UNIVERSITY OF COPENHAGEN

DEPARTMENT OF NORDIC STUDIES AND LINGUISTICS

Decoding Cognitive Events: Predicting Event Boundaries with Eye Tracking and Machine Learning Models

Argyri Morfakidou (xmp875)
Miłosz Holeksa (tmk107)
Weronika Kopytko (mgr983)

Computational Cognitive Science 3

MSc in IT & Cognition
Autumn 2024

Division of work

The authors attest that they contributed equally to preparing the project, accordingly to information presented in the table below.

Section	**Author**
Abstract	Argyri, Weronika, Milosz
1. Introduction	Argyri, Weronika, Milosz
1.2 and 1.3 Literature review	Argyri, Weronika, Milosz
2.1 Data	Weronika, Milosz
2.2 Baseline HMM Model	Weronika, Milosz
2.3 Main LSTM Model	Weronika, Milosz
2.4 Pupil Diameter Exploration	Argyri
3. Results	Argyri, Weronika, Milosz
4. Discussion	Argyri, Weronika, Milosz
6. Conclusions	Argyri
7. Limitations and Future Work	Argyri

Code & Data Availability Statement

The source code used in this study is available publicly at this GitHub repository The data can be accessed through the following link: OSF.

Abstract

This study is grounded in Event Segmentation Theory (EST) which posits that when processing information, people inherently segment their conscious experience into distinct events. Scientific literature stresses out a crucial role that event segmentation plays in bridging perception and memory, by making the observation that additional cognitive processing is involved around perceived event boundaries, which leads to better memory encoding. We used a dataset containing information gathered from a study where participants watched a video with pre-determined event boundaries. After watching the video participants' memory regarding the events was measured using two metrics. Eye tracking data of the gaze position and pupil size were gathered. We built upon the findings of the study suggesting that event boundaries lead to distinct gaze patterns and (a) developed a Hidden Markov Model (HMM) and a Long Short-Term Memory model (LSTM) to predict the occurrence of event boundaries based on the eye tracking data, (b) attempted to associate cognitive load and memory performance with the pupillometry data during the processing of the events and (c) validate our models with internally collected data. The results of our study showed better performance of the LSTM model compared to the HMM architecture used in previous research. These results were partially confirmed by our internal validation experiment, where the LSTM model had a higher accuracy than HMM, but both models performed worse than on the external data.

1. Introduction

Event Segmentation Theory (EST), initially articulated by Zacks and Tversky (2001), proposes that individuals naturally and spontaneously divide continuous experiences into discrete events during naturalistic information processing. This segmentation process is critical to making sense of the complex stream of sensory information that we experience every day. The theory suggests that event boundaries - specific points where individuals identify the beginning or end of an event - are associated with increased cognitive processing, which helps facilitate the encoding of these boundaries into long-term memory. Event segmentation acts as a bridge between perception and memory, whereby segmenting experiences into meaningful units allows individuals to better organize, encode, and later retrieve these experiences (Zacks et al., 2007).

Recent studies have confirmed that event boundaries are more cognitively salient than other parts of an experience, often requiring additional cognitive load, as evidenced by physiological markers such as pupil dilation and synchronized gaze patterns (Smith et al., 2024b). Although traditional neuroimaging techniques such as fMRI and EEG have been extensively used to investigate the neural foundations of event segmentation (Ben-Yakov and Henson, 2021; Silva et al., 2019), eye tracking has emerged as a novel tool that offers unique insights into how individuals visually parse and encode event boundaries. Changes in gaze behavior, pupil dilation, and synchrony among viewers are important indicators of how the human brain processes event boundaries.

1.2. Literature review

The concept of event segmentation has its roots in understanding how physiological metrics can reflect the mental workload during task performance. Kahneman and Beatty (1966) were among the first to propose that physiological responses, such as pupil dilation, can be indicators of cognitive effort. They stated, that during a short-term memory task, pupil diameter might be a measure of the amount of material which is under active processing at given time. In the following decade, Newtson (1973) conducted pivotal research on event perception, which established that people naturally segment continuous actions into meaningful units, a phenomenon he referred to as *unitization* of perception. Newtson's work revealed that participants who

segmented observed activities into smaller units had better memory recall of those activities, emphasizing the importance of segmentation in memory processes. Zacks and Tversky (2001) formalized EST, proposing that humans naturally break continuous experiences into meaningful units called *events*, with boundaries marking transitions that require heightened cognitive resources for effective memory encoding. Research into cognitive aging, such as (Smith et al., 2024a), revealed that while segmentation improved memory recall in young and middle-aged adults, it was ineffective in older adults.

Subsequent studies provided neural evidence for EST. Zacks et al. (2007) used fMRI to show increased activation in the frontal and temporal cortices at event boundaries during video viewing and Silva et al. (2019) used EEG along with a Hidden Markov Model (HMM) to demonstrate how the memory systems segment a continuous long stream of experience into episodic events by observing a left-lateralized anterior negative ERP effect.

Expanding these findings, Li et al. (2024) employed a novel approach with high-resolution eye-tracking and HMM to show that gaze patterns become more distinct at event boundaries, correlating with improved memory recall. This underscores the role of event boundaries as cognitive "anchors" that trigger additional processing to enhance memory encoding and retrieval, integrating physiological, behavioral insights and machine learning into a cohesive framework for researching and understanding event segmentation.

1.3. Literature review: Exploration of the Pupil Diameter measurement

Pupil dilation refers to the increase in pupil size triggered by a stimulus, measured relative to the baseline diameter observed before the stimulus was presented. According to the literature pupil diameter measurements can serve as a metric for several physiological and psychological factors. A strong link has been observed between the diameter of the pupil and the cognitive effort. In particular, pupil dilation increases in proportion to the demands of the task being performed (Hess and Polt, 1964). The pupillary response evoked by the task-related cognitive load is described as task-evoked pupillary response (TEPR) (Hess and Polt, 1964). Szulewski et al. (2015) used the TEPR metric to measure the mental effort associated with either trained physicians or novices answering clinical questions. TEPRs in novices were greater than those in trained physicians suggesting that pupillometry provides objective

measures of mental effort. In another study, Mitre-Hernandez et al. (2020) successfully used pupillometry to measure the cognitive load related to different difficulty levels in a video game.

Pupil diameter was also found to be a biomarker of memory performance. In a study conducted by Kucewicz et al. (2018) words that were successfully remembered exhibited notably different pupil responses during their encoding phase compared to those that were not retained. Similarly, in a short-term memory task pupil size was found to exhibit an increase proportional to the amount of information retained (Kahneman and Beatty, 1966). In another study, pupil diameter was found to be associated with accuracy of memory representations (Starc et al., 2017).

Drawing from the aforementioned literature we set out to explore whether we could distinguish between experimental conditions based on the pupil diameter measurements. Given that pupil diameter could be indicative of cognitive load and its relation to memory we also expected pupil diameter to serve as a mediatory variable between the distinct experimental groups and the memory scores.

Overall, our research adds to the findings of the study on event boundaries leading to distinct gaze patterns (Li et al., 2024) in multiple ways:

1. Our approach focuses on building machine learning (ML) models that predict the occurrence of event boundaries based on both pupil diameter and gaze speed. Instead of solely relying on the commonly used HMM architecture for similar tasks, we also experimented with a Long Short-Term Memory model (LSTM) to explore whether it could offer a better approach for this type of prediction.

2. We further explored the pupillometry data collected during the processing of events, building upon the literature that suggests a link between cognitive load and pupil dilation.

3. Finally, we attempted to validate our findings by employing our model with internally gathered data.

2. Methodology

2.1. Data

In our analysis we reuse readily available data, collected in a pre-printed study conducted by Li et al. (2024). The data is publicly available at OSF. The data consists of gaze location (x and y) and pupil size collected during a video watching experiment conducted on 130 participants. It was collected at a frequency of 1000 Hz, however we downsample it to 10 Hz, following the approach of the authors of the paper we use as a source. The data was collected for around 21,5 minutes while participants were watching a part of BBC's "Sherlock: A Study in Pink". Prior to viewing the film participants were exposed to four different types of online content. The x and y coordinates were used to calculate gaze speed.

Since we wanted to measure the occurrence of the event boundaries we formulated the task as a binary classification problem. This meant that we needed to split the data into fragments containing event boundaries and fragments not containing event boundaries. We did that by sampling an equal number of fixed-length (100 time points) time series, divided into two groups: containing and not containing event boundaries. We applied this procedure the same number of times for each participant, which gave us a dataset of 21000 labeled samples. We then standardized the data by subtracting the mean and dividing by the standard deviation.

2.2. Baseline HMM Model

Hidden Markov Models (HMMs) have been widely used for tasks involving temporal data due to their ability to model sequences of observations and transitions between hidden states. In this study, we implemented an HMM as a baseline model for predicting the occurrence of event boundaries. The HMM was trained on gaze speed and pupil size data, assuming that the event boundary transitions could be captured as latent states. By leveraging the probabilistic framework of HMMs, the model assigned likelihood scores to sequences, enabling the detection of event boundaries based on observed patterns. While effective, HMMs rely on predefined assumptions about state transitions and emissions, which may limit their adaptability to complex, non-linear patterns in the data.

The Hidden Markov Model (HMM) used in this study consisted of three

hidden states, which controlled the model's complexity and allowed it to capture the temporal dependencies in the sequential data. Each class (containing vs not containing an event boundary) was trained with its own HMM, where the model parameters - transition probabilities between states, emission probabilities modeled by a Gaussian distribution, and initial state probabilities - were estimated using the Expectation-Maximization (EM) algorithm. The number of iterations for training was set to 1000. During prediction, the log-likelihood of each test sequence was calculated for both class-specific models, and the sequence was classified into the class with the higher log-likelihood. The model's performance was assessed based on classification accuracy. The model was developed using a dedicated package: "HMMLearn".

2.3 Main LSTM Model

To explore a more flexible and potentially superior approach, we implemented a Long Short-Term Memory model (LSTM) model. LSTMs are a type particularly suited for sequential data as they can capture temporal dependencies through their ability to retain information from previous time steps. Our LSTM was designed to process the downsampled gaze speed and pupil data, learning intricate temporal patterns associated with event boundaries. Unlike HMMs, LSTMs do not require explicit assumptions about state transitions, making them better equipped to handle the non-linear and dynamic nature of the task. The model's architecture allowed for a more data-driven approach to predicting event boundaries, providing an opportunity to compare its performance against the baseline HMM.

We developed an LSTM Model consisting of two parallel long short-term memory layers, one processing the gaze speed and the other processing the pupil size. Each LSTM layer had an input size of 1, a hidden size of 64 units, and consisted of 2 layers, which allowed the model to capture complex temporal dependencies in the data. Both LSTMs outputted the hidden state from the last time step, which captured the temporal information from the entire sequence. These final hidden states were concatenated and passed through a fully connected layer, which had 128 input features (the combined hidden states from both channels). The output of the fully connected layer was passed through a sigmoid activation function, producing a binary classification output. The model was developed using PyTorch, trained on 100 epochs and batches of size 64.

2.4 Pupil Diameter Exploration

Additionally, we employed a mixed-method approach to investigate the effects of different video viewing conditions on pupil dilation and subsequent memory performance. Participants were divided into four experimental groups according to what type of content they viewed before their exposure to "Sherlock: A Study in Pink" movie clip: Short-R, Short-P, Long, and Schema. The Short-R group viewed random TikTok videos, while the Short-P group watched personalized TikTok videos. In contrast, the Long group was exposed to a movie-unrelated documentary, and the Schema group viewed a movie-related clip. During the subsequent viewing period, pupillometry data were collected at multiple time points, signifying event segmentation points, and memory performance was assessed via two distinct tasks administered afterwards: a recall task (Memory1 score) and a detail-oriented task (Memory2 score).

To explore the pupillometry data, KMeans clustering was applied to standardized pupil diameter values across the specified time points. Clustering participants into four groups provided insights into shared patterns of pupillary responses. The silhouette score was calculated to assess the clustering quality, and these clusters were compared with the original experimental groups in order to explore whether there is alignment between physiological responses and the pre-defined conditions.

Subsequently, we applied inferential statistics, including Analysis of Variance (ANOVA) as a means to investigate whether differences in pupil dilation existed across the experimental groups. In order to assess if the differences between pairs of group means were statistically significant we employed the post-hoc Tukey HSD test. Descriptive statistics were then calculated to explore the variability in physiological performance across the experimental groups. These analyses were the base for examining the connection between physiological responses during video viewing and subsequent memory outcomes.

Regression analyses were then conducted to predict memory performance based on pupil dilation metrics and group conditions. Separate models were trained for Memory1 and Memory2 scores, using both pupil size data and experimental group information as predictors.

2.5 Internal Validation

To validate our models, we tested them using internally collected eye tracking data gathered from eight participants aged 24–25, all of whom were familiar with viewing TikTok videos and had no prior exposure to the selected movie clip (Sherlock, Season 1, Episode 1: A Study in Pink). Participants provided informed consent for recording and anonymously using their eye tracking data. The data collection was conducted using a Tobii Pro eye tracker, which operates at a sampling rate of 1000 Hz and includes a 17-inch integrated LCD monitor for stimulus presentation. It utilizes two binocular infrared cameras to capture various eye movements with an average accuracy of 0.5 degrees. Data was documented using Tobii Studio software. The metrics selected were gaze position used for calculating the gaze speed, and pupil size. Participants viewed personalized content for five minutes and were subsequently exposed to the movie clip. Eye tracking data were only collected during the movie exposure. The data was processed similarly as the data from the previous study. We calculated the gaze speed based on the x and y coordinates. We sampled time sequences both containing and not containing event boundaries and standardized the data. However, we faced an issue of large number of missing data (10%-50% depending on participant). We approached this by: (a) excluding data of participant with highest amount of missing data (50%) (b) sampling only sequences which had less than 10% of missing data (c) performing linear imputation of these missing fragments.

3. Results

3.1. Machine Learning Models' Results

After training our models we got the results shown in Table 1.

Model	**Accuracy**
LSTM on external dataset	99%
HMM on external dataset	62%
LSTM on internal dataset	59%
HMM on internal dataset	52%

Table 1: Accuracy results for machine learning models on test data. Where: LSTM - Long Short-Term Memory model, HMM - Hidden Makrkov Model.

3.2. Results of the exploratory analysis of pupil dilation metrics

The analysis revealed notable differences in pupillary responses and memory performance across the experimental groups. KMeans clustering of pupil dilation data resulted in four distinct clusters, suggesting variability in the pupil dilation metric during video viewing. The clustering patterns showed that participants in the Short-P and Schema groups were more likely to exhibit similar pupillary responses compared to the Short-R and Long groups. The Silhouette score for the clustering was 0.527 suggesting that the clusters captured distinct differences in pupillary response patterns across the groups. From the computing of the descriptive statistics emerged distinct differences in mean pupil dilation values across the experimental groups. The Short-R group exhibited the lowest mean pupil dilation, with an average value of 1608.047 pixels^2, significantly lower than the Short-P group (M=2089.205) and the Schema group (M=2383.119). The Long group recorded a mean dilation of 2406.965, a comparable level of pupil dilation to the Schema group. These results, depicted in Figure 1 highlight the reduced physiological engagement in the condition including viewing of random TikTok videos.

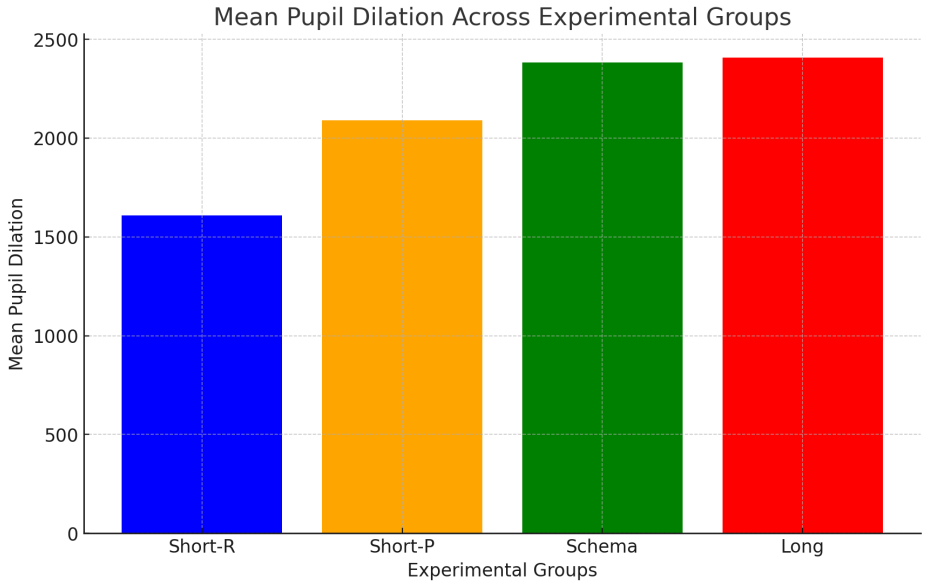

Figure 1: Visual Representation of Mean Differences in Pupil Dilation

ANOVA results confirmed that the differences in mean pupil dilation across groups were statistically significant ($F=60.326$, $p<0.001$). Tukey HSD post-hoc tests revealed significant differences between the Short-R group and all other groups: Short-P ($p<0.001$), Schema ($p<0.001$), and Long ($p<0.001$). Similarly, the Short-P group significantly differed from both the Schema ($p=0.0003$) and Long ($p<0.001$) groups. However, the difference between the Schema and Long groups was not statistically significant ($p=0.968$).

According to the results of the regression analyses performed pupil dilation metrics are only weakly predictive of memory scores. Even within the Short-R group, no significant relationship was found between dilation and memory performance. Specifically, within the Short-R group, the mean squared error (MSE) for Memory1 (recall) was 0.033, and for Memory2 (detail), it was 0.359. Similarly, for the Schema group, the MSE for Memory1 was 0.028, reflecting slightly better predictive performance, but Memory2 scores showed poorer predictive power, since the MSE was 0.519. The Short-P group demonstrated modest prediction accuracy for Memory1, with an MSE of 0.030, but Memory2 performance was weakly predicted, with an MSE of 0.658. The Long group exhibited the lowest prediction accuracy for Memory1 at 0.060 while Memory2 scores showed an MSE of 0.274 (Figure 2).

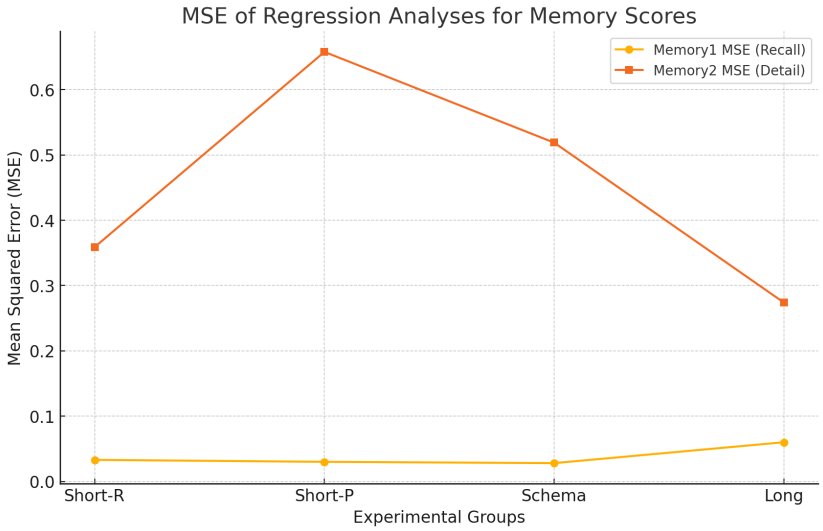

Figure 2: Visual Representation of Regression Analysis Results

4. Discussion

Machine Learning Models

Our results highlight the potential of machine learning models in leveraging eye tracking data for predicting the occurrence of event boundaries, as proposed by Event Segmentation Theory (EST). The Hidden Markov Model (HMM), a commonly used approach in previous studies, demonstrated moderate success in identifying event boundaries with an accuracy of 62% on the external dataset. However, its performance was constrained by the reliance on predefined assumptions about state transitions and emissions, which are inherently limited in their ability to capture complex, non-linear patterns in sequential data.

In contrast, the Long Short-Term Memory (LSTM) model achieved remarkable accuracy, surpassing 99% on the external dataset while not overfitting to the training data. This result underscores the strength of deep learning models in capturing temporal dependencies and dynamic patterns within eye tracking data. Unlike HMMs, LSTMs do not require explicit assumptions about the structure of the data, enabling a more flexible and data-driven approach. The dual-channel architecture of the LSTM, which separately processed gaze speed and pupil size, allowed the model to learn intricate relationships between these features and event boundaries.

When applied to internally collected data, both models showed performance differences, with LSTM maintaining higher accuracy compared to HMM. However, the obtained accuracies were significantly lower than for the external data. We wanted to further investigate the discrepancy between the accuracies by testing our LSTM model on each participant's data separately. We noticed that calculated accuracies were ranging from 50% up to 72% and depended on the amount of missing data in sampled sequences, with highest accuracy obtained for the participant with the lowest amount of missing data (10%). We conclude that relatively low accuracy of the model tested on our internal data was likely due to the faulty data collection process that the linear imputation of data was not sufficient to account for. Further data collection is required to confirm the model's performance.

Pupilometry Data Analysis

The main observation derived from the pupillometry related exploration that we conducted was that the type of content viewed significantly influences

pupil dilation. The larger pupil dilation observed in the Short-P and Schema groups likely reflects greater mental effort and engagement. Given the distinct pupil dilation scores per experimental condition and the literature pointing to the influence of the specific metric on cognitive load and memory performance (van der Wel and van Steenbergen, 2018; Kucewicz et al., 2018) we expected pupil dilation measurements to be predictive of the memory scores. However, from our analyses a weak relationship between pupil size and memory performance emerged within each experimental condition. This lack of statistical power could be attributed to the relatively small number of observations. Interestingly, there was a weaker association between pupil diameter and Memory 2 compared to Memory1 suggesting that the metric may be more closely tied to general recall (Memory1) rather than the encoding of fine details.

5. Conclusions

Our findings demonstrate the feasibility of using machine learning models to analyze eye tracking data for the prediction of cognitive events, presenting an efficient alternative to resource-intensive methods such as fMRI or EEG. The high accuracy achieved by the LSTM model emphasizes the potential for deep learning approaches in cognitive science research. In contrast to what was anticipated, we did not find a direct association between pupil diameter and cognitive load. However, we observed distinctly different pupil diameter patterns linked to different content viewing conditions. This observation could be further explored in a future study.

6. Limitations & Future Work

The main limitation of this study was the small sample size, especially in the validation experiment, which reduces the generalizability of our findings. Future studies could include larger and more diverse participant pools to improve statistical power and ensure the generalizability of the outcomes. Additionally, in future work it would be interesting to investigate the temporal precision of event boundary predictions, moving beyond binary classification to predicting the exact timing of event transitions. Furthermore, addressing the issue of missing data is crucial, as our analyses revealed that the amount

of missing data significantly impacted model performance. Future studies should employ more robust data collection protocols and explore advanced imputation methods to minimize the impact of missing data. Finally, the weak predictive relationship observed between pupil dilation and memory performance in this study could be addressed by exploring alternative - possibly multimodal - physiological markers alongside pupil dilation. Integrating additional features could better capture the nuances of cognitive load.

Bibliography

Ben-Yakov, A. and Henson, R. N. (2021). The hippocampal film editor: Sensitivity and specificity to event boundaries in continuous experience. *Journal of Neuroscience*, 41(8):1845–1847.

Hess, E. H. and Polt, J. M. (1964). Pupil size in relation to mental activity during simple problem-solving. *Science*, 143:1190–1192.

Kahneman, D. and Beatty, J. (1966). Pupil diameter and load on memory. *Science*, 154:1583–1585.

Kucewicz, M. T., Dolezal, J., Kremen, V., Berry, B. M., Miller, L. R., Magee, A. L., Fabian, V., and Worrell, G. A. (2018). Pupil size reflects successful encoding and recall of memory in humans. *Scientific Reports*, 8.

Li, J., Cheng, Z., Hao, X., and Liu, W. (2024). Boundaries in the eyes: measure event segmentation during naturalistic video watching using eye tracking. *bioRxiv (Cold Spring Harbor Laboratory)*.

Mitre-Hernandez, H., Covarrubias-Carrillo, R., and Lara-Alvarez, C. (2020). Pupillary responses for cognitive load measurement: Classifying difficulty levels in an educational video game (preprint). *JMIR Serious Games*.

Newtson, D. (1973). Attribution and the unit of perception of ongoing behavior. *Journal of Personality and Social Psychology*, 28:28–38.

Silva, M., Baldassano, C., and Fuentemilla, L. (2019). Rapid memory reactivation at movie event boundaries promotes episodic encoding. *Journal of Neuroscience*, 39(43):8538–8548.

Smith, M. E., Hall, C. S., Membreno, R., Quintero, D., and Zacks, J. M. (2024a). Attention to event segmentation improves memory in young adults: A lifespan study. *Psychology and Aging*, 39(7):750–769.

Smith, M. E., Loschky, L. C., and Bailey, H. R. (2024b). Eye movements and event segmentation: Eye movements reveal age-related differences in event model updating. *Psychology and Aging*, 39(2):180–187.

Starc, M., Anticevic, A., and Repovš, G. (2017). Fine-grained versus categorical: Pupil size differentiates between strategies for spatial working memory performance. *Psychophysiology*, 54:724–735.

Szulewski, A., Roth, N., and Howes, D. (2015). The use of task-evoked pupillary response as an objective measure of cognitive load in novices and trained physicians. *Academic Medicine*, 90:981–987.

van der Wel, P. and van Steenbergen, H. (2018). Pupil dilation as an index of effort in cognitive control tasks: A review. *Psychonomic Bulletin Review*, 25:2005–2015.

Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., and Reynolds, J. R. (2007). Event perception: A mind-brain perspective. *Psychological Bulletin*, 133:273–293.

Zacks, J. M. and Tversky, B. (2001). Event structure in perception and conception. *Psychological Bulletin*, 127:3–21.