Rational Episodic Memory: When Should Wearable AI Remember?

Joseph Letobar[1]

Department of Computer Science, University of Central Florida, Orlando, FL 32816, USA
joseph.letobar@ucf.edu

Abstract. Wearable AI systems rely on long-term memory to support retrieval, question answering, and contextual assistance. Yet an always-on system must decide what to preserve before knowing which experiences will later become relevant. We investigate two cognitively inspired principles for question-agnostic memory admission under fixed storage budgets. The first uses visible engagement as a rational-action-inspired proxy for relevance to the wearer's goals. The second uses semantic surprise from pretrained visual representations as evidence of eventful change.

We evaluate these signals as write policies on long egocentric recordings, measuring both whether future-question-relevant intervals are substantially retained and how much of their duration is preserved. On eight long recordings averaging 4.14 hours, both policies outperform uniform sampling, while engagement achieves the strongest retention across all tested budgets. A secondary analysis on shorter, more event-dense recordings generally favors surprise, consistent with prediction-error-based accounts of event segmentation. These results suggest that memory admission may depend on the temporal structure of experience: surprise is effective for detecting densely occurring event changes, whereas engagement is useful for preserving relevant episodes over extended everyday activity.

Keywords: Wearable AI · Egocentric Vision · Episodic Memory · Human-Centered AI

1 Introduction

Wearable AI systems promise persistent assistants capable of answering questions such as "Where did I leave my wallet?" or "When was the last time I took my medication?" Realizing this vision requires an always-on system that continuously observes the user's daily life. However, storing every observation is computationally impractical and quickly exceeds realistic memory budgets.

Humans face the same challenge. Although we receive a continuous stream of rich sensory information, we do not encode every detail into episodic memory. We rarely remember the color of every person's shoes or every crack in the sidewalk, even though those sensory signals were available to us. Instead, episodic memory is selective, preferentially encoding experiences that may be useful in the future.

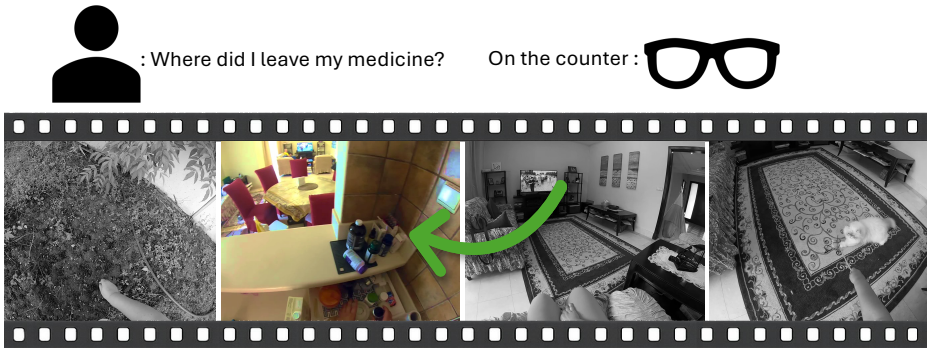

Fig. 1: A future user query requires localizing the relevant experience within a full day of egocentric observations.

We draw inspiration from this selectivity to study an *upstream* question in wearable AI memory: what should be written to memory in the first place? Prior wearable and egocentric systems construct memory through continuous capture, attention- or content-based selection, and subsequent compression or consolidation [1, 24, 25, 42, 46, 47]. However, comparatively little work treats memory admission by deciding which incoming experiences warrant long-term retention before the system knows what the user will later ask.

Cognitive Science offers many explanations for why an experience may be worth preserving. We draw on two well-established ideas that lend themselves to computational implementation: evidence of sustained goal-directed behavior and prediction error indicating eventful change.

Research on Theory of Mind and action understanding suggests that humans infer the goals of others by assuming that intelligent agents select actions rationally to achieve those goals [12]. Observed behavior can therefore provide evidence about what another person is trying to accomplish. An egocentric memory system similarly observes the wearer's behavior without direct access to their internal goals. This motivates visible engagement as a proxy for goal relevance: sustained interaction with an object, person, or environment suggests that the ongoing experience may be relevant to what the wearer is trying to accomplish.

Predictive-processing theories propose that perception continually compares incoming observations with expectations about what is likely to occur next [11], while Event Segmentation Theory argues that continuous experience is divided into discrete events when those expectations are violated [50]. This motivates our second candidate signal: change in the current visual representation relative to recent experience, used as a proxy for prediction-error-driven event boundaries. Our measure approximates the effect described by Itti and Baldi [20] and supported empirically by Kumar et al. [22] through distance in a pretrained latent space.

Motivated by these perspectives, we examine visible engagement and surprise as alternative signals for question-agnostic memory admission. On long recordings, engagement retains more future-question-relevant evidence than surprise or uniform sampling, supporting the value of an agent-relative signal over extended experience. In a secondary analysis of shorter, more event-dense recordings, surprise instead performs best. These recordings contain more frequent discrete actions and transitions, a setting in which prediction error may provide a particularly effective boundary signal, consistent with Event Segmentation Theory. Together, these results suggest that engagement is well suited to preserving sustained goal-relevant episodes, whereas surprise becomes more useful when experience is densely structured around distinct events.

Applications in wearable AI extend beyond episodic memory formation to decisions about when a wearable assistant should surface context or intervene proactively. The same question-agnostic cues could also support long-duration robots and egocentric dataset curation, with engagement emphasizing sustained goal-directed behavior and surprise emphasizing event transitions.

2 Related Work

2.1 Egocentric Summarization and Memory

Retrospective Selection and Retrieval. Early egocentric summarization methods select representative frames or segments from an already recorded video using interaction, attention, coherence, diversity, or visual uniqueness [5, 26, 28, 33, 47]. Query-focused methods instead retrieve portions of an existing recording that match a supplied concept or question [34–37, 42]. These approaches improve summarization or retrieval after capture, but do not determine what should be stored before the user's future information needs are known.

Other systems compress or organize complete experiences into latent states, knowledge graphs, or user models [4, 41, 45]. Such representations can support efficient reasoning, but memory admission is typically learned indirectly from downstream supervision or replaced by continual structuring of the observed stream.

Online Memory Formation. Several systems make capture decisions as experience unfolds. StartleCam monitors skin conductivity and saves a buffer of recent images when it detects a startle response, while SenseCam captures images periodically and in response to changes detected by onboard light, motion, temperature, and passive-infrared sensors [18, 19]. Lin et al. classify the activity or scene of each incoming segment and apply a context-specific model to decide whether it should be retained in the summary, without first processing the complete recording [27]. AMEGO more directly constructs an active memory online by initiating object memories from hand-object interactions and identifying activity-centric locations [13]. These systems establish precedents for causal

selection, but their admission criteria rely on additional sensors, highlight objectives, or predefined interaction and object cues.

Related egocentric work studies wearer-centered signals without necessarily applying them to memory admission. Su and Grauman infer engagement from egomotion, while other approaches use gaze or social engagement as evidence of wearer attention and subjective relevance [39, 40, 47]. These findings motivate our examination of engagement as a possible memory-admission signal, alongside surprise.

2.2 Memory in Embodied and Language Agents

Embodied-agent memory systems commonly store observations in semantic maps, scene graphs, or timestamped databases for later retrieval and reasoning [2, 8, 14, 17]. Related work learns compressed latent memories through question-answering supervision or controls consolidation using repetition, similarity, and forgetting [4, 32].

Most closely related, *Worth Remembering* formulates admission as an explicit query-independent gating problem and stores episodes around moments of Bayesian surprise computed from predictive visual representations [15]. Its results show that prediction error can outperform uniform and random storage for robot question answering. We extend this framing by comparing surprise with visible engagement for wearable memory admission.

Language-agent systems likewise manage memories through ranking, consolidation, and updating [7, 30, 31, 48, 52], often using downstream answer quality, utility, novelty, or semantic surprise [38, 49, 51]. Unlike these approaches, we study the *upstream* admission of raw egocentric observations before future queries are known, comparing engagement and surprise as stream-computable signals for question-agnostic memory admission.

3 Method

We evaluate two signals for determining which observations from a continuous egocentric video stream should be retained before future queries are known. First, a supervised engagement estimator predicts whether the wearer is behaviorally engaged with the current activity, providing a proxy for sustained goal relevance. Second, an unsupervised semantic-surprise measure estimates how strongly the current visual representation departs from recent experience, providing a signal of eventful change. Neither signal requires a persistent user model or identity-specific calibration at inference, though personalization is a possible extension discussed in Section 5. We evaluate these signals independently as question-agnostic memory write policies under fixed storage budgets. The following sections describe their computation in detail.

3.1 Engagement

Definition and Model. We treat engagement as a supervised temporal classification problem: given an egocentric video window, predict whether it belongs to an annotated engagement episode, following the formulation of Su and Grauman [39]. Whereas their method proposes variable-length engagement intervals, we use fixed-length sliding windows to produce continuous estimates suitable for memory admission, where exact temporal boundaries are less important than retaining the relevant episode and limited surrounding context.

We represent each window using frozen V-JEPA 2 embeddings [3]. Because V-JEPA 2 is pretrained to predict masked spatiotemporal representations, its embeddings capture motion, semantic continuity, and scene structure useful for engagement recognition. These embeddings are provided to a Random Forest classifier trained on binary engagement labels using 64-frame windows with a stride of 32 frames.

The classifier contains 200 trees and uses random seed 0. At inference, it outputs the probability that each window belongs to an engagement episode. Predictions from overlapping windows are aggregated to produce a continuous engagement signal over time. The primary evaluation uses this Random Forest head. The effects of the frozen representation and classifier choice are examined in the ablation study in Section 4.2.

Annotations. We annotated a subset of the Wearable AI benchmark [23, 44], which contains short egocentric videos spanning activities such as shopping, sightseeing, cooking, home improvement, exercise, and tourism. Our subset contains 40 videos totaling 7.05 hours, divided into 31 training videos and 9 held-out videos. All windows from a video remain within the same partition to prevent leakage.

Videos were selected to include both routine periods, such as walking or commuting, and sustained goal-directed activities, such as cooking or cleaning. Engagement intervals were annotated using the operational definition of Su and Grauman [39]: the wearer is engaged when they purposefully gather information about an object, scene, or ongoing task. Examples include inspecting products, reading signs, manipulating objects, or examining an unfamiliar environment. For activities not explicitly represented in the original dataset, we applied the same definition to the broader context. For example, an extended LEGO assembly sequence was treated as one continuous engagement episode because the wearer remained focused on locating pieces and constructing the model.

Engagement is inherently subjective, particularly at episode boundaries. Su and Grauman reported substantially higher agreement on whether engagement occurred (F1 = 0.914) than on its exact temporal boundaries (frame-level F1 = 0.818; interval-boundary F1 = 0.837) [39]. Although we did not conduct a multi-annotator consistency study, our objective is similarly to identify meaningful engagement episodes rather than recover frame-exact onset and offset times. Small boundary differences are therefore unlikely to materially affect the intended memory-write decision.

3.2 Surprise

Surprise is an unsupervised temporal novelty signal computed from the same V-JEPA 2 embeddings used for engagement estimation.

Let $z_t \in \mathbb{R}^d$ denote the pooled V-JEPA 2 embedding for a 64-frame window. Consecutive windows use a stride of 32 frames and therefore overlap by half their duration.

We define latent surprise as the standardized change between consecutive embeddings:

$$R_t = \frac{1}{2d} \sum_{i=1}^{d} \left(\frac{z_{t,i} - z_{t-1,i}}{\max(\sigma_i, \epsilon)} \right)^2,$$

where σ_i is estimated from a development set and $\epsilon > 0$ prevents division by zero. This is a dimension-normalized half squared Mahalanobis distance under a shared diagonal covariance model. We use it as a Bayesian-surprise-inspired proxy for latent visual change, rather than an exact Bayesian surprise measure.

We map R_t to $[0, 1]$ using fixed calibration values estimated on development data disjoint from the evaluation set:

$$\widetilde{S}_t = \text{clip}\left(\frac{R_t - r_{\min}}{r_{\max} - r_{\min}}, 0, 1 \right),$$

where $r_{\min} = 0.01$ and $r_{\max} = 0.25$. Finally, we apply causal exponential smoothing:

$$S_t = 0.25\widetilde{S}_t + 0.75S_{t-1}, \qquad S_1 = \widetilde{S}_1.$$

The resulting $S_t \in [0, 1]$ is held constant until the next 32-frame window is available.

4 Experiments

We evaluate engagement and surprise as question-agnostic memory-admission signals on 16 Ego4D [16] recordings totaling 42.71 hours. The evaluation includes a long-recording regime and an analysis of shorter recordings. We compare both signals with uniform sampling under matched storage budgets. We additionally report ablations of the engagement estimator's representation and classifier head.

4.1 Memory Write Policy Evaluation

Setup. The main evaluation contains eight shorter recordings averaging 1.20 hours and eight long recordings averaging 4.14 hours. Because our central research question concerns selective memory over extended experience, the long-recording subset serves as the primary evaluation regime, while the shorter recordings provide a secondary comparison across temporal scales.

Although the subsets are defined by duration, they also differ qualitatively in their within-recording activity composition. The shorter recordings are generally *event-dense*: most footage is devoted to one sustained activity or a closely related sequence, with relatively frequent subactions or transitions within that activity, such as baking, carpentry, gardening, cleaning, or exercise. The longer recordings are more *activity-heterogeneous*, spanning multiple activities and transitions across broader portions of everyday experience, including walking, conversation, sightseeing, restaurants, phone use, transportation, household tasks, and social interaction. We use these terms descriptively rather than treating activity composition as an independently controlled factor. We therefore report the subsets separately, while treating the long recordings as the setting most directly aligned with long-horizon memory.

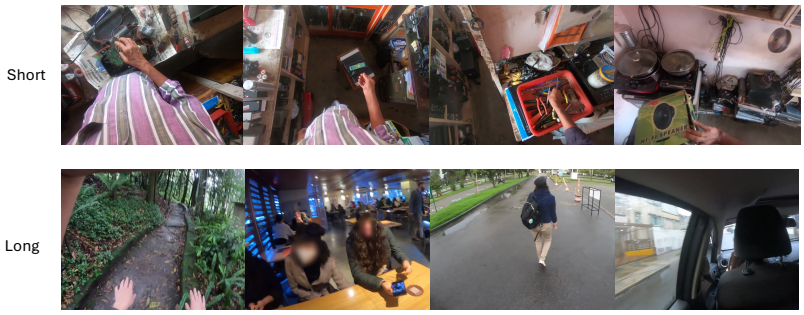

Fig. 2: Qualitative comparison across temporal regimes. Top: A *shorter*, event-dense recording centered on one sustained activity. Bottom: A *longer*, activity-heterogeneous recording spanning multiple activities and transitions. These examples illustrate the descriptive distinction between the subsets.

To evaluate whether a write policy retains useful information under a limited storage budget, we use the temporal intervals from Ego4D Natural Language Queries (NLQ) as targets. NLQ pairs natural-language questions about past egocentric experience with the target evidence intervals needed to answer them [16]. Retaining these intervals therefore provides a proxy for whether the memory contains information that could support plausible future queries while discarding portions of the stream less relevant to those information needs.

Of the 16 selected recordings, 10 contain official NLQ intervals, but only two of the eight long recordings do. For the remaining six long recordings, we construct 149 narration-derived question–interval pairs following the same format. We select question-worthy narrated events distributed throughout each recording, convert them into natural-language questions, and assign the target evidence intervals needed to answer them. Selection is performed without access to engagement or surprise scores, and results using official and researcher-created intervals are reported separately.

The engagement and surprise signals are computed incrementally as temporal windows become available, without access to future queries or evaluation annotations. To compare the signals under precisely matched storage budgets, we use an offline top-K protocol: after scoring a complete recording, we retain exactly the number of windows permitted by each budget. This controlled protocol isolates the experimental question of how well each signal ranks potentially useful observations from deployment-specific choices such as threshold calibration, quota accounting, and boundary handling. Uniform sampling is given the same window count; translating these scores into a causal streaming controller is a complementary systems step left to future work.

Results. For each matched storage budget, Table 1 reports Hit@50, which counts a target evidence interval as retained only when at least half of its duration is preserved, and annotated-time recall, which measures the fraction of total target evidence duration retained. These metrics capture both the number of substantially preserved intervals and the temporal extent of relevant evidence maintained in memory.

Table 1: Memory retention across temporal regimes.* The long subset contains 8 recordings averaging 4.14 hours with 215 target evidence intervals; the short subset contains 8 recordings averaging 1.20 hours with 292 target evidence intervals. Entries report Hit@50 / annotated-time recall, in percentages.

	Long recordings			Short recordings		
Budget	**Engagement**	**Surprise**	**Uniform**	**Engagement**	**Surprise**	**Uniform**
30%	**42.3** / **44.6**	33.0 / 41.5	26.4 / 29.7	15.8 / 15.2	**28.8** / **31.6**	22.1 / 30.1
25%	**33.5** / **38.1**	28.8 / 35.9	22.4 / 25.2	12.0 / 13.0	**24.0** / **27.8**	17.6 / 25.0
20%	**30.7** / **32.6**	20.5 / 29.6	18.6 / 20.2	7.9 / 8.8	**17.5** / **23.3**	14.5 / 19.9
15%	**25.6** / **27.5**	13.5 / 21.5	13.0 / 15.0	7.2 / 7.2	**11.0** / **15.9**	10.9 / 15.0
10%	**16.7** / **18.1**	9.3 / 14.0	9.0 / 10.0	3.8 / 5.2	6.2 / **10.9**	**7.3** / 10.0
5%	**10.7** / **11.1**	4.2 / 7.3	4.2 / 4.8	2.7 / 3.2	2.4 / **5.4**	**3.5** / 4.9

* Annotation-source sensitivity is evaluated within the long-recording subset (2 recordings with official NLQ intervals, 6 with researcher-created intervals). Engagement outperforms Surprise on Hit@50 at all six budgets for both sources: the Engagement–Surprise gap is +5.8–+11.5 percentage points for recordings with researcher-created intervals and +10.6–+25.8 points for recordings with official intervals. The ordering of this gap across budgets correlates only moderately between the two sources ($\rho = 0.580$, mean absolute discrepancy 9.54 percentage points), so the size of the advantage should not be treated as stable, even though its direction is consistent. Annotated-time recall agrees at four of the six budgets; the researcher-created intervals weakly favor Surprise at the 25% and 30% budgets, by 0.5–0.9 percentage points, which, given the official comparison covers only two recordings, more plausibly reflects video-level variance than a systematic annotation-source effect.

The two temporal regimes produce a reversal in relative performance. On long recordings, engagement achieves the highest Hit@50 and annotated-time

recall at every tested budget, with the largest gains appearing at tighter storage constraints. On shorter, event-dense recordings, surprise performs best across most budgets, while uniform sampling remains competitive only at the lowest budgets. These results indicate that the relative effectiveness of the two signals may depend on the temporal structure of the evaluated recording.

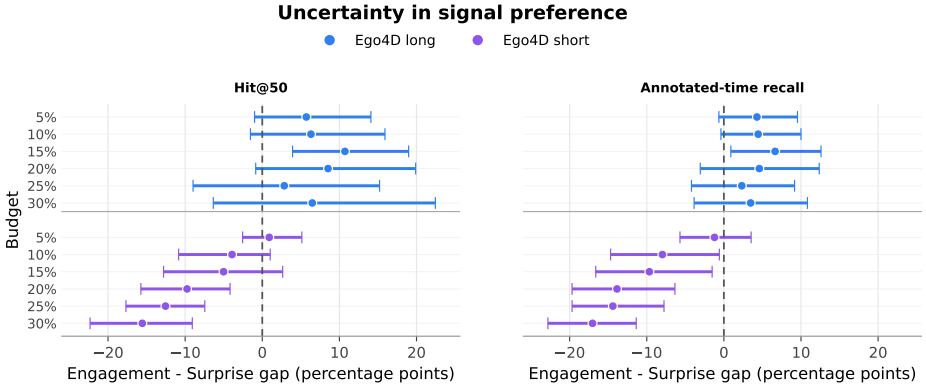

Fig. 3: Uncertainty in signal preference. Points show the mean paired Engagement–Surprise gap in percentage points; whiskers show 95% confidence intervals. Positive values favor Engagement.

Figure 3 shows that the Engagement-minus-Surprise differences are directionally consistent with the pattern observed in Table 1. The long-recording subset generally favors Engagement, whereas the short-recording subset generally favors Surprise. However, the intervals are wide and several cross zero, reflecting video-level variation and the limited sample size. Importantly, we do not interpret recording duration itself as the explanatory variable; the long/short split is instead a coarse descriptive contrast that often coincides with differences in temporal activity density.

4.2 Engagement Ablations

Engagement Representation Ablation. We additionally compared V-JEPA 2 embeddings with VideoMAE [43] and with a motion-based engagement representation following Su and Grauman [39]. Dense Farnebäck optical flow is computed between successive frames using OpenCV [6,10], temporally smoothed, and summarized with a hierarchical spatiotemporal pyramid that pools horizontal and vertical motion statistics across temporal segments and spatial cells. All three representations use the same engagement annotations and the same 31-video training / 9-video held-out split. Table 2 summarizes the comparison.

On this held-out split, V-JEPA 2 achieved the highest F1, recall, and ROC AUC, while optical flow achieved the highest precision and specificity; VideoMAE was lower across the reported metrics. Because these representations were

Table 2: Engagement representation ablation. Results report the best evaluated window-size/stride configurations for optical flow (128/64) and V-JEPA 2 (64/32), while VideoMAE uses its standard 16-frame configuration (16/4), all on the 31-video training / 9-video held-out split.

Representation	Win./Stride	F1	Prec.	Recall	Spec.	AUC
Optical Flow	128/64	0.861	**0.891**	0.833	**0.851**	0.908
VideoMAE	16/4	0.740	0.762	0.719	0.671	0.807
V-JEPA 2	64/32	**0.883**	0.863	**0.904**	0.788	**0.931**

evaluated only as engagement estimators and not in the downstream memory-retention task, we treat this as an implementation ablation rather than a general claim about representation quality.

Engagement Classifier Ablation. To examine the effect of the classifier head independently of the frozen representation, we trained a linear probe, a small multilayer perceptron (MLP), and the Random Forest on the same cached V-JEPA 2 embeddings, annotations, and 31/9 video split. As shown in Table 3, the Random Forest achieved the strongest held-out accuracy, F1, recall, and ROC AUC, and is therefore used for the primary memory-retention evaluation.

Table 3: Engagement classifier ablation. Results use frozen V-JEPA 2 embeddings with a 64-frame window and 32-frame stride on the 31-video training / 9-video held-out split.

Classifier	Acc.	F1	Prec.	Recall	AUC
Linear probe	0.817	0.835	**0.902**	0.778	0.913
MLP	0.806	0.825	0.891	0.767	0.915
Random Forest	**0.857**	**0.883**	0.863	**0.904**	**0.931**

5 Discussion

5.1 Interpretation

The observed reversal across temporal regimes is broadly consistent with the intended roles of the two signals. Engagement asks whether the wearer appears to be acting purposefully in pursuit of an ongoing goal. This signal is most informative in long, activity-heterogeneous recordings that contain a mixture of routine periods and distinct activities, such as walking, driving, eating, working, and interacting with other people. In this setting, engagement can distinguish relatively sparse periods of sustained goal-directed behavior from the broader

background of everyday experience, which may explain its stronger performance on the long recordings.

The shorter recordings are qualitatively event-dense. They are more often organized around a single sustained activity, such as cooking, cleaning, gardening, or exercise, with relatively little inactive or unrelated footage. We hypothesize that the engagement signal becomes saturated in this setting: when the wearer appears engaged throughout most of the recording, engagement provides little discrimination about which moments should receive a limited memory budget.

Figure 4 shows the distribution of predicted engagement across the original recordings. The short recordings have higher mean predicted engagement than the long recordings, consistent with the possibility that engagement is more saturated in the short subset. This figure is descriptive and does not establish a recording-level correlation or causal relationship between engagement saturation and memory discrimination.

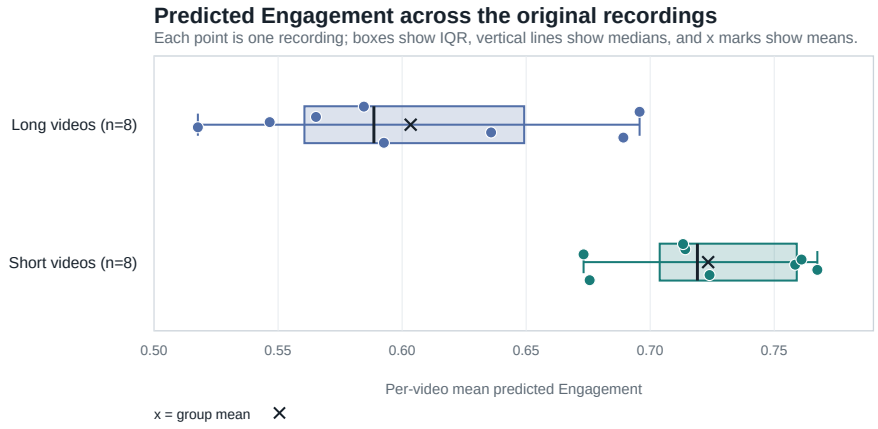

Fig. 4: Predicted engagement across the original recordings. Each point is one recording; boxes show the interquartile range, vertical lines show medians, and crosses show group means.

Long, task-saturated case. In attempt to separate recording duration from activity structure, we additionally analyze a 5.2 hour welding recording dominated by one sustained, highly engaged task. Because this recording differs qualitatively from the activity-heterogeneous long recordings, it is excluded from the primary aggregate and treated as a separate case study. Representative frames from this case study are shown in Figure 5.

As shown in Table 4, the welding case produces a budget-dependent result. Surprise achieves substantially higher Hit@50 at moderate and high storage budgets, despite the recording being long, while engagement becomes stronger

Fig. 5: Qualitative samples from the long, task-saturated welding recording. Although the recording spans several hours, the wearer remains engaged in a single sustained activity. Engagement is therefore broadly elevated, while surprise provides greater discrimination among changes in tools, materials, viewpoints, and welding subactions.

Table 4: Memory retention on a long, task-saturated welding recording. The video is dominated by one sustained activity. Entries are percentages.

	Hit@50		Annotated-time recall	
Budget	**Engagement**	**Surprise**	**Engagement**	**Surprise**
30%	26.3	**61.4**	42.1	**52.2**
25%	24.6	**42.1**	**41.5**	40.8
20%	17.5	**29.8**	**34.8**	30.1
15%	15.8	**19.3**	**29.4**	22.2
10%	**14.0**	3.5	**22.8**	10.8
5%	**7.0**	1.8	**10.6**	0.3

under the tightest budgets and preserves more annotated duration at most operating points. This provides evidence against duration alone explaining the difference between the two main subsets. Instead, the temporal structure of the recording appears important: when one sustained task causes engagement to remain broadly elevated, surprise can provide greater discrimination by identifying changes within that activity. As a single-video case study, however, this result should be interpreted as supporting evidence rather than a general conclusion.

5.2 Limitations

This work is an early exploration conducted on a limited number of recordings and does not yet evaluate complete day-long wearable deployments. The current evaluation also uses the known recording duration to determine the number of retained windows at each percentage budget and does not evaluate a causal online quota controller or an end-to-end question-answering system. The study also depends on researcher-created target intervals. Engagement intervals were labeled by a single annotator using the operational definition of Su and Grauman [39].

For selected Ego4D recordings without official NLQ intervals, we additionally constructed narration-derived question–interval pairs representing information that a wearer might reasonably ask about later. These intervals were selected without access to engagement or surprise scores.

Both the learned engagement model and the retention evaluation may therefore reflect subjective annotation decisions. We compare against uniform sampling only; existing memory-selection methods are not evaluated, limiting direct comparison with prior selection strategies. We did not explicitly evaluate cross-user generalization, so user-independent performance remains an open question.

The proposed explanation for the difference between short and long recordings also remains preliminary. We hypothesize that engagement becomes saturated in shorter, event-dense recordings, while surprise benefits from their denser sequence of event transitions. Because duration and activity composition vary together in the present evaluation, we cannot truly isolate which factor causes the observed reversal.

5.3 Future Work

Future work should expand the evaluation scale, improve annotation reliability, and better separate recording duration from activity composition.

Towards Deployment. A natural next step is to evaluate a deployment-oriented system in which engagement and surprise scores are converted into memory writes using causal thresholding, quota management, or buffering as the stream arrives. Such a system could be evaluated end-to-end on a larger standardized video question-answering benchmark, using downstream answer quality rather than retrospective interval retention as the primary metric. Because question answering is a common evaluation task for memory systems, this would also enable more direct comparisons with existing memory-admission and selection methods. Question-answer annotations are more widely available than the fine-grained evidence intervals used here, potentially expanding the evaluation set, but this would require an additional retrieval-and-answering pipeline beyond the scope of this preliminary study.

The behavior of engagement and surprise motivates further study of fusion strategies. Although we explored simple combinations, the results were not sufficiently consistent to support a fusion claim in this work. Future systems could therefore investigate whether using both signals as complementary evidence improves memory admission across different temporal structures.

Beyond Visual Memory. While this work estimates engagement primarily from visual behavior, future systems could fuse visual cues with gaze, inertial, physiological, and other wearable measurements to improve engagement estimation. Wearable-memory and egocentric systems have already explored physiological cues and gaze for identifying memorable experiences and visual attention [18, 47], while platforms such as Project Aria demonstrate the broader multimodal sensing stack available for egocentric research [9]. Recent work has demonstrated user-generalizable surface electromyography for wearable human–computer interaction [21], as demonstrated by Meta's Ray-Ban Display glasses,

which pair egocentric cameras with a wrist-worn EMG interface [29]. Such signals could provide additional evidence of manipulation intent, similar in spirit to AMEGO, which builds an online memory around hand–object interactions using visual cues alone [13].

Towards User Models. While this work approximates surprise from local feature dynamics, future systems could model it relative to a persistent personalized user model encoding recurring environments, objects, routines, and individuals. Recent work has begun constructing persistent egocentric user models [45], while Gorlo et al. identify habituation as an important direction for avoiding repeated surprise responses to familiar observations [15]. Combining these ideas could support more human-like memory formation, where novel observations are initially salient but gradually become unsurprising as the internal model adapts.

6 Conclusion

We present preliminary evidence that visible engagement and surprise can support question-agnostic memory admission under fixed storage budgets. Their relative effectiveness varies with temporal structure: engagement performs best on long, activity-heterogeneous recordings, whereas surprise performs better on shorter, event-dense recordings. These results do not establish a universal write policy, but suggest that cognitively motivated signals can provide useful alternatives relative to a uniform-sampling baseline while motivating larger online and downstream evaluations.

Acknowledgements

The author thanks Clint Johnson at the Georgia Tech Research Institute for valuable discussions and feedback during the development of this work, which helped refine the paper's framing and presentation.

References

1. Aizawa, K., Ishijima, K., Shiina, M.: Summarizing wearable video. In: Proceedings of the 2001 International Conference on Image Processing. vol. 3, pp. 398–401. IEEE (2001). https://doi.org/10.1109/ICIP.2001.958135
2. Anwar, A., Welsh, J., Biswas, J., Pouya, S., Chang, Y.: ReMEmbR: Building and reasoning over long-horizon spatio-temporal memory for robot navigation. In: 2025 IEEE International Conference on Robotics and Automation (ICRA). pp. 2838–2845 (2025). https://doi.org/10.1109/ICRA55743.2025.11127706
3. Assran, M., Bardes, A., Fan, D., Garrido, Q., Howes, R., Komeili, M., Muckley, M., Rizvi, A., Roberts, C., Sinha, K., Zholus, A., Arnaud, S., Gejji, A., Martin, A., Hogan, F.R., Dugas, D., Bojanowski, P., Khalidov, V., Labatut, P., Massa, F., Szafraniec, M., Krishnakumar, K., Li, Y., Ma, X., Chandar, S., Meier, F., LeCun, Y., Rabbat, M., Ballas, N.: V-jepa 2: Self-supervised video models enable understanding, prediction and planning. arXiv preprint arXiv:2506.09985 (2025)

4. Bärmann, L., Waibel, A.: 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 Workshops (CVPRW). pp. 1560–1568 (2022). https://doi.org/10.1109/CVPRW56347.2022.00162
5. Bolaños, M., Mestre, R., Talavera, E., Giró-i Nieto, X., Radeva, P.: Visual summary of egocentric photostreams by representative keyframes. In: 2015 IEEE International Conference on Multimedia and Expo Workshops (ICMEW). pp. 1–6 (2015). https://doi.org/10.1109/ICMEW.2015.7169863
6. Bradski, G.: The opencv library. Dr. Dobb's Journal of Software Tools (2000)
7. Chhikara, P., Khant, D., Aryan, S., Singh, T., Yadav, D.: Mem0: Building production-ready ai agents with scalable long-term memory (2025). https://doi.org/10.48550/arXiv.2504.19413
8. Datta, S., Dharur, S., Cartillier, V., Desai, R., Khanna, M., Batra, D., Parikh, D.: Episodic memory question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19119–19128 (2022)
9. Engel, J., Somasundaram, K., Goesele, M., Sun, A., Gamino, A., Turner, A., Talattof, A., Yuan, A., Souti, B., Meredith, B., Peng, C., Sweeney, C., Wilson, C., Barnes, D., DeTone, D., Caruso, D., Valleroy, D., Ginjupalli, D., Frost, D., Miller, E., Mueggler, E., Oleinik, E., Zhang, F., Somasundaram, G., Solaira, G., Lanaras, H., Howard-Jenkins, H., Tang, H., Kim, H.J., Rivera, J., Luo, J., Dong, J., Straub, J., Bailey, K., Eckenhoff, K., Ma, L., Pesqueira, L., Schwesinger, M., Monge, M., Yang, N., Charron, N., Raina, N., Parkhi, O., Borschowa, P., Moulon, P., Gupta, P., Mur-Artal, R., Pennington, R., Kulkarni, S., Miglani, S., Gondi, S., Solanki, S., Diener, S., Cheng, S., Green, S., Saarinen, S., Patra, S., Mourikis, T., Whelan, T., Singh, T., Balntas, V., Baiyya, V., Dreewes, W., Pan, X., Lou, Y., Zhao, Y., Mansour, Y., Zou, Y., Lv, Z., Wang, Z., Yan, M., Ren, C., Nardi, R.D., Newcombe, R.: Project aria: A new tool for egocentric multi-modal ai research (2023)
10. Farnebäck, G.: Two-frame motion estimation based on polynomial expansion. In: Proceedings of the 13th Scandinavian Conference on Image Analysis. Lecture Notes in Computer Science, vol. 2749, pp. 363–370. Springer (2003). https://doi.org/10.1007/3-540-45103-X_50
11. Friston, K.: A theory of cortical responses. Philosophical Transactions of the Royal Society B: Biological Sciences **360**(1456), 815–836 (2005). https://doi.org/10.1098/rstb.2005.1622
12. Gergely, G., Csibra, G.: Teleological reasoning in infancy: The naïve theory of rational action. Trends in Cognitive Sciences **7**(7), 287–292 (2003). https://doi.org/10.1016/S1364-6613(03)00128-1
13. Goletto, G., Nagarajan, T., Averta, G., Damen, D.: AMEGO: Active memory from long egocentric videos. In: Computer Vision–ECCV 2024. Lecture Notes in Computer Science, vol. 15071, pp. 92–110. Springer (2024). https://doi.org/10.1007/978-3-031-72624-8_6
14. Gorlo, N., Schmid, L., Carlone, L.: Describe anything anywhere at any moment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 35002–35013 (2026)
15. Gorlo, N., Wise, D.K., Speranzon, A., Carlone, L.: Worth remembering: Surprise-gated robot episodic memory (2026). https://doi.org/10.48550/arXiv.2606.03787
16. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., Martin, M., Nagarajan, T., Radosavovic, I.,

Ramakrishnan, S.K., Ryan, F., Sharma, J., Wray, M., Xu, M., Xu, E.Z., Zhao, C., Bansal, S., Batra, D., Cartillier, V., Crane, S., Do, T., Doulaty, M., Erapalli, A., Feichtenhofer, C., Fragomeni, A., Fu, Q., Gebreselasie, A., González, C., Hillis, J., Huang, X., Huang, Y., Jia, W., Khoo, W., Kolář, J., Kottur, S., Kumar, A., Landini, F., Li, C., Li, Y., Li, Z., Mangalam, K., Modhugu, R., Munro, J., Murrell, T., Nishiyasu, T., Price, W., Ruiz, P., Ramazanova, M., Sari, L., Somasundaram, K., Southerland, A., Sugano, Y., Tao, R., Vo, M., Wang, Y., Wu, X., Yagi, T., Zhao, Z., Zhu, Y., Arbeláez, P., Crandall, D., Damen, D., Farinella, G.M., Fuegen, C., Ghanem, B., Ithapu, V.K., Jawahar, C.V., Joo, H., Kitani, K., Li, H., Newcombe, R., Oliva, A., Park, H.S., Rehg, J.M., Sato, Y., Shi, J., Shou, M.Z., Torralba, A., Torresani, L., Yan, M., Malik, J.: Ego4d: Around the world in 3,000 hours of egocentric video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18995–19012 (June 2022)

17. Gu, Q., Kuwajerwala, A., Morin, S., Jatavallabhula, K.M., Sen, B., Agarwal, A., Rivera, C., Paul, W., Ellis, K., Chellappa, R., Gan, C., de Melo, C.M., Tenenbaum, J.B., Torralba, A., Shkurti, F., Paull, L.: ConceptGraphs: Open-vocabulary 3d scene graphs for perception and planning. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 5021–5028 (2024). https://doi.org/10.1109/ICRA57147.2024.10610243
18. Healey, J., Picard, R.W.: StartleCam: A cybernetic wearable camera. In: Proceedings of the Second International Symposium on Wearable Computers (ISWC). pp. 42–49. IEEE Computer Society (1998). https://doi.org/10.1109/ISWC.1998.729528
19. Hodges, S., Williams, L., Berry, E., Izadi, S., Srinivasan, J., Butler, A., Smyth, G., Kapur, N., Wood, K.R.: SenseCam: A retrospective memory aid. In: UbiComp 2006: Ubiquitous Computing. Lecture Notes in Computer Science, vol. 4206, pp. 177–193. Springer (2006). https://doi.org/10.1007/11853565_11
20. Itti, L., Baldi, P.: Bayesian surprise attracts human attention. Vision Research **49**(10), 1295–1306 (2009). https://doi.org/10.1016/j.visres.2008.09.007
21. Kaifosh, P., Reardon, T., et al.: A generic non-invasive neuromotor interface for human–computer interaction. Nature **645**, 702–711 (2025). https://doi.org/10.1038/s41586-025-09255-w
22. Kumar, M., Goldstein, A., Michelmann, S., Zacks, J.M., Hasson, U., Norman, K.A.: Bayesian surprise predicts human event segmentation in story listening. Cognitive Science **47**(10), e13343 (2023). https://doi.org/10.1111/cogs.13343
23. Kundu, K., Shrivastava, R., Arap, M., Wang, N., Zhu, X., Fettes, Q., Tiwari, G., Suresh, P., Moutakanni, T., Munoz, A.C., Bolourchi, A., Fung, P., Donmez, P., Damavandi, B., Kumar, A., Moon, S.: Plan, watch, recover: A benchmark and architectures for proactive procedural assistance (2026), https://arxiv.org/abs/2606.04970
24. Lando, G., Forte, R., Farinella, G.M., Furnari, A.: How far can off-the-shelf multimodal large language models go in online episodic memory question answering? In: Image Analysis and Processing – ICIAP 2025. Lecture Notes in Computer Science, vol. 16168, pp. 499–511. Springer (2026). https://doi.org/10.1007/978-3-032-10192-1_42, https://doi.org/10.1007/978-3-032-10192-1_42
25. Lee, Y.J., Ghosh, J., Grauman, K.: Discovering important people and objects for egocentric video summarization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1346–1353. IEEE (2012). https://doi.org/10.1109/CVPR.2012.6247820

26. Lee, Y.J., Grauman, K.: Predicting important objects for egocentric video summarization. International Journal of Computer Vision **114**(1), 38–55 (2015). https://doi.org/10.1007/s11263-014-0794-5
27. Lin, Y.L., Morariu, V.I., Hsu, W.: Summarizing while recording: Context-based highlight detection for egocentric videos. In: Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW). pp. 443–451 (2015). https://doi.org/10.1109/ICCVW.2015.65
28. Lu, Z., Grauman, K.: Story-driven summarization for egocentric video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2714–2721 (2013). https://doi.org/10.1109/CVPR.2013.350
29. Meta: Meta Ray-Ban Display: AI Glasses With an EMG Wristband (Sep 2025), https://about.fb.com/news/2025/09/meta-ray-ban-display-ai-glasses-emg-wristband/
30. Packer, C., Wooders, S., Lin, K., Fang, V., Patil, S.G., Stoica, I., Gonzalez, J.E.: MemGPT: Towards llms as operating systems. arXiv preprint arXiv:2310.08560 (2023). https://doi.org/10.48550/arXiv.2310.08560
31. Park, J.S., O'Brien, J.C., Cai, C.J., Morris, M.R., Liang, P., Bernstein, M.S.: Generative agents: Interactive simulacra of human behavior. In: Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST). pp. 1–22 (2023). https://doi.org/10.1145/3586183.3606763
32. Plewnia, J., Peller-Konrad, F., Asfour, T.: Forgetting in robotic episodic long-term memory. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 6711–6717. IEEE (2024). https://doi.org/10.1109/ICRA57147.2024.10610299
33. Poleg, Y., Halperin, T., Arora, C., Peleg, S.: EgoSampling: Fast-forward and stereo for egocentric videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4768–4776 (2015)
34. Pramanick, S., Song, Y., Nag, S., Lin, K.Q., Shah, H., Shou, M.Z., Chellappa, R., Zhang, P.: EgoVLPv2: Egocentric video-language pre-training with fusion in the backbone. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5285–5297 (2023)
35. Ramakrishnan, S.K., Al-Halah, Z., Grauman, K.: NaQ: Leveraging narrations as queries to supervise episodic memory. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6694–6703 (2023)
36. Sharghi, A., Gong, B., Shah, M.: Query-focused extractive video summarization. In: Computer Vision–ECCV 2016. Lecture Notes in Computer Science, vol. 9912, pp. 3–19. Springer (2016). https://doi.org/10.1007/978-3-319-46484-8_1
37. Sharghi, A., Laurel, J.S., Gong, B.: Query-focused video summarization: Dataset, evaluation, and a memory network based approach. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4788–4797 (2017). https://doi.org/10.1109/CVPR.2017.229
38. Song, Y., Xin, Q.: D-MEM: Dopamine-gated agentic memory via reward prediction error routing (2026). https://doi.org/10.48550/arXiv.2603.14597
39. Su, Y.C., Grauman, K.: Detecting engagement in egocentric video. In: Computer Vision – ECCV 2016. Lecture Notes in Computer Science, vol. 9909, pp. 454–471. Springer International Publishing (2016). https://doi.org/10.1007/978-3-319-46454-1_28
40. Subramaniam, V., Subbaraju, V., Roy, D., Krishna, P., Kandappu, T., Xu, Q.: Detecting social engagement of elderly from lifelog image-streams to identify effective cues for autobiographic recall. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 3380–3389 (2026)

41. Sun, S., Han, K., Huang, Y., Cai, W., Song, J.: EgoGraph: Temporal knowledge graph for egocentric video understanding (2026). https://doi.org/10.48550/arXiv.2602.23709
42. Tang, Y., Zhang, J., Qin, X., Yu, J., Qiu, M., Gou, G., Xiong, G., Lin, Q., Rajmohan, S., Zhang, D., Wu, Q.: EgoMemory: Memory-augmented personalized retrieval for long-context egocentric video. In: Findings of the Association for Computational Linguistics: ACL 2026. pp. 7317–7349. Association for Computational Linguistics, San Diego, California, United States (Jul 2026). https://doi.org/10.18653/v1/2026.findings-acl.362, https://aclanthology.org/2026.findings-acl.362/
43. Tong, Z., Song, Y., Wang, J., Wang, L.: Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. Advances in Neural Information Processing Systems **35**, 10078–10093 (2022)
44. Tran, T.H., Arap, M., Moon, S., Hamid, R., Suglia, A., Kira, Z., Fung, P., Shah, M.: Wearable ai workshop at eccv 2026. https://wearable-ai-workshop.github.io/ (2026), workshop at the European Conference on Computer Vision (ECCV) 2026
45. Wang, Y., Xu, Y., Li, X., Hong, J., Wang, Y., Chen, C.W., Zhu, W.: Egoself: From memory to personalized egocentric assistant. arXiv preprint arXiv:2604.19564 (2026)
46. Wang, Y., Yang, Y., Ren, M.: LifelongMemory: Leveraging LLMs for answering queries in long-form egocentric videos. arXiv preprint arXiv:2312.05269 (2023). https://doi.org/10.48550/arXiv.2312.05269, https://arxiv.org/abs/2312.05269
47. Xu, J., Mukherjee, L., Li, Y., Warner, J., Rehg, J.M., Singh, V.: Gaze-enabled egocentric video summarization via constrained submodular maximization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2235–2244. IEEE (2015). https://doi.org/10.1109/CVPR.2015.7298836, https://openaccess.thecvf.com/content_cvpr_2015/html/Xu_Gaze-Enabled_Egocentric_Video_2015_CVPR_paper.html
48. Xu, W., Liang, Z., Mei, K., Gao, H., Tan, J., Zhang, Y.: A-MEM: Agentic memory for llm agents. In: Advances in Neural Information Processing Systems (2025)
49. Yan, S., Yang, X., Huang, Z., Nie, E., Ding, Z., Li, Z., Ma, X., Schütze, H., Tresp, V., Ma, Y.: Memory-R1: Enhancing large language model agents to manage and utilize memories via reinforcement learning (2025). https://doi.org/10.48550/arXiv.2508.19828
50. Zacks, J.M., Speer, N.K., Swallow, K.M., Braver, T.S., Reynolds, J.R.: Event perception: A mind–brain perspective. Psychological Bulletin **133**(2), 273–293 (2007). https://doi.org/10.1037/0033-2909.133.2.273
51. Zhang, G., Jiang, W., Wang, X., Behr, A., Zhao, K., Friedman, J., Chu, X., Anoun, A.: Adaptive memory admission control for llm agents (2026). https://doi.org/10.48550/arXiv.2603.04549
52. Zhong, W., Guo, L., Gao, Q., Ye, H., Wang, Y.: MemoryBank: Enhancing large language models with long-term memory. Proceedings of the AAAI Conference on Artificial Intelligence **38**(17), 19724–19731 (2024). https://doi.org/10.1609/aaai.v38i17.29946