I leave my keys?". Building on this, Encode-Store-Retrieve [45] introduces a system that encodes egocentric videos into language representations stored in a vector database, enabling efficient natural language querying and outperforming traditional models in episodic recall tasks.

Computationally efficient approaches that leverage less expensive sensor modalities, such as IMU in the case of EgoDistill [47] and audio in the case of Listen to Look [17], aim to reduce visual redundancy and enhance action recognition in egocentric videos. Additionally, emerging datasets such as EgoSchema [35], EgoTempo [37], and Ego-Life [50] aim to provide richer contextual information, including temporal dynamics and daily activity patterns, to support the development of more robust memory augmentation systems. These advancements highlight the potential of integrating multimodal egocentric data to create personalized, efficient, and context-aware memory aids.

In this paper, our proposed system, *EgoTrigger*, advances the vision of all-day human memory enhancement by using lightweight audio

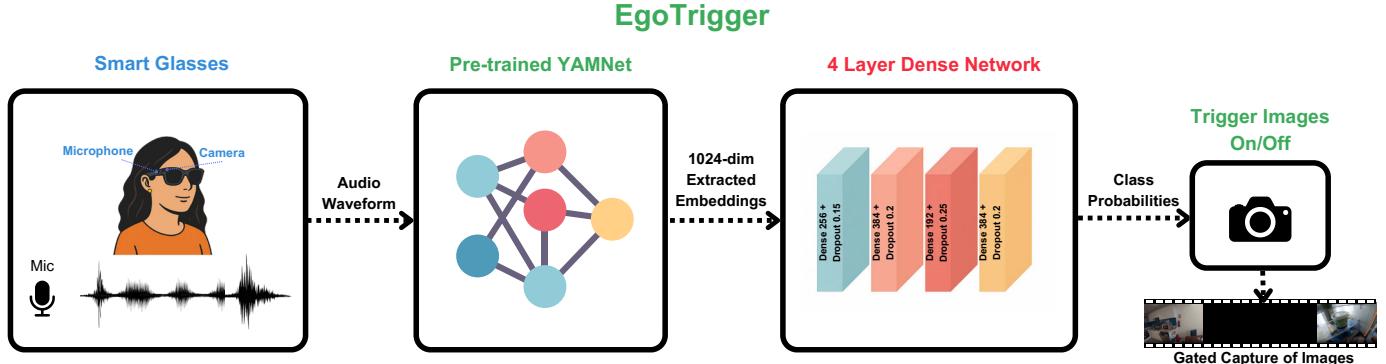


Fig. 2: **EgoTrigger**. Our approach focuses on audio from smart glasses that are processed by a pre-trained YAMNet model to extract 1024-dimensional embeddings, which are passed through a 4-layer dense network that serves as our custom, binary classification head. The resulting class probabilities for C_0 (indicates no hand-object interaction, or HOI) and C_1 (HOI) are used to trigger image capture, enabling gated visual sensing based on salient HOI audio events.

classification capable of on-device usage to selectively trigger image capture during hand-object interaction moments, which in turn serve as key cues for episodic memory and subsequently reduce the need for continuous visual sensing while maintaining downstream performance for human memory enhancement.

3 OUR APPROACH

3.1 EgoTrigger

We hypothesize that inexpensive sensing modalities (e.g., audio from a microphone) will be readily available in future smart glasses and can be leveraged to effectively trigger capture of imagery from the camera. Prior research underscores the significance of hand-object interactions (HOIs) in shaping our memories, emphasizing that bodily engagement, particularly involving our hands, plays a crucial role in how we encode, recall, and structure key life experiences [24]. Our approach centers on utilizing audio cues characteristic of Hand-Object Interactions (HOI) as a trigger mechanism for egocentric data capture. The core component is a binary audio classifier, f_θ , parameterized by θ , trained to differentiate between audio segments containing HOI events (C_1) and those without HOI events (C_0). This classifier forms the foundation of our *EgoTrigger* system. *EgoTrigger* processes a continuous audio stream X by applying f_θ within a sliding window framework. Specifically, it analyzes short, overlapping audio segments x_i extracted from X at time intervals corresponding to the window hop. Let w_d be the window duration (e.g., 4 seconds) and w_h be the hop size (e.g., 2 seconds).

$$x_i = X[i \cdot w_h : i \cdot w_h + w_d]$$

The classifier outputs the posterior probability:

$$P(C_1 | x_i) = f_\theta(x_i)$$

EgoTrigger activates a designated function when this probability exceeds a predefined threshold τ :

$$P(C_1 | x_i) \geq \tau$$

This provides contextually relevant activation based on audible user interactions. *EgoTrigger* then translates the classifier's output into practical event triggers. When $P(C_1 | x_i) \geq \tau$, the system enters an active state. Two primary control strategies manage this state. The first employs a fixed 'OFF' duration, T_{fixed} (e.g., 1 second). If the trigger activates at the start time $t_{start} = i \cdot w_h$ of window i , the system remains active for the interval:

$$[t_{start}, t_{start} + T_{fixed}]$$

The second strategy uses hysteresis with distinct 'ON' and 'OFF' thresholds, τ_{on} and τ_{off} respectively, where $\tau_{on} > \tau_{off}$ (e.g., $\tau_{on} = 0.8$, $\tau_{off} = 0.7$). The system state $S_t \in \{\text{ON}, \text{OFF}\}$ at time t is

updated based on the classifier's output using the following hysteresis rule:

$$S_t = \begin{cases} \text{ON}, & \text{if } S_{t-1} = \text{OFF} \text{ and } P(C_1 | x_t) \geq \tau_{on} \\ \text{OFF}, & \text{if } S_{t-1} = \text{ON} \text{ and } P(C_1 | x_t) < \tau_{off} \\ S_{t-1}, & \text{otherwise} \end{cases}$$

This prevents rapid state changes due to minor probability fluctuations. Both strategies output time intervals:

$$[t_{start}, t_{stop}]$$

defining periods of system activation for reference.

As a part of our evaluation in Section 6.1, we evaluate *EgoTrigger*'s ability to differentiate between audio segments containing HOI events (C_1) and those without HOI events (C_0), additionally taking into account weighted results due to the inherent class imbalances in most egocentric datasets involving audio. We also analyze trigger robustness and false positives, including provision of a false positive rate. In Section 6.2, we evaluate on the episodic memory task using both the fixed 'OFF' duration, T_{fixed} , strategy and the hysteresis with distinct 'ON' and 'OFF' thresholds strategy in contrast to continuous capture and fixed decimation baselines.

3.2 Implementation

The classifier f_θ was implemented using TensorFlow [14], leveraging transfer learning from a pretrained YAMNet [22] model. The final classification layer of YAMNet was replaced with a custom four-layer dense network, fine-tuned for binary hand-object interaction (HOI) detection (C_1 vs C_0). The classifier head consists of a sequence of fully connected layers with ReLU activations and dropout for regularization: $1024 \rightarrow 256 \rightarrow 384 \rightarrow 192 \rightarrow 384 \rightarrow 2$, with dropout rates of 0.15, 0.2, 0.25, and 0.2 respectively after each hidden layer.

At training time, input audio segments are pre-processed via a standardized pipeline that includes YAMNet as a base model to extract embeddings. This process includes conversion to mono, resampling to a target sampling rate $F_s = 16$ kHz, and amplitude normalization to the range $[-1.0, 1.0]$. The model outputs logits $z = [z_0, z_1]$, which are converted to class probabilities using the softmax function:

$$P(C_k | x_i) = \frac{e^{z_k}}{e^{z_0} + e^{z_1}}, \quad k \in \{0, 1\}$$

A decision threshold τ (e.g., $\tau = 0.4$) is then applied to $P(C_1 | x_i)$ to determine whether to activate the image capture trigger. We train *EgoTrigger* using the binary cross-entropy loss:

$$\mathcal{L} = -[y \log f_\theta(x) + (1-y) \log(1 - f_\theta(x))]$$

where $y \in \{0, 1\}$ is the ground-truth label indicating the presence (C_1) or absence (C_0) of a hand-object interaction, and $f_\theta(x)$ is the predicted probability of C_1 .

In order to train *EgoTrigger*, we used a subset of the Ego4DSounds dataset from [8]. Ego4DSounds is derived from Ego4D [20], an existing large-scale egocentric video dataset, and contains video clips spanning hundreds of different scenes and actions. Clips have a high action-audio correspondence and contain time-stamped narrations indicating actions performed by the camera-wearer, making them an excellent candidate to build training and test data for our particular classification task. We randomly sampled approximately 14,000 audio waveform files for training and approximately 2,300 samples for testing. Using Gemini 1.5 Pro, we additionally ran an audio understanding pass on all samples in order to annotate them with a label, based on both the audio and corresponding narration, indicating whether or not the audio sample corresponded to a hand-object interaction (HOI). We include further details regarding this process, including the prompt we utilized, in our supplementary materials. To address class imbalance in the Ego4DSounds dataset—where hand-object interaction (HOI) events can dominate—we applied a class weighting strategy. Class weights were computed from training set frequencies, resulting in a weight of $w_0 = 5.6577$ for the negative class (C_0) and $w_1 = 0.5485$ for the positive class (C_1). These weights were incorporated directly into the loss function.

For deployment, particularly targeting resource-constrained edge devices, the TensorFlow model was converted to the TensorFlow Lite (TFLite) format [12], more recently known as LiteRT. Default optimizations were applied to enhance efficiency. This resulted in a model size reduction of 86.23%, bringing the model size from 23.60MB to 3.77MB. Additionally, it should be noted that YAMNet-based models and the LiteRT optimized variants are highly capable of being deployed on-device, with prior task benchmarks indicating an average latency on a mobile device using CPU / GPU being 12.29 milliseconds [15]. This can be further optimized by the usage of known techniques, such as streaming convolutions [44] or quantization [32] that enables usage on specific hardware components such as the "low power island" of the AR1 Gen1 [41].

4 HME-QA

4.1 Motivation

Recent advancements in egocentric datasets, notably Ego4D [20], have catalyzed the development of task-specific extensions aimed at enhancing various aspects of video understanding. For instance, EgoSchema [35] focuses on long-form video question answering, requiring models to comprehend extended temporal contexts. EgoTempo [37] emphasizes temporal reasoning capabilities, challenging models to integrate information across entire videos. Additionally, datasets like QA-Ego4D [5] extend Ego4D's tasks by pairing egocentric videos with natural language questions and answers, facilitating research in natural language query (NLQ) understanding.

Despite the richness of these datasets, a significant limitation persists: the lack of comprehensive audio data and corresponding QA pairs. Audio cues can be crucial for tasks involving hand-object interactions (HOI), which are central to our approach. An analysis of the QA-Ego4D dataset reveals that nearly 40% of its test split videos lack audio tracks, rendering them unusable for audio-centric research. This gap underscores the necessity for a curated dataset that ensures the presence of meaningful audio, thereby enabling more robust exploration of multimodal egocentric data.

4.2 Dataset

To address this limitation, we curated the HME-QA dataset, focusing on videos that encompass both high-quality audio and IMU modalities within the episodic memory benchmark of Ego4D. Our curation process began by filtering the Ego4D dataset to identify 82 full-length videos containing complete narrations and moments annotations. We then employed FFmpeg to verify the presence of actual audio tracks, resulting in a refined set of 50 valid videos. For these 50 videos, we utilized Gemini 2.0 Flash to generate initial question-answer pairs, emphasizing semantic relevance to the camera wearer's interactions. These AI-generated pairs underwent meticulous review by an expert annotator specializing in egocentric video datasets. The annotator discarded

low-quality pairs and retained those deemed high-quality, culminating in a collection of 340 question-answer pairs.

The HME-QA dataset comprises 50 full-length Ego4D videos, on average 12 minutes in duration, with some extending up to 30 minutes. This dataset can serve as a valuable resource for advancing research in human memory enhancement through multimodal egocentric data, with a focus on ensuring high quality audio data is available in videos utilized for evaluation.



Fig. 3: **Sample QA Pairs from Datasets.** Example question-answer pairs from QA-Ego4D (top) and HME-QA (bottom), each paired with the relevant image sequence from which the answer is derived.

5 EXPERIMENTAL SETUP

We trained the EgoTrigger audio classifier using a batch size of 64, categorical cross-entropy loss, and the AdamW [31, 33] optimizer with a learning rate $\eta = 3 \times 10^{-3}$ and weight decay $\lambda = 0.01$. Training was conducted for 50 epochs. Model performance on the Ego4DSounds test set was evaluated using standard classification metrics: precision (Pr), recall (Re), and F1-score, defined as $F_1 = \frac{2 \cdot \text{Pr} \cdot \text{Re}}{\text{Pr} + \text{Re}}$. We analyzed performance trade-offs across various decision thresholds τ through Figure 4.

For evaluation, we performed the same filtering of videos for audio done in Section 4 for the QA-Ego4D dataset - taking us from an initial test set comprised of 1854 QA pairs across 166 unique videos to a final test set comprised of 1071 QA pairs across 104 unique videos. We paired these processed videos with their corresponding question-answer pairs and used a large language model (LLM) to score the correctness of predicted answers. Following prior work [29, 34, 37, 40], we adopted an LLM-as-a-judge evaluation protocol. Specifically, we employed Gemini models with a fixed seed of 1337 and a temperature of 0.2 to ensure consistent results. For generating answers from processed videos and questions, we utilized Gemini 1.5 Pro, while for judging we utilized Gemini 1.5 Pro in a separate pass. To isolate attention and ensure precise evaluation, each prompt contained a single question paired with its corresponding video segment. Further details, including the exact prompts used, are provided in the supplementary materials.

6 EVALUATION

6.1 Performance Evaluation of EgoTrigger

We evaluate *EgoTrigger*'s ability to detect hand-object interaction (HOI) events by comparing several strategies to handle the imbalanced Ego4DSounds test set, with detailed results shown in Tab. 1. Our baseline approach using class weights achieves strong performance on the positive majority class (C_1), attaining an F1-score of 0.91. However, performance on the negative minority class (C_0) is poor, with an F1-score of only 0.15. To address this, we implemented an oversampling strategy using SMOTE. This approach dramatically improves performance for the minority class, boosting the F1-score for C_0 from 0.15 to

0.54, while maintaining excellent performance on C_1 (0.90 F1-score). This results in a superior overall model, increasing the weighted average F1-score from 0.83 to 0.86. For completeness, we also evaluated random undersampling, which improved the C_0 F1-score to 0.47 but degraded the C_1 F1-score significantly, lowering the overall weighted average. Given these results, we will adopt the model trained with SMOTE for our subsequent experiments.

Table 1: HOI Classification Performance. Evaluation of binary hand-object interaction (HOI) classification on the test set using three different training strategies. The support numbers remain constant as resampling is only applied to training data.

Strategy	Class	Prec.	Recall	F1	Supp.
Class Weights (Original)	C_0 (No HOI)	0.17	0.14	0.15	251
	C_1 (HOI)	0.90	0.92	0.91	2095
	Weighted Avg	0.82	0.84	0.83	2346
Oversampling (SMOTE)	C_0 (No HOI)	0.46	0.65	0.54	251
	C_1 (HOI)	0.93	0.87	0.90	2095
	Weighted Avg	0.88	0.85	0.86	2346
Undersampling (Random)	C_0 (No HOI)	0.35	0.70	0.47	251
	C_1 (HOI)	0.94	0.74	0.83	2095
	Weighted Avg	0.88	0.73	0.79	2346

To analyze the effect of the decision threshold τ , we report precision, recall, and F1-score for class C_1 across a range of thresholds in [Fig. 4](#). Results indicate that *EgoTrigger* maintains robust F1-scores (≥ 0.9) for a wide range of thresholds ($\tau \leq 0.5$), highlighting the reliability of the classifier across deployment settings. As τ increases, the model favors precision over recall, enabling applications to trade off between false positives and false negatives depending on the energy constraints and tolerance for missed events.

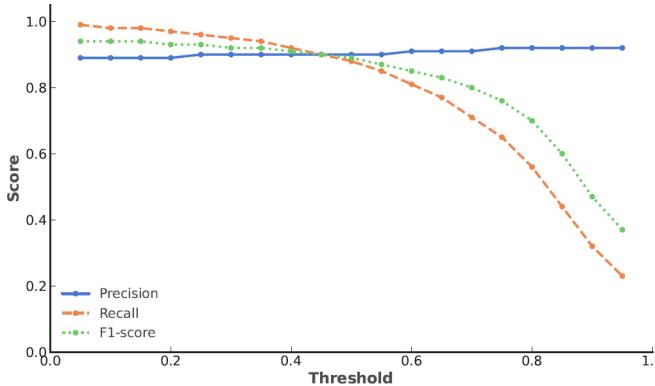


Fig. 4: C_1 Metrics vs. Threshold. Precision, recall, and F1-score for class C_1 as a function of decision threshold. Precision remains high across thresholds, while recall decreases more rapidly. F1-score peaks near the region where precision and recall are balanced, guiding threshold selection for optimal tradeoff.

To further assess the robustness and real-world viability of our HOI trigger, we conducted a series of experiments to quantify its performance against non-HOI sounds, additive noise, and in continuous, uncurated video streams. The results are summarized in [Tab. 2](#).

First, we evaluated the trigger’s resilience to noise by adding white noise to our Ego4DSounds test set at various Signal-to-Noise Ratios (SNRs). While performance degrades in noisier conditions, the model maintains a strong weighted F1-score of 0.78 even at a higher noise level ($\sigma^2 = 0.1$).

Second, we tested the trigger’s resilience to noise by adding white noise to our Ego4DSounds test set at various Signal-to-Noise Ratios (SNRs). While performance degrades in noisier conditions, the model maintains a strong weighted F1-score of 0.78 even at a higher noise level ($\sigma^2 = 0.1$).

Finally, to evaluate real-world performance, we tested the system on five full-length videos from our HME-QA dataset, manually identifying false triggers unrelated to human-object interactions. The system generated an average of 2.4 false positives per minute (FPPM). This rate allows the system to remain in standby mode most of the time while avoiding continuous recording, making it viable for practical deployment despite remaining challenges with noise robustness.

Table 2: Trigger Robustness and False Positive Analysis. False Positive Rate (FPR) is shown on non-HOI datasets, alongside F1-score degradation with additive noise and the in-the-wild False Positives Per Minute (FPPM) rate.

Experiment / Condition	Value
Specificity (FPR)	
ESC-50 Environmental Sounds	7.1%
Ego4DSounds (Non-HOI Test Clips)	2.3%
Robustness to Noise (F1-score)	
Clean ($\sigma^2=0$)	0.86
Low Noise ($\sigma^2 = 0.01$)	0.78
High Noise ($\sigma^2 = 0.1$)	0.59
In-the-Wild Perf. (FPPM)	
5 HME-QA Videos (avg. 12 min)	2.4

6.2 Performance on Egocentric QA Datasets

To simulate *EgoTrigger*-based image capture on input video, we first processed the input audio waveform using a sliding window approach with a window duration $w_d = 4$ seconds and a hop size $w_h = 2$ seconds. *EgoTrigger* includes an offline operation mode that processes pre-recorded video by simulating trigger behavior using predicted timestamps for when image capture would be activated or suppressed. Using the predicted activation intervals, we generate timestamped JSON files for each video. These are then used to modify the original videos via FFmpeg by blacking out all frames that would not have been captured by *EgoTrigger*.

We evaluated four approaches—continuous full capture (**Full**), naive decimation (**Decimated**), *EgoTrigger* with fixed-duration triggering (**ET-1s**), and *EgoTrigger* with hysteresis-based control (**ET-Hyst**)—on both the HME-QA and QA-Ego4D datasets. Naive decimation—uniformly dropping frames at fixed intervals—can reduce bitrate and energy consumption but often sacrifices temporal coherence, leading to degraded performance on downstream video understanding tasks [21]. Our naive decimation approach involves processing a video such that frames are only captured once every five seconds, effectively operating at 0.2 FPS. On HME-QA, the Full baseline achieved the highest accuracy at 77.3%. Both *EgoTrigger* variants maintained more comparable performance, with ET-1s and ET-Hyst. achieving 75.7% and 74.7% accuracy, respectively. In contrast, decimation dropped performance to 71.9%. On QA-Ego4D, the Full approach reached 41.08% accuracy, followed closely by ET-1s (40.1%) and ET-Hyst. (39.2%), while decimation again underperformed at 34.6%. These results indicate that *EgoTrigger* can significantly reduce visual data while preserving QA performance close to the full capture baseline. Qualitative examples are shown in [Figure 5](#).

6.3 Performance versus Energy Efficiency

Tables 4 and 5 present a comparative analysis of four image capture strategies across two egocentric QA datasets: HME-QA and QA-Ego4D. These include continuous full capture (**Full**), uniform frame decimation (**Decimated**), *EgoTrigger* with a fixed 1-second OFF duration (**ET-1s**), and *EgoTrigger* with a hysteresis-based OFF trigger

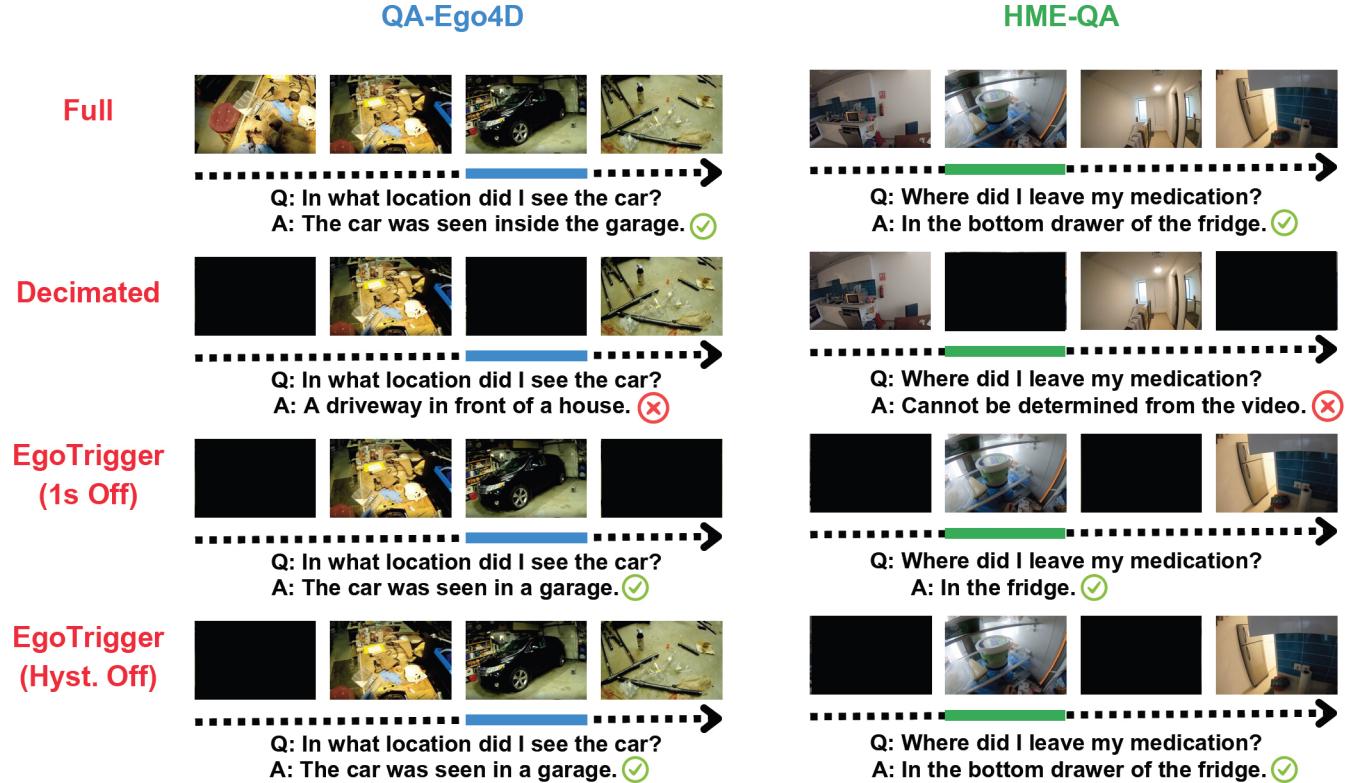


Fig. 5: **EgoTrigger QA Results.** Sample question-answer outcomes on QA-Ego4D (left) and HME-QA (right) using different video sampling strategies. EgoTrigger variants preserve key visual content necessary for correct answers while significantly reducing the number of frames, compared to full or naively decimated baselines.

(ET-Hyst.). Figure 6 contains a visualization of the accuracy versus frame reduction trade-off.

Table 3: **Accuracy (%) across Datasets.** Comparison of correctness (rating ≥ 3) scores across approaches on HME-QA ($N = 340$) and QA-Ego4D ($N = 1071$). The best result is in bold, while the second best result is underlined.

Approach	HME-QA	QA-Ego4D
Full	77.3	41.08
Decimated	71.9	34.6
ET-1s	<u>75.7</u>	<u>40.1</u>
ET-Hyst.	74.7	39.2

Across both datasets, EgoTrigger variants consistently reduce frame usage by a substantial margin while maintaining QA accuracy close to that of full capture, in contrast to the decimated baseline. For HME-QA, ET-1s achieves a 54.39% reduction in frames and only a 0.98 percentage point drop in accuracy compared to Full. Similarly, on QA-Ego4D, ET-1s reduces frame count by 54.28% with only a 0.98 percentage point drop in accuracy. In contrast, decimation saves the most frames (79.5%) but at a greater accuracy cost, especially on QA-Ego4D where it drops by nearly 6.48 percentage points.

These reductions can translate directly to bitrate savings. ET-1s approximately halves the bitrate on both datasets (HME-QA: 2.50 Mbps vs. 5.47 Mbps Full; QA-Ego4D: 0.60 Mbps vs. 1.31 Mbps Full). We inferred these values using black frame durations calculated via FFmpeg and video file sizes. Black frames replace periods where the trigger was off, and thus serve as a reliable estimate of when camera capture can be off and effective data transmission avoided.

7 DISCUSSION

Our evaluation demonstrates the potential of *EgoTrigger* as an energy-efficient strategy for capturing visually relevant moments for memory

Table 4: **Performance and Efficiency on HME-QA.** Comparison of accuracy, frame reduction, and estimated bitrate across baselines and EgoTrigger variants on the HME-QA dataset.

Approach	Correct (%)	Frames Reduced (%)	Est. Bitrate (Mbps)
Full	77.3	–	5.47
Decimated	71.9	79.48	1.13
ET-1s	75.7	54.39	2.50
ET-Hyst.	74.7	27.88	3.96

Table 5: **Performance and Efficiency on QA-Ego4D.** Comparison of accuracy, frame reduction, and estimated bitrate across baselines and EgoTrigger variants on the QA-Ego4D dataset.

Approach	Correct (%)	Frames Reduced (%)	Est. Bitrate (Mbps)
Full	41.08	–	1.31
Decimated	34.6	79.41	0.27
ET-1s	40.1	54.28	0.60
ET-Hyst.	39.2	23.22	1.01

enhancement, triggered by audio cues. We now discuss the implications of our findings, framed around the specific evaluation areas, followed by limitations and future directions.

7.1 Performance Evaluation of EgoTrigger

The core technical feasibility of our approach hinges on reliably detecting Hand-Object Interactions (HOIs) from audio using a lightweight model suitable for on-device deployment. Our evaluation (Section 6.1) confirmed that the fine-tuned YAMNet-based classifier can effectively identify characteristic HOI sounds within egocentric audio streams, even amidst significant class imbalance (Tab. 1). This validates the fundamental premise that salient audio cues can serve as effective triggers for visual capture. Furthermore, the classifier’s robustness across

various decision thresholds (Fig. 4) suggests practical deployability, allowing system designers to tune the trade-off between trigger sensitivity (recall) and specificity (precision) based on application needs and energy constraints. While classification performance on non-HOI sounds (C_0) was lower, the downstream task results indicate that the positive HOI signal is sufficiently discriminative for the intended memory enhancement application.

7.2 Performance on Egocentric QA Datasets

The key utility of *EgoTrigger* lies in its ability to support downstream applications while reducing resource consumption. Our evaluation on the HME-QA and QA-Ego4D datasets (Section 6.2) reveals a crucial insight: selectively capturing visual frames around detected HOIs can preserve the essential information required to answer challenging episodic memory questions. Both *EgoTrigger* variants maintained question-answering correctness remarkably close to that achieved using the full, continuous video stream (Section 6.3). This outcome contrasts sharply with naive temporal decimation, which, despite potentially higher frame reduction, incurred a more significant performance penalty. This suggests that the context-aware nature of *EgoTrigger*—focusing capture on moments of interaction—is substantially more effective at retaining task-relevant visual details than uniform subsampling. These findings support the hypothesis that HOIs often coincide with moments of high visual salience critical for memory encoding and retrieval, at least for the interaction-centric queries common in egocentric memory tasks. The evaluation also underscores the value of curated datasets like HME-QA, which guarantee audio availability for multimodal analysis (Section 4).

To further analyze these post-trigger recognition characteristics, it is important to clarify how intermittent video is processed. Our method does not create disjointed clips; rather, it blacks out frames from the original continuous video stream (Section 6.2). This means the downstream Large Multimodal Model (LMM) receives a video that retains its original duration and temporal structure, ensuring that the timing and distance between triggered events remain coherent. Our approach is effectively "training-free" for the downstream task; no architectural changes or fine-tuning were performed on the Gemini 1.5 Pro model. The minimal performance degradation observed between *EgoTrigger* and the "Full" capture baseline (e.g., a <2% drop for ET-1s on HME-QA) can be seen as the empirical answer to how performance fundamentally differs. We attribute this robustness to two factors: 1) the inherent capability of large foundation models, trained on vast and varied web-scale data, to perform reasoning with incomplete or sparse information, and 2) our system's success in preserving the most information-dense segments (HOIs) critical for this specific episodic memory task.

7.3 Performance versus Energy-Efficiency

Achieving all-day operation in form-factor constrained smart glasses necessitates aggressive energy management. The substantial frame reduction achieved by *EgoTrigger* (e.g., over 54% reduction for ET-1s, see Tab. 4, Tab. 5) directly translates to potential energy savings by minimizing the active time of the most power-hungry components in the visual pipeline: the camera sensor, image signal processor (ISP), compression hardware, and wireless transmitter. Given the significant power draw associated with continuous high-resolution video streaming (estimated at 770 mW in our analysis, rapidly depleting typical smart glass batteries), the savings realized by gating these components during periods deemed less critical by the audio trigger can drastically extend device longevity.

However, a complete energy analysis must also consider the cost of the trigger mechanism itself. While *EgoTrigger* saves power by deactivating the visual pipeline, it requires the audio pipeline (microphone, front-end processing) and the audio classifier to run continuously or near-continuously. The energy consumption associated with this – particularly the on-device model's inference cost – represents an overhead compared to a completely idle system. Quantifying this overhead precisely requires measurements on target hardware, as it depends heavily on the chosen platform (e.g., specific chipset capabilities like low-power DSPs or AI accelerators), the final model architecture, and the level

of optimization applied (potentially involving frameworks like TFLite Micro for highly constrained devices).

Despite this inference cost, the energy savings from selectively deactivating the significantly more power-intensive visual pipeline are expected to yield substantial net power reduction compared to continuous visual capture. The favorable position of *EgoTrigger* variants on the accuracy-versus-reduction Pareto frontier (Figure 6) indicates a promising balance. Furthermore, the different triggering strategies (ET-1s vs. ET-Hyst.) offer tunability, enabling system designers to adapt the balance between energy conservation and potential information loss based on specific device limitations and user requirements, paving the way towards practical all-day assistive smart glasses.

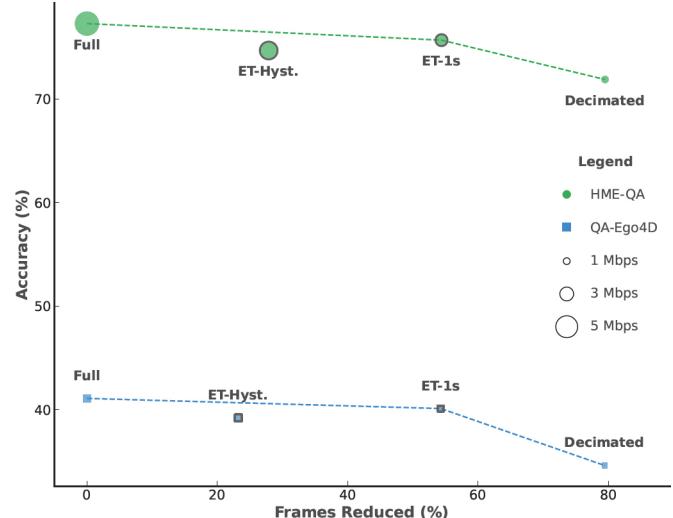


Fig. 6: **Accuracy vs. Frame Reduction Tradeoff.** *EgoTrigger* variants (ET-1s and ET-Hyst.) lie close to the Pareto frontier, balancing QA accuracy with reduced frame count. Results are shown for both HME-QA and QA-Ego4D datasets, with dataset indicated by marker shape (circles for HME-QA, squares for QA-Ego4D) and estimated bitrate reflected by marker size. Outlined markers highlight our proposed methods. Variant ET-1s in particular reduces frame usage and subsequently bitrate while preserving accuracy close to full-frame baselines.

7.4 Empirical Power Consumption Analysis

To empirically validate energy savings, which is infeasible on closed-platform commercial smart glasses, we constructed a simple hardware prototype. This system consists of a Raspberry Pi 5, a USB camera, and a USB microphone, with total power measured using a high-precision USB multimeter. Using this hardware, we measured the total system power draw across several key operating states: an idle baseline (0 FPS), our *EgoTrigger* (ET-1s) approach, a naive decimation baseline, and two high-power baselines representing continuous capture and continuous capture with Wi-Fi transmission.

The empirical results in Tab. 6 demonstrate tangible power savings. Our prototype establishes a 2.37 W idle power floor. From this, the average power for *EgoTrigger* is 4.11 W, representing a **17.3% power reduction** compared to the 4.97 W continuous capture baseline. This shows substantial savings from selective sensing alone. Furthermore, by reducing captured data by 54%, *EgoTrigger* avoids the high overhead of continuous wireless transmission (5.51 W), a critical factor for enabling all-day battery life. While this prototype provides a valuable comparison, we note that a commercial product would use a highly optimized embedded system (e.g., an MCU or ASIC) with a much lower absolute power draw. However, we expect the *relative* energy savings demonstrated by our selective triggering strategy to be comparable in such a specialized system.

Table 6: **Empirical Power Analysis.** Measured power (Watts) of the Raspberry Pi 5 prototype, showing total system draw for each operating state and the contribution of its core components.

Operating Condition / Component	Power (W)
System Idle (0 FPS Capture)	2.37
Raspberry Pi 5 System (Idle)	2.20
Microphone (Connected)	0.17
Baseline (Continuous Capture)	4.97
Raspberry Pi 5 System (Processing)	4.20
Microphone (Always-On)	0.17
Camera (Always-On)	0.60
Baseline + Continuous Wi-Fi Tx	5.51
Raspberry Pi 5 System (Proc. + Tx)	4.74
Microphone (Always-On)	0.17
Camera (Always-On)	0.60
Decimated (0.2 FPS)	3.02
Raspberry Pi 5 System (Processing)	2.60
Microphone (Always-On)	0.17
Camera (Gated, Avg.)	0.25
EgoTrigger (ET-1s, Avg. Power)	4.11
Raspberry Pi 5 System (Processing)	3.50
Microphone (Always-On)	0.17
Camera (Gated, Avg.)	0.44

8 LIMITATIONS AND FUTURE WORK

Despite the promising results, our work highlights several limitations and opens avenues for future investigation:

Hardware Validation, Optimization, and Power Modeling: Our current evaluation relied on offline processing of datasets and high-level energy estimations based on component datasheets and a simple hardware prototype. Validating *EgoTrigger* on real smart glasses hardware is essential to measure true end-to-end performance, latency, and power consumption under real-world conditions. This includes accurately quantifying the on-device model’s inference cost. Furthermore, deploying the machine learning model effectively onto resource-constrained smart glasses hardware will necessitate significant optimization efforts, potentially leveraging frameworks like TensorFlow Lite Micro (TFLM) which are designed for such environments, going beyond the TFLite conversion explored here. Crucially, such optimizations often involve trade-offs, and the potential impact on model accuracy or other performance metrics resulting from these necessary optimization steps has not been quantified in this study.

Trigger Robustness and Scope: The current system depends exclusively on audible Hand-Object Interactions (HOIs) as triggers. This limits its applicability in scenarios involving silent actions (e.g., picking up a soft object), purely visual memories (e.g., observing something without interaction), or environments with high levels of masking noise or unfamiliar sounds that could affect trigger accuracy. Future work should explore integrating complementary low-power sensor modalities. This could include onboard Inertial Measurement Unit (IMU) data from the glasses to detect motion patterns associated with interactions, or even fusing data streams from other wearable devices the user might be wearing concurrently, such as smartwatches, fitness trackers, or potentially IMUs integrated into footwear. Combining these signals could help detect non-auditory interactions, disambiguate sounds, and improve overall trigger robustness. Additionally, exploring more advanced or personalized on-device audio models could enhance robustness to noise and user-specific interaction sounds. Furthermore, future work could explore learning audio representations from unlabeled video, inspired by seminal works like SoundNet [3]. While our current system relies on a pre-trained YAMNet and supervised fine-tuning, a cross-modal self-supervised approach could learn more robust features by leveraging the natural synchronization between visual data and audio. For example, by using a visual model to detect hand-object

interactions in unlabeled egocentric videos, we could automatically generate weak labels to train an audio model to recognize the corresponding HOI sounds. This could significantly enhance the trigger’s robustness across diverse environments and reduce the dependency on manually curated audio datasets, representing a promising direction for scalable and adaptive on-device sensing.

Evaluation Scope: While the LLM-based Question-Answering evaluation provides valuable insights into downstream task performance, it should be complemented by real-time user studies. Assessing the practical usability and effectiveness of *EgoTrigger* in diverse, real-world scenarios with human participants is crucial. Exploring integration with active user interaction models, such as allowing users to explicitly query their memories or providing feedback via a display, would also be valuable future directions.

Privacy Considerations: As with any system involving continuous environmental sensing, particularly audio, careful consideration of user privacy is paramount. Deploying such technology requires robust safeguards, transparent user controls regarding data capture and processing, and adherence to ethical guidelines to ensure user trust and acceptance. *EgoTrigger* shows promise in this direction by gating the transmission of sensor data unless required by a specific experience - in this case, particular human-object interactions for human memory enhancement.

Addressing these points will be crucial for realizing the full potential of passive, intelligent multimodal sensing systems for applications such as human memory enhancement, through wearable devices such as smart glasses.

9 CONCLUSION

Enabling all-day human memory enhancement via smart glasses demands solutions to the significant energy cost associated with continuous camera operation. This paper introduced *EgoTrigger*, an approach leveraging lightweight audio classification of hand-object interactions (HOIs) to intelligently trigger selective image capture. Our evaluations demonstrated that *EgoTrigger* can reduce the number of captured visual frames by an average of 54% while maintaining comparable performance on demanding episodic memory question-answering tasks, significantly outperforming naive decimation strategies. By selectively activating power-hungry visual sensors primarily when contextually relevant audio cues indicate moments likely critical for memory encoding, *EgoTrigger* offers a practical pathway towards realizing energy-efficient smart glasses capable of passive, all-day sensing. This supports applications such as helping users recall daily activities or object locations. The proposed context-aware triggering strategy represents a meaningful step forward for assistive wearable computing, highlighting the potential of using low-power sensor modalities to intelligently manage high-power ones.

10 ACKNOWLEDGMENTS

Akshay Paruchuri conducted this research during an internship under the mentorship of Ishan Chatterjee at Google AR in Seattle, WA, USA. The authors thank the following colleagues at Google for insightful discussions around multimodal sensing and energy efficient computing: Zhihan Zhang, Jake Garrison, Shamik Ganguly, Zixuan Qu, and Anish Prabhu.

REFERENCES

[1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1

[2] A. Anthropic. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*, 1:1, 2024. 1

[3] Y. Aytar, C. Vondrick, and A. Torralba. Soundnet: Learning sound representations from unlabeled video. *Advances in neural information processing systems*, 29, 2016. 8

[4] J.-H. Bahk, H. Fang, K. Yazawa, and A. Shakouri. Flexible thermoelectric materials and device optimization for wearable energy harvesting. *Journal of Materials Chemistry C*, 3(40):10362–10374, 2015. 2

[5] L. Bärmann and A. Waibel. 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*, pp. 1560–1568, 2022. 2, 4

[6] M. Billinghurst and T. Starner. Wearable devices: new ways to manage information. *Computer*, 32(1):57–64, 1999. doi: 10.1109/2.738305 1, 2

[7] B. Chakravarthi, A. A. Verma, K. Daniilidis, C. Fermuller, and Y. Yang. Recent event camera innovations: A survey. *arXiv preprint arXiv:2408.13627*, 2024. 2

[8] C. Chen, P. Peng, A. Baid, Z. Xue, W.-N. Hsu, D. Harwath, and K. Grauman. Action2sound: Ambient-aware generation of action sounds from egocentric videos. In *European Conference on Computer Vision*, pp. 277–295. Springer, 2024. 4

[9] K. Chen, T. Wan, N. Matsuda, A. Ninan, A. Chapiro, and Q. Sun. Peapods: Perceptual evaluation of algorithms for power optimization in xr displays. *ACM Transactions on Graphics (TOG)*, 43(4):1–17, 2024. 2

[10] Y. Chu, J. Xu, X. Zhou, Q. Yang, S. Zhang, Z. Yan, C. Zhou, and J. Zhou. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*, 2023. 2

[11] V. Corporation. Vuzix m400™ smart glasses. Accessed: 2025-04-10. 1, 2

[12] R. David, J. Duke, A. Jain, V. Janapa Reddi, N. Jeffries, J. Li, N. Kreeger, I. Nappier, M. Natraj, T. Wang, et al. Tensorflow lite micro: Embedded machine learning for tinyml systems. *Proceedings of Machine Learning and Systems*, 3:800–811, 2021. 4

[13] R. W. DeVaul, A. S. Pentland, and V. R. Corey. The memory glasses: Subliminal vs. overt memory support with imperfect information. In *Proceedings of the 7th IEEE International Symposium on Wearable Computers*, ISWC '03, p. 146. IEEE Computer Society, USA, 2003. 1

[14] T. Developers. Tensorflow. *Zenodo*, 2022. 3

[15] G. A. Edge. Audio classification guide, 2024. Accessed: 2025-04-11. 4

[16] F. Eyben, M. Wöllmer, and B. Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pp. 1459–1462, 2010. 2

[17] R. Gao, T.-H. Oh, K. Grauman, and L. Torresani. Listen to look: Action recognition by previewing audio. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10457–10467, 2020. 2

[18] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 776–780. IEEE, 2017. 2

[19] E. J. Gonzalez, I. Chatterjee, M. Gonzalez-Franco, A. Colaço, and K. Ahuja. Intent-driven input device arbitration for xr. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pp. 1–5, 2024. 2

[20] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18995–19012, 2022. 2, 4

[21] C. Herglotz, M. Kränzler, R. Ludwig, and A. Kaup. Video decoding energy reduction using temporal-domain filtering. In *Proceedings of the First International Workshop on Green Multimedia Systems*, pp. 22–27, 2023. 5

[22] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. Weiss, and K. Wilson. Cnn architectures for large-scale audio classification. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2017. 2, 3

[23] J. Hoisko. Context triggered visual episodic memory prosthesis. In *Proceedings of the 4th IEEE International Symposium on Wearable Computers*, ISWC '00, p. 185. IEEE Computer Society, USA, 2000. 1

[24] F. Ianì. Embodied memories: Reviewing the role of the body in memory processes. *Psychonomic bulletin & review*, 26(6):1747–1766, 2019. 2, 3

[25] A. Jindal, K. Wolski, K. Myszkowski, and R. K. Mantiuk. Perceptual model for adaptive local shading and refresh rate. *ACM Transactions on Graphics (TOG)*, 40(6):1–18, 2021. 2

[26] H. Kalantarian, N. Alshurafa, M. Pourhomayoun, and M. Sarrafzadeh. Power optimization for wearable devices. In *2015 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)*, pp. 568–573. IEEE, 2015. 2

[27] E. Kazakos, A. Nagrani, A. Zisserman, and D. Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5492–5501, 2019. 2

[28] I. Khan, S. Khusro, S. Ali, and J. Ahmad. Sensors are power hungry: An investigation of smartphone sensors impact on battery power from lifelogging perspective. *Bahria University Journal of Information & Communication Technology*, 9(2):8–19, 2016. 1, 2

[29] M. U. Khattak, M. F. Naeem, J. Hassan, M. Naseer, F. Tombari, F. S. Khan, and S. Khan. How good is my video lmm? complex video reasoning and robustness evaluation suite for video-lmms. *arXiv preprint arXiv:2405.03690*, 2024. 4

[30] P. Kindt, D. Yunge, M. Gopp, and S. Chakraborty. Adaptive online power-management for bluetooth low energy. In *2015 IEEE Conference on Computer Communications (INFOCOM)*, pp. 2695–2703. IEEE, 2015. 2

[31] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4

[32] A. Krishna, H. Shankaranarayanan, H. P. Oleti, A. Chauhan, A. van Schaik, M. Mehendale, and C. S. Thakur. Tinyml acoustic classification using raman accelerator and neuromorphic cochlea. In *2023 IEEE Asia Pacific Conference On Postgraduate Research In Microelectronics And Electronics (PRIMEAsia)*, pp. 44–45. IEEE, 2023. 4

[33] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 4

[34] M. Maaz, H. Rasheed, S. Khan, and F. S. Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023. 4

[35] K. Mangalam, R. Akshulakov, and J. Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023. 2, 4

[36] Meta. Introducing orion, our first true augmented reality glasses, Sept. 2024. Accessed: 2025-04-10. 1

[37] C. Plizzari, A. Tonioni, Y. Xian, A. Kulshrestha, and F. Tombari. Omnia de egotempo: Benchmarking temporal understanding of multi-modal llms in egocentric videos. *arXiv preprint arXiv:2503.13646*, 2025. 2, 4

[38] W. B. Qaim, A. Ometov, A. Molinaro, I. Lener, C. Campolo, E. S. Lohan, and J. Nurmi. Towards energy efficiency in the internet of wearable things: A systematic review. *IEEE Access*, 8:175412–175435, 2020. 2

[39] X. Qi, M. Keally, G. Zhou, Y. Li, and Z. Ren. Adasense: Adapting sampling rates for activity recognition in body sensor networks. In *2013 IEEE 19th Real-Time and Embedded Technology and Applications Symposium (RTAS)*, pp. 163–172. IEEE, 2013. 2

[40] Y. Qian, H. Zhang, Y. Yang, and Z. Gan. How easy is it to fool your multimodal llms? an empirical analysis on deceptive prompts. *arXiv preprint arXiv:2402.13220*, 2(7), 2024. 4

[41] I. Qualcomm Technologies. Snapdragon ar1 gen 1 platform. <https://www.qualcomm.com/products/mobile/snapdragon/xr-vr-ar/snapdragon-ar1-gen-1-platform>, 2023. Accessed: 2025-04-11. 4

[42] Ray-Ban and Meta. Ray-Ban Meta Smart Glasses. <https://www.ray-ban.com/usa/discover-ray-ban-meta-smart-glasses>. Accessed: 2025-04-07. 1, 2

[43] B. J. Rhodes. The wearable remembrance agent: a system for augmented memory. In *Proceedings of the 1st IEEE International Symposium on Wearable Computers*, ISWC '97, p. 123. IEEE Computer Society, USA, 1997. 1

[44] O. Rybakov, N. Kononenko, N. Subrahmanyam, M. Visontai, and S. Laurenzo. Streaming keyword spotting on mobile devices. *arXiv preprint arXiv:2005.06720*, 2020. 4

[45] J. Shen, J. J. Dudley, and P. O. Kristensson. Encode-store-retrieve: Aug-

menting human memory through language-encoded egocentric perception. In *2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 923–931. IEEE, 2024. 2

[46] R. Sunder, U. K. Lilhore, A. K. Rai, E. Ghith, M. Tlija, S. Simaiya, and A. H. Majeed. Smartapm framework for adaptive power management in wearable devices using deep reinforcement learning. *Scientific Reports*, 15(1):6911, 2025. 2

[47] S. Tan, T. Nagarajan, and K. Grauman. Egodistill: Egocentric head motion distillation for efficient video understanding. *Advances in Neural Information Processing Systems*, 36:33485–33498, 2023. 2

[48] G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 1

[49] B. Tian, Y. Pang, M. Huzefa, S. Wang, and S. Adve. Towards energy-efficiency by navigating the trilemma of energy, latency, and accuracy. In *2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 913–922. IEEE, 2024. 2

[50] J. Yang, S. Liu, H. Guo, Y. Dong, X. Zhang, S. Zhang, P. Wang, Z. Zhou, B. Xie, Z. Wang, et al. Egolife: Towards egocentric life assistant. *arXiv preprint arXiv:2503.03803*, 2025. 2