A neural network with episodic memory learns causal relationships between narrative events

Hayoung Song[1]*, Qihong Lu[2], Tan T. Nguyen[3], Janice Chen[4], Yuan Chang Leong[5,6,7], Monica D. Rosenberg[5,6,7], ShiNung Ching[8], Jeffrey M. Zacks[3]

[1]Center for Theoretical and Computational Neuroscience, Washington University in St Louis
[2]Zuckerman Institute, Columbia University
[3]Department of Psychological and Brain Sciences, Washington University in St. Louis
[4]Department of Psychological and Brain Sciences, Johns Hopkins University
[5]Department of Psychology, University of Chicago
[6]Neuroscience Institute, University of Chicago
[7]Institute of Mind and Biology, University of Chicago
[8]Department of Electrical and Systems Engineering, Washington University in St. Louis

*Correspondence: Hayoung Song (hayoung@wustl.edu)

Abstract
Humans reflect on past memories to make sense of an ongoing event. Past work has shown that people retrieve causally related past events during comprehension, but the exact process by which this causal inference occurs remains elusive. Here, we employed a recurrent neural network augmented with an episodic memory buffer to examine how memories are retrieved and integrated based on causal relationships between events. The model was trained to predict upcoming scenes as it watched a television episode. At every time step, the model transformed the current scene into two distinct representations—“value” representing memory content and “key” representing memory address—both of which were stored as episodic memory. The model learned to retrieve selective past values by applying self-attention over stored keys, and it integrated these retrieved values with the current scene representation to predict an upcoming scene. By separating representations used for encoding from those used for retrieval, the model learned to retrieve memories in ways that go beyond simple pattern similarity. In turn, the model represented causally related events with similar patterns beyond perceptual or semantic similarities, suggesting that it organized event representations based on latent causal structure. Memories retrieved by the model were similar to those retrieved by human participants who watched the same television episode. The model also exhibited hippocampus-like pattern separation and pattern completion, and its representational structure aligned more closely with human fMRI data than a comparison model without an episodic memory buffer. These findings suggest that the model captures the way humans represent events and retrieve memories based on causal relationships. Together, this work proposes a key-value episodic memory system as a candidate computational mechanism for how humans retrieve causally related memories to comprehend naturalistic events.

Introduction

Understanding ongoing experience is not just about making sense of what is happening now. People reason about why an event has happened by remembering events that occurred in the past. Since the 1970s, theories have described narrative comprehension as a process of constructing a coherent representation of a situation by inferring cause-and-effect relationships between events (McNamara and Magliano, 2009). Causal inference occurs throughout comprehension, as people retrieve causally related past events to understand and integrate them with the current event (Trabasso and van den Broek, 1985; Graesser et al., 1994; Langston and Trabasso, 1999).

Initial evidence for this view came from work that asked participants to describe their thought process during comprehension or recall story events afterwards. When participants were asked to think aloud while reading a story, they were more likely to mention past events that were causally related to the current event (Suh and Trabasso, 1993). Studies also found that events that are more causally connected to other events were more likely to be remembered (Trabasso and Sperry, 1985; Cohn-Sheehy et al., 2022; Lee and Chen, 2022; Antony et al., 2024). Complementing these behavioral findings, fMRI work showed that neural patterns of related past events were reinstated in the brain during narrative comprehension (Chang et al., 2021; Hahamy et al., 2023). Recent work specifically showed that *causally related* past events were reinstated in the brain, more so than semantically or perceptually similar events (Song et al., 2025).

These findings suggest that people reason about causal relationships between events through memory retrieval. But what are the computational underpinnings of this causal reasoning process? In other words, how does the human brain search through a vast store of memories to identify and retrieve causally related ones, and then integrate them with an ongoing event to update situational representations? Through this memory retrieval and integration process, how does the brain come to represent events based on their causal relationships? These questions call for formal computational models of causal inference during event comprehension.

One prevailing theory characterizes causal inference as a process of probabilistic inference (Griffiths and Tenenbaum, 2009). In this framework, people infer the causes of an event by performing Bayesian inference over hypothesis spaces generated by prior knowledge. Although this model has successfully explained human behavior in intuitive physics and early cognitive development (Griffiths et al., 2011; Xu et al., 2021; Smith et al., 2024), it has not yet been generalized to the domain of event comprehension where events and causal links are continuously updated over time—which makes it a computationally intense problem. Another line of work has used variants of recurrent neural networks (RNNs) to directly simulate the event comprehension process. Studies have shown that the model segments continuous experience into meaningful events (Franklin et al., 2020; Gumbsch et al., 2022), selectively encodes and retrieves memories at event boundaries (Lu et al., 2022), and represents event schemas that give rise to structured event representations (Elman and McRae, 2019; Lu et al., 2024b). Yet, no event comprehension model thus far has examined the causal inference process.

Here, we introduce a model to test the proposal that causal inference emerges from the continuous encoding and retrieval of episodic memories. By training the model to predict the next scene while "watching" a television episode, we examined whether the model would retrieve causally related past events to guide its predictions. Because exhaustive Bayesian inference is computationally costly, we considered another possibility that pre-clustering event representations based on their causal links may provide an efficient solution to the causal inference problem. We therefore hypothesized that, by selectively retrieving causally related memories, the model would learn to represent causally related events with similar patterns, organizing event representations according to their causal relationships.

Our model is built on a gated recurrent unit (GRU) RNN to process dynamic inputs (Cho et al., 2014). The RNN is augmented with an episodic memory (EM) buffer in which encoding and retrieval occur at every time step. This design builds on recent memory-augmented neural networks that leverage an external memory buffer to store multiple memories and later retrieve them when long-term dependencies are needed (Santoro et al., 2016; Pritzel et al., 2017; Ritter et al., 2018; Li et al., 2024; Lu et al., 2024a; Dong et al., 2025). Encoding and retrieval are controlled by a key-value memory system that utilizes a self-attention mechanism in Transformers (Vaswani et al., 2017). In this system, values represent memory contents whereas keys serve as memory addresses or indices (Gershman et al., 2025). This dissociation is analogous to a library search, where one uses an index (keys) to locate a book (value). At each time step, a query surveys the memory buffer and assigns high attention weights to keys most similar to the query, but then retrieves values that are associated with those keys. In other words, the model dissociates representations used for memory encoding (values) and retrieval (keys). We hypothesized that this dissociation would allow the system to retrieve memories based on relationships beyond linear pattern similarity, allowing it to capture higher-order causal relationships. This critically distinguishes the key-value framework from more commonly used content-addressable memory models, which retrieve memories by directly matching the current state to stored contents (Hopfield, 1982; Howard and Kahana, 2002; Polyn et al., 2009; Li et al., 2024; Lu et al., 2024a). The key-value memory system has recently been proposed as a candidate mechanism of human memory and hippocampal function (Whittington et al., 2020; Krotov and Hopfield, 2021; Chandra et al., 2025; Fang and Rajan, 2025; Gershman et al., 2025) and has shown success in various cognitive tasks when integrated with an RNN (Whittington et al., 2024; Zheng et al., 2025). To formally test its role in causal comprehension, we “lesioned” different components of the model, such as keys, queries, or the entire EM buffer, to assess their respective contributions to event representations and memory retrieval.

The model’s internal representations and memory retrieval behavior were examined in relation to human behavioral and fMRI data collected by Song et al. (2025). In this empirical study, 36 participants watched a hard-to-understand television episode that was scrambled in time. Participants were asked to press an “aha” button whenever they understood something new about the narratives and were later asked to verbally explain their insights at those moments. Participants retrieved causally related past events when explaining their insights. The study also found that the brain represents causally related events with similar neural patterns, suggesting that the brain formed a coherent representation of events based on their causal relationships. In our study, the model watched the same episode to match human experience. We asked if the model, like humans, retrieves past events that are causally related to the ongoing event. We further asked if the model represents events like the human brains do—by representing causally related events with similar representation patterns.

A recurrent neural network with a key-value episodic memory buffer performs well in the next-scene prediction task and generalizes well to an unseen episode

We employed a GRU RNN augmented with a key-value EM buffer, hereby called the EM-GRU model (**Figure 1**). At each time step t, the model receives an input vector x_t which is transformed into three distinct representations: a hidden state h_t, a key k_t, and a query q_t. The hidden state h_t is produced by a standard GRU update, based on the current input x_t and the previous hidden state h_{t-1}. This representation serves as a “value" in the episodic memory system and is stored in the EM buffer (EM[V]) at each time step. In parallel, the input x_t is linearly transformed into a key representation k_t, which serves as a memory address and is also encoded in the EM (EM[K]). To retrieve relevant past information, a query q_t surveys every stored key in EM[K]. High attention weights are given to selective past keys that are high in similarity with the query via the softmax function, following the self-attention mechanism used in Transformers (Vaswani et al., 2017). Stored values (memory contents) associated with these keys

(memory addresses) are retrieved as memory m_t, computed as a weighted sum of past values with weights determined by the attention scores over the keys. The h_t representing the current input and m_t representing the retrieved memory are integrated via weighted summation: $h_t \times \alpha + m_t \times (1 - \alpha)$, with α set to 0.5. This integrated representation is then used to predict the next scene x_{t+1}. The model was optimized for the next-scene prediction, with the goal of minimizing the loss between the predicted and observed x_{t+1}.

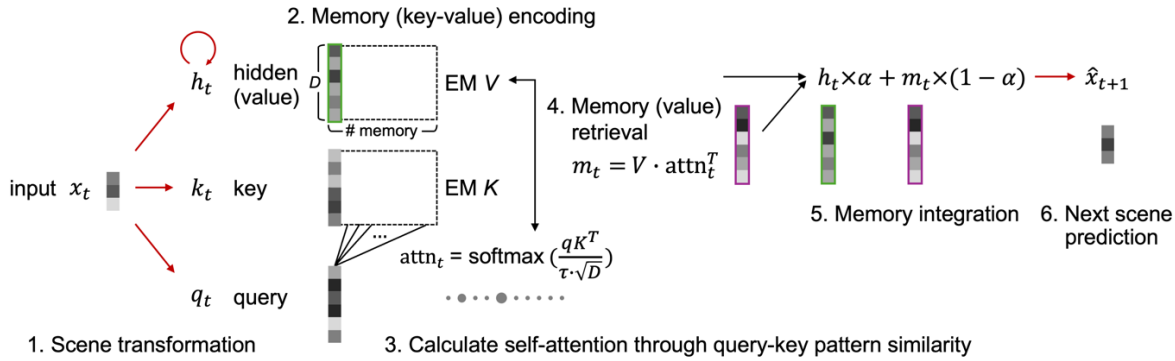


Figure 1. Model architecture. This figure illustrates a set of processes taking place at one time step. 1) *Scene transformation*. The model receives an input vector x_t at time t and transforms it into three distinct representations, a hidden state h_t, a key k_t, and a query q_t. Input-to-hidden transformation follows a nonlinear GRU RNN, and the hidden state corresponds to a value in the key-value system. Input-to-key and input-to-query mappings follow a linear transformation. 2) *Memory (key-value) encoding*. The hidden state and key vectors are encoded in the EM. The EM buffers are thus matrices in which the number of stored memories (rows) increases over time up to a predefined limit set by a hyperparameter. In our model, the number of memories corresponds to the number of scenes in episode 1, which means the model can store memories of all past scenes at test without forgetting. 3) *Calculate self-attention through query-key pattern similarity*. The query vector surveys all past keys stored in the EM and assigns attention weights based on the softmax function. The lower the τ, the more selective the model becomes in assigning weights. In our model, τ was set to 0.1 to increase sparsity. Circles illustrate sparse attention weights given to past keys. See **Supplementary Figure S1** for results with varying τ. 4) *Memory (value) retrieval*. Stored values in EM are summated, weighted by attention weights, which results in the retrieved memory m_t. 5) *Memory integration*. The current scene representation h_t is combined with the retrieved memory m_t. α is set to 0.5, such that the past and present representations are averaged. See **Supplementary Figure S1** for results with varying α. 6) *Next scene prediction*. The model learns a linear transformation from an integrated representation to the predicted next scene representation. The model parameters, indicated with red arrows, are optimized using gradient descent to minimize the loss between the predicted and observed next scenes. See ***Methods*** for details.

We asked whether the model, while watching a television episode, would learn the underlying event structure and retrieve relevant past memories to guide comprehension. We used eighteen episodes from *This Is Us* Season 1, a character-focused television drama that follows the lives of members of a family, jumping back and forth in time from their early childhood to older adulthood. In turn, the narrative revolves around the same main characters, but at different stages of their lives. To represent the narratives, each episode was segmented into a sequence of ~4-second scenes based on shot changes, which resulted in, on average, 614 $\pm$ 68 scenes comprising an episode. Visual and semantic contents of each scene were represented as a 512-dimensional embedding vector using a pretrained Contrastive Language Image Pretraining (CLIP) model (Radford et al., 2021), which was reduced to the top 50 principal components. These embedding time series served as inputs to the model.

The model was trained to repeatedly watch episodes 2-18, during which model parameters were optimized through a gradient descent. We used these episodes because we wanted the model to acquire some general knowledge, such as how this television series was filmed and how characters in this series usually behave or interact with one another. We then fixed the learned model parameters and tested the model on the held-out episode 1. In Song et al. (2025), episode 1 was segmented into 48 events (mean 52 $\pm$ 12s) and presented in three different scrambled sequences to three groups of fMRI participants (N=12 per group). For the model, the same stimulus was also presented in three different scrambled sequences at test, such that the results could be comparable to human participants' neural and behavioral results.

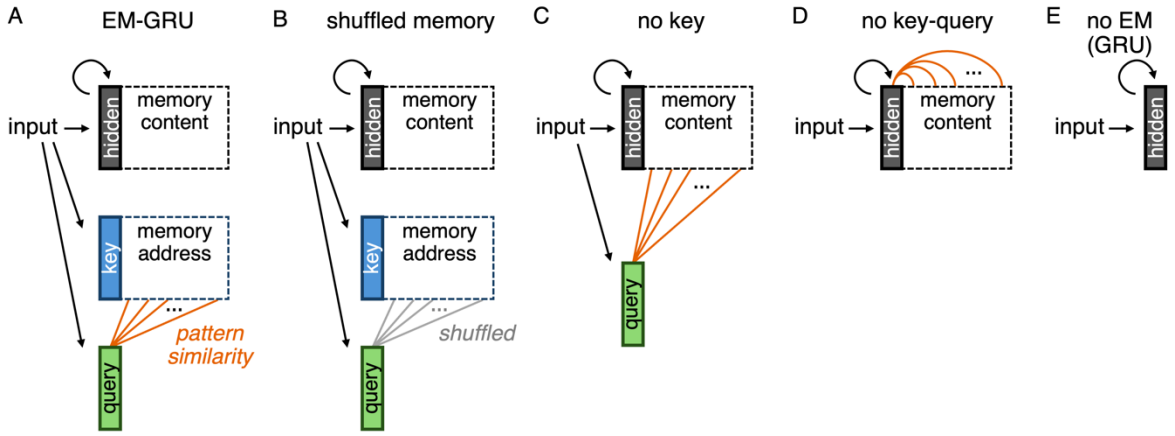


Figure 2. EM-GRU and comparison models. **(A)** EM-GRU model. A simplified visualization of our main model in **Figure 1**, focused on input transformation and memory retrieval based on pattern similarity. **(B)** *Shuffled memory* model. The architecture is the same as EM-GRU, but we shuffled attention weights calculated between a query and keys, so that the model retrieves random memories. This serves as a null model that tests whether memories retrieved by the model are informative for task performance and event representations. **(C)** *No key* model. The model does not have keys that address memory contents. Thus, pattern similarity is calculated directly between a query and hidden states. This model tests the usefulness of dissociating representations used for encoding and retrieval. **(D)** *No key-query* model. The model has neither a key nor a query. Instead, the hidden state of the current moment acts like a query and searches through past hidden states stored in the EM. This design follows the conventional approach in memory research, retrieving information by matching the current state directly to past stored states. **(E)** *No EM (GRU)* model. The model is a GRU RNN that is not augmented with EM. It is designed to test the utility of retrieving memories from the distant past in addition to the recurrence implemented by the GRU. Black arrows indicate learnable model parameters. Orange lines indicate attention calculation based on pattern similarity, such that memories that are assigned high attention weights due to high pattern similarity are selectively retrieved.

We compared the EM-GRU to four control models (**Figure 2**). First, to test the contribution of the self-attention mechanism, we shuffled the attention weights—after calculating them as in EM-GRU—so that the model retrieves random memories (*shuffled memory*; **Figure 2B**). This tested whether EM-GRU's attention to selective memories was meaningful and led to better performance than retrieving random ones. Second, to test the effect of having distinct representations for encoding and retrieval, we lesioned keys from the EM (*no key*; **Figure 2C**). In this variant, inputs were transformed into values and queries, and the query directly searched stored values based on pattern similarity, using the same attention calculation as EM-GRU. This provided a comparison between the key-value mechanism and content-addressable memory retrieval (Kanerva, 2003). Third, to test the effect of removing both keys and queries, we implemented a model frequently used in prior literature (*no key-query*; **Figure 2D**) (Hopfield, 1982; Howard and Kahana, 2002; Polyn et al., 2009; Li et al., 2024; Lu et al., 2024a). In this design, the current hidden state acts like

the query, searching past hidden states to find those with high pattern similarities. Notably, there was no additional transformation beyond the input-to-hidden state mapping, meaning pattern similarity was computed within the same representational space. Finally, to assess the benefit of having an episodic memory buffer with the RNN, we trained a standard GRU RNN without the external EM (*no EM*; F**igure 2E**). This baseline model did not have keys, queries, or memory encoding and retrieval processes that took place at every time step.

Figure 3 shows EM-GRU and the control models' next-scene prediction performances over the course of multiple training iterations. All five models learned well (**Figure 3A**) and generalized well (**Figure 3B**). All models learned rapidly up to near 10th training iteration, after which the loss decreased more gradually. Generalization loss at test was minimal near the 35th training iteration and rose after that. Thus, we stopped the training at the 50th iteration and report selective findings from the 35th iteration in this paper. The *no EM (GRU)* model learned the fastest among the five but eventually converged to a similar level of performance as the EM-GRU. The *shuffled memory* model performed the worst among the five models.

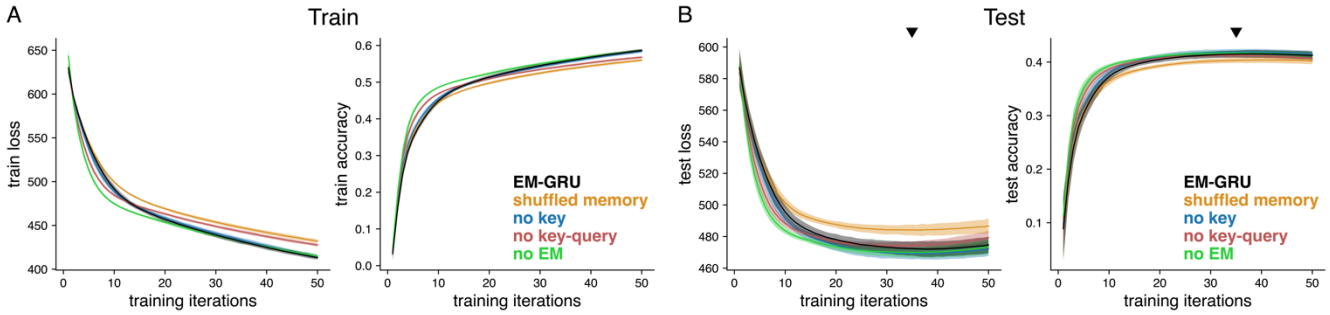


Figure 3. Model performance at train and test. **(A)** Train performance. One training iteration includes 17 rounds of parameter updates as the model watches episodes 2 to 18. **(B)** Test performance. The model was tested on the held-out episode 1 after different rounds of training iterations. Loss indicates the mean squared error between the predicted and observed next scene vectors, averaged across time. Accuracy indicates rank correlation between the predicted and observed next scene vectors, averaged across time. Black lines indicate EM-GRU results, and the colored lines indicate results from the comparison models. Lines indicate the mean and error bars indicate 95% confidence intervals across 20 repetitions using different random seeds. The triangle points to the 35th training iteration where the loss is minimized at test.

The model retrieves memories similar to those retrieved by humans

The EM-GRU and the control models all performed and generalized well in the next-scene prediction task, which was explicitly optimized in training. More importantly, we asked whether the model retrieved memories that are meaningful for comprehension. To ask this, we compared memories retrieved by the model to memories retrieved by human participants as they both watched the same sitcom episode.

Human memory retrieval was directly coded from participants' verbal responses made during the experiment. In Song et al. (2025), 41.15 ± 16.64 % of people's insight explanations included mentions of past events. If an event A was retrieved at an insight during event B, +1 was coded in the events A-B pair in the memory retrieval matrix. The memory retrieval matrices created for all participants were averaged (**Figure 4A** *Left*). The study has found that participants were more likely to retrieve causally related past memories, even more so than semantically or perceptually similar memories.

Unlike the human memory retrieval matrix, which reflected selective retrieval at insight moments, the model's memory retrieval matrix reflected retrieval occurring at every time step throughout comprehension. Our self-attention mechanism was based on pattern similarity between the current query and past keys, such that keys with higher pattern similarity with the query received higher attention weights. To assess whether one event was more or less likely to be retrieved by another event, we averaged the query and key representation patterns across all scenes corresponding to each of the 48 events. We then computed query-key pattern similarities for all event pairs, generating an event-by-event memory retrieval matrix of the model (**Figure 4A** *Right*). In control models that did not have queries or keys, the representations serving those roles were instead used to compute memory retrieval matrices (i.e., query-value similarity for the *no key* model and value-value similarity for the *no key-query* model).

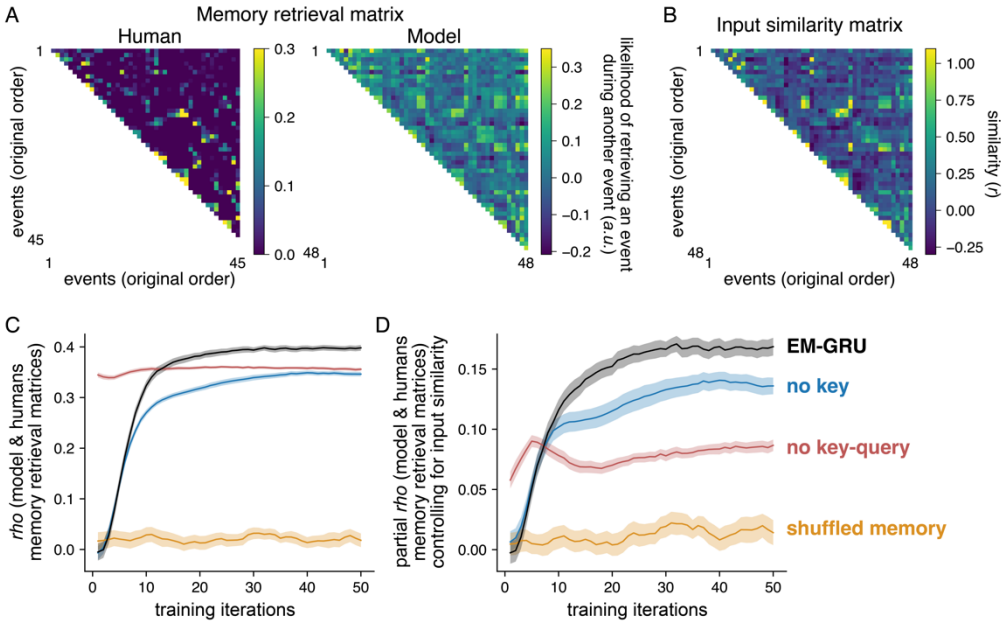


Figure 4. Humans and model's memory retrieval. **(A)** Memory retrieval matrices of the model and human participants. (*Left*) Humans' memory retrieval matrix from Song et al. (2025) based on participants' verbal explanations of their insight moments. If an event was retrieved when explaining an insight during another event, +1 was noted for that event pair. Memory retrieval matrices of 32 participants assigned to three different scrambled order groups were averaged. (*Right*) Because the likelihood of retrieving one event during another event was determined by pattern similarity between the query and the key, the model's memory retrieval was represented by the query-key pattern similarity (Spearman's rank correlation). Memory retrieval matrices created for three different scrambled sequences and 20 different random seeds were averaged. **(B)** Input similarity matrix. The CLIP embeddings of all scenes corresponding to one event was averaged to represent an input pattern for that event. We calculated an event-by-event input pattern similarity, which was controlled for in subsequent analyses. **(C)** Relating the model's memory retrieval with humans' memory retrieval. The model and humans' memory retrieval matrices were correlated across different training iterations. High correlation between the two matrices indicates that the model retrieved similar event memories as humans did. **(D)** Relating the model's memory retrieval with humans' memory retrieval, while controlling for input pattern similarity. Note that the *no EM (GRU)* model was not analyzed, as this model does not retrieve memories. Lines indicate the mean and error bars indicate the standard error of the mean across 20 repetitions using different random seeds.

Over the course of training, the correlation between the model's and humans' memory retrieval matrices increased up to near ρ = 0.4 for EM-GRU (ρ = 0.397 $\pm$ 0.032 at training iteration = 35; **Figure 4C**). This

indicates that the model became increasingly human-like in terms of which memories it selected for retrieval. The resemblance was stronger than that of the *no key* model (ρ = 0.345 $\pm$ 0.026) and the *shuffled memory* model (ρ = 0.022 $\pm$ 0.042). These results suggest that including keys, and thus having separate representations for encoding and retrieval, was useful for memory retrieval. They also suggest that the self-attention mechanism selected memories that meaningfully mapped onto human behavior. Notably, the *no key-query* model, which directly retrieves past hidden states based on their similarity with the current hidden state, immediately produced human-like retrieval even before training. We reasoned that this could be due to an intrinsic correlation between memory retrieval and the model inputs' pattern similarity (i.e., CLIP embedding similarity), which reflects semantic and perceptual similarities between events (ρ = 0.472, p = 5e-56; **Figure 4B**). To account for this, we controlled for the input pattern similarity matrix when correlating the model's and humans' memory retrieval matrices. Even after this control, the EM-GRU model showed the highest correlation with humans compared to the control models (partial ρ = 0.169 $\pm$ 0.032; **Figure 4D**). As expected, the correlation of the *no key-query* model dropped substantially (partial ρ = 0.081 $\pm$ 0.019), to a level below that of the *no key* model (partial ρ = 0.137 $\pm$ 0.036). Together, these findings suggest that the EM-GRU model—optimized for next-step prediction—retrieved memories that were relevant to comprehending ongoing events.

The model represents causally related events with similar patterns

The EM-GRU learned to retrieve memories like humans during the next-scene prediction task. We then asked whether, through retrieving these memories, the model also came to represent causally related events with more similar patterns. Song et al. (2025) showed that causally related events are represented by similar brain patterns, beyond semantic and perceptual similarities. Here, we replicated this analysis, but by replacing brain representations with the model's internal representations.

At test, we extracted the time series of hidden states h_T, keys k_T, and queries q_T. Representation patterns of all scenes corresponding to each of the 48 events were averaged to represent that one event. By taking correlations of all pairs of these 48 events, we created an event-by-event representation similarity matrix (RSM; **Figure 5A**). The RSM was then correlated with the episode's causal relationship matrix, which denotes a causal relationship between every event pair on a scale of 0 to 4, rated by 3 raters with high reliability (Song et al., 2025) (**Figure 5B**).

Across training, the EM-GRU increasingly reflected events' causal relationship structure in its representations (**Figure 5C**). This effect was evident for hidden states, keys, and queries, with the EM-GRU model RSMs consistently showing stronger correlations with the causal relationship matrix than comparison models. In other words, the model learned to represent causally related events with similar patterns, whereas this trend was weaker in control models. Although EM-GRU (ρ = 0.425 $\pm$ 0.026) showed numerically higher correlations than the *no key* model (ρ = 0.419 $\pm$ 0.020), the two were statistically not significant in how much their hidden state RSMs captured causal relationship structure (paired *t*-test: $t(19)$ = 1.028, FDR-p = 0.317; whereas EM-GRU compared to other models: FDR-p values < 0.02). This suggests that a model where queries directly search through memory contents without the keys has sufficiently captured the causal relationship structure as the EM-GRU.

This pattern of results persisted after controlling for input pattern similarity which was highly correlated with the causal relationship (ρ = 0.542, p = 3e-87; **Figure 5D**). Although correlation values decreased overall, partial correlations remained significantly above chance, as confirmed by comparison with a null distribution generated by shuffling hidden state RSMs 1,000 times (z = 15.295, p < 0.001). Together, the findings indicate that the EM-GRU captured not only perceptual and semantic similarities explicit in the

inputs but also a latent causal structure, reflecting a higher-order organization of events that goes beyond content similarity.

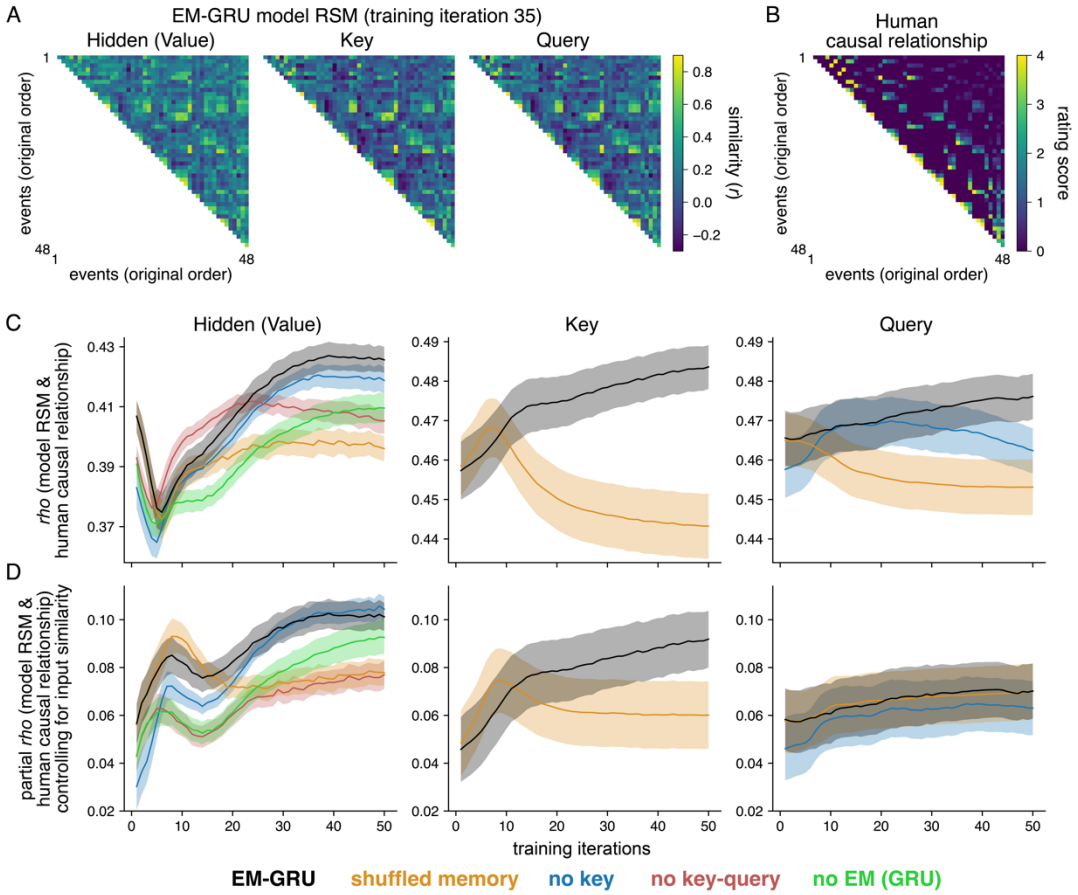


Figure 5. Representation of causal event structure. **(A)** Event-by-event model representation similarity. Upon training the model 35 times and testing it on episode 1, we extracted the representational time series of h_T, k_T, and q_T. Scene patterns pertaining to an event were averaged, and the resulting patterns representing 48 events were correlated in pairs. This resulted in an event-by-event representation similarity matrix (RSM) for each model representation. RSMs of 20 different random seeds were averaged for visualization. **(B)** Event-by-event human-rated causal relationships. Manual ratings of causal relationships between all pairs of 48 events were obtained from Song et al. (2025). High values indicate that the two events are highly causally related to one another. **(C)** Relating model representation similarity with causal relationships. The model RSMs were correlated with the human-rated causal relationship matrix across different training iterations. High correlation between the two matrices indicates that the model represents causally related events with similar patterns. **(D)** Relating model representation similarity with causal relationships, while controlling for input pattern similarity. Lines indicate the mean and error bars indicate the standard error of the mean across 20 repetitions using different random seeds.

The model uses distinct representations for encoding and retrieval

We showed that the EM-GRU retrieves memories and represents causal relationship structure like humans, more so than the comparison models that do not distinguish value and key representations for encoding and retrieval. To examine how the model utilizes these representations, we measured similarities among values, keys, and queries across training iterations (**Figure 6**). Correlations between value–query and value–key representations remained near zero, suggesting that the model maintained orthogonal

representations for encoding versus retrieval. In contrast, key–query similarity increased over training and stabilized around ρ = 0.3 (ρ = 0.277 $\pm$ 0.015 at training iteration = 35).

Hypothetically, if keys and queries were identical in EM-GRU, self-attention would be computed among scene representations linearly transformed by the same weight matrix from the input (akin to the pattern similarity schematics in **Figure 2D**, but in the key space). This design can be useful, specifically in this context in which memory retrieval, causal relationship, and input similarity are all correlated with one another. This likely explains why EM-GRU's keys and queries became moderately similar over training—yet the similarity did not reach 1.0 but instead reached an asymptote near 0.3.

To test whether moderate key–query similarity was more beneficial than setting them equal, we trained another variant of EM-GRU where the input-to-key and input-to-query transformation matrices were forced to be identical (i.e., key–query similarity of 1.0). This model's memory retrieval was significantly less similar to humans' memory retrieval than the original EM-GRU upon controlling for input similarity (paired *t*-test: $t(19) = 8.227$, p = 1e-07; partial $\rho = 0.092 \pm 0.034$, compared to original EM-GRU of partial $\rho = 0.169 \pm 0.032$). No significant difference was found in how much the two models' hidden state RSMs captured causal relationships (paired *t*-test: $t(19) = 1.132$, $p = 0.272$; partial $\rho = 0.111 \pm 0.020$, compared to original EM-GRU of partial $\rho = 0.101 \pm 0.029$). These findings suggest that having similar keys and queries can be useful in contexts where input similarity is tightly coupled with the causal relationship and memory retrieval. Nonetheless, our EM-GRU model that learned distinct key and query transformations from scratch better captured human-like memory retrieval behavior when controlling for the explicit input similarity.

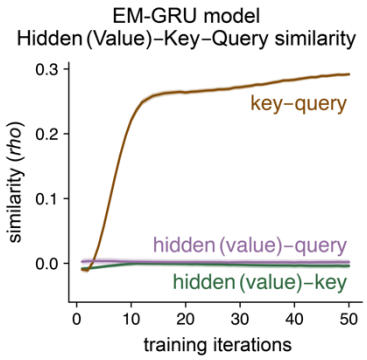


Figure 6. Pairwise similarity between the hidden state (value), key, and query of the EM-GRU. Lines indicate the mean and error bars indicate the standard error of the mean across 20 repetitions using different random seeds.

Model representations exhibit pattern separation and pattern completion

The EM-GRU is a cognitive model, meaning its design is agnostic to neurobiological circuits. Yet, we still asked if this model, which closely resembles human behavior, exhibits similar information coding principles as the human brain. A characteristic feature of the hippocampus is the coexistence of pattern separation and pattern completion (O'Reilly and McClelland, 1994; Rolls, 2016). At encoding, the hippocampus distinguishes two similar inputs by representing them with more distinctive patterns, called pattern separation. In contrast, during retrieval, the hippocampus represents the two related inputs with similar patterns, which may allow one to cue the other through pattern completion. The dentate gyrus and CA3 subregions of the hippocampus have been proposed to support pattern separation and pattern completion, respectively (Neunuebel and Knierim, 2014).

To search for evidence of pattern separation and pattern completion in the EM-GRU model, we calculated pattern correlations between scene pairs that correspond to the same event (**Figure 7A**). The scenes' hidden state representations became dissimilar to one another over the course of learning, meaning scene representations within an event became more distinctive from one another, paralleling pattern separation (ρ = 0.629 $\pm$ 0.016 at training iteration = 35, with respect to within-event scene correlation of the raw input embeddings: mean ρ = 0.406). In contrast, the within-event scenes' key and query representations became more similar to one another over time, paralleling pattern completion (key: ρ = 0.377 $\pm$ 0.009, query: ρ = 0.381 $\pm$ 0.008 at training iteration = 35) (**Figure 7B**). It is notable that values represent memory contents stored in encoding, which is when pattern separation occurs, while keys and queries are used during retrieval, which is when pattern completion occurs. Note that we are not arguing that this is the model of the hippocampus, because we did not enforce any hippocampus-like architectural constraint in model design as has been done in previous studies (Whittington et al., 2020; Chandra et al., 2025). Yet, the way this normative model represents the scenes in two distinct ways within a system suggests that this may be a computational mechanism by which the hippocampus pattern separates during encoding and pattern completes during retrieval. By distinguishing representations used for encoding and retrieval, the model exhibits both pattern separation and pattern completion.

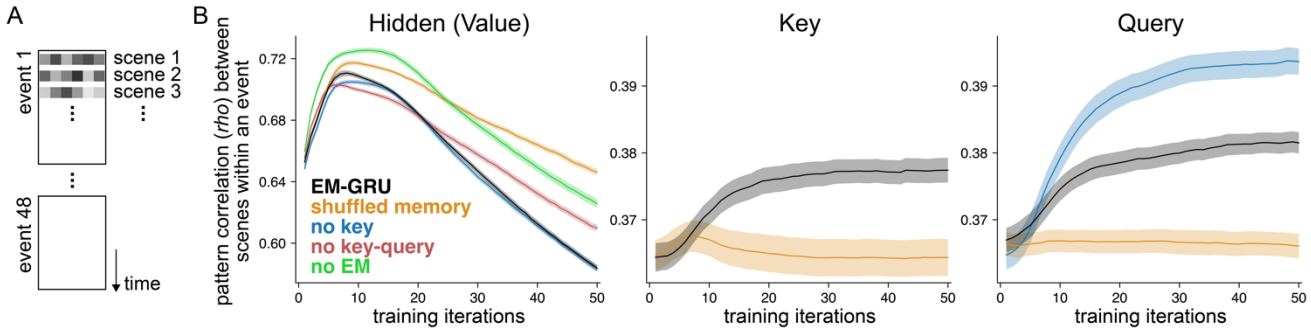


Figure 7. Pattern similarities among scenes within an event. **(A)** Schematics of the data structure. For three internal representations, the hidden state, key, and query, we grouped scenes into 48 predefined events and estimated pattern correlations of the scene pairs belonging to the same event. Correlation values were averaged across all scene pairs. **(B)** Pattern correlations between scenes within an event across different training iterations. The increase in correlation indicates that the scenes became more similar to one another within an event, whereas the decrease indicates that the scenes became more discriminable within an event over the course of training. Lines indicate the mean and error bars indicate the standard error of the mean across 20 repetitions using different random seeds.

The model's event representation resembles that of the human brain

Finally, we asked if the model represents narrative events in a way that resembles how the human brain represents the same events during fMRI. For each participant in Song et al. (2025), we extracted voxel-wise BOLD activity time series for each of the 200 cortical (Schaefer et al., 2018) and 32 subcortical regions (Tian et al., 2020) and averaged the time series per event to extract voxel activity patterns representing 48 events. For every brain region, we calculated an event-by-event RSM, estimating how similar the voxel activity patterns are for every pair of 48 events. We correlated the brain RSM to the model RSM (**Figure 5A**), to ask if the model's event representation resembles that of the functional brain.

Among the 232 areas tested, we found that most brain areas' RSMs positively and significantly correlated with the EM-GRU model RSMs (at training iterations = 35), compared to chance distributions where brain RSMs were randomly shuffled (1,000 iterations). Specifically, the representation of the model's hidden states resembled 207 out of 232 brain regions' representational patterns significantly above chance, 224

regions for keys, and 223 regions for queries (FDR-corrected $p < .05$). These suggest that the model's representations resemble the representational patterns observed in the majority of brain areas.

Importantly, we asked whether the inclusion of the EM buffer boosted the model's representational similarity with the brain. We tested whether the hidden state RSM of the EM-GRU model is more similar to brain RSMs, compared to the hidden state RSM of the GRU model. We selected the GRU as the comparison model because this model does not have an EM buffer (but see **Supplementary Figure S2** for comparisons with different control models). Most brain regions—except for the sensorimotor areas—exhibited significantly higher similarity with the EM-GRU's hidden states compared to the GRU's hidden states (paired *t*-tests, FDR-corrected *p* values < .05) (**Figure 8**). This suggests that the EM improved the RNN's hidden state representations, making them not only better reflect causal event structure but also more closely match the brain's representational patterns.

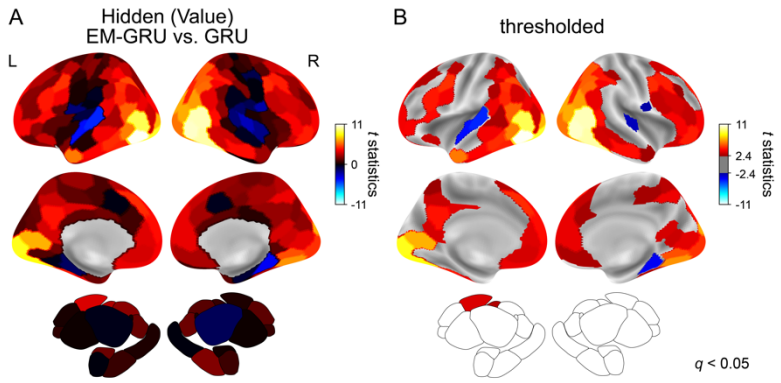


Figure 8. Model and human brain's event representations. **(A)** Paired comparison between the EM-GRU and GRU models' hidden states in how much they resemble 232 brain regions' representation patterns. The EM-GRU and GRU models' hidden state RSMs were correlated with the voxel pattern RSMs of each brain area of 33 participants. These 33 correlation values of the EM-GRU and GRU models were compared using paired *t*-tests, repeatedly across 232 brain regions. Color denotes *t* statistics. Red-yellow colors indicate that the region's representation was more similar to the EM-GRU, whereas blue-cyan colors indicate that the region's representation was more similar to the GRU. **(B)** Areas that hold statistical significance. The same result as **A** but thresholded at FDR-corrected $p < 0.05$, corrected for 232 brain regions. The figure shows the lateral (*Top*) and medial (*Middle*) visualizations of the cortex and the lateral visualization of the subcortex (*Bottom*).

Discussion

In this study, we introduced an RNN augmented with a key-value EM buffer that continuously encodes and retrieves memories to comprehend an ongoing event. We found that the model represents causally related events with similar patterns and retrieves past events that are comparable to what human participants retrieve. The model exhibits both pattern separation and pattern completion in distinct components of the architecture, mirroring known functions of the hippocampus. Moreover, its representational structure aligns more closely with human brain patterns than that of the control models. These findings highlight EM-GRU as a promising candidate model of human memory during event comprehension and suggest a computational mechanism by which we remember and reason about ongoing naturalistic events.

We developed EM-GRU to understand how causally related past events are selectively retrieved during event comprehension. Notably, the model was not explicitly trained to perform causal inference. Instead, it was optimized for next-scene prediction, while memory encoding and retrieval took place as inputs were processed. Nevertheless, it developed structured representations following causal relationships that

extended beyond semantic or perceptual similarities, and its retrieved memories closely aligned with those of human participants. This suggests that causal inference during comprehension emerges from continuous memory encoding and retrieval.

What features of the model allowed it to capture causal dependencies and assign high attention weights to causally related past events? Several features of the EM-GRU appear critical for the observed correspondence with the human brain and behavior. First, the model separates representations for encoding and retrieval. A query surveys the keys based on similarity, but the actual retrieved contents come from values that are encoded with the keys. This means having distinct representations for encoding and retrieval allows the model to retrieve memories not just based on linear pattern similarity but something beyond that—which captures the underlying causal relationships. This feature distinguishes the EM-GRU from classic frameworks of episodic memory—such as the Hopfield network (Hopfield, 1982), the temporal context model (Howard and Kahana, 2002; Polyn et al., 2009), and a range of EM-augmented RNN models (Li et al., 2024; Lu et al., 2024a)—which directly store internal representations in episodic memory and later retrieve those with high pattern similarity to the current state.

Second, EM-GRU includes two notable integration processes. One involves integrating past memories based on attention weights calculated from the self-attention mechanism. This is motivated by a view that people retrieve not just one memory, but a blend of multiple memories that are assigned different weights (Manning, 2021). Yet, there is a sweet spot—the model performs better if it is attending to a few selective memories by assigning sparse attention weights compared to broadly summing up arrays of memories using diffuse attention weights (**Supplementary Figure S1**) (Anderson and Reder, 1999; Norman and O'Reilly, 2003; Norman, 2010). The second integration combines the current representation (h_t) with the retrieved memory (m_t) to predict the upcoming scene. This integration of the past and the present did not improve accuracy in the next-step prediction task (**Figure 3**). However, it was critical for emergent causal reasoning and contributed significantly to the model's alignment with human behavior (**Supplementary Figure S1**).

Third, the model simulates both working memory and episodic memory processes within its architecture. Working memory was implemented via recurrence and nonlinear transformations of the inputs to the hidden states in the RNN, allowing the model to maintain memories over relatively short timescales (Whittington et al., 2024; Zheng et al., 2025). Episodic memory was supported by the key-value EM buffer, which grants full access to memories from the distant past.

We have emphasized that EM-GRU is a normative model of cognition that does not directly simulate neurobiological circuits. Nevertheless, its ability to exhibit hippocampus-like pattern separation and completion, as well as its close alignment with the human brain's representation patterns, suggest that the model can offer insights into functions of the human brain. The model's key-value memory system was initially designed based on the hippocampal memory indexing theory, which proposes that hippocampal engram cells index memory contents stored in neocortical areas (Teyler and DiScenna, 1986). However, interpretations of model components vary depending on the neural system being modeled—whether hippocampal-entorhinal (Whittington et al., 2020; Chandra et al., 2025) or hippocampal-prefrontal cortical system (Whittington et al., 2024; Zheng et al., 2025). Future work can continue to understand and develop this model in relation to neurobiological circuits.

In sum, the EM-GRU model, though not explicitly trained for causal inference, spontaneously developed internal representations that reflect causal relationships between events. The model also exhibited memory retrieval behaviors that resemble those of humans, and brain-like coding properties such as pattern

separation and completion. These findings provide insights into the computational underpinnings of how people reason about causal relationships between events through memory encoding and retrieval.

Data and Code Availability

Model and analysis codes and human behavioral data are openly available in: https://github.com/hyssong/memorymodel.

Acknowledgment

We thank JeongJun Park with conceptualization and helpful discussions and comments on the manuscript. The research was supported by the McDonnell Center for Systems Neuroscience and the McDonnell Center for Cellular and Molecular Neurobiology at Washington University in St. Louis (HS).

Methods

Stimuli

Eighteen episodes of a television series, *This is Us* Season 1 (2016, directed by Requa, J. & Ficarra, G. and written by Fogelman, D.) were used as stimuli. Episodes 2-18 were used as training data (mean 41m 37s, ranging from 40m 16s to 43m 31s) and episode 1 (41m 40s) was used as test data. During test, the episode was segmented into 48 events (mean 52 $\pm$ 12s, ranging from 8s to 1m 37s) and ordered in a temporally scrambled sequence. Temporal scrambling followed the manipulation of Song et al. (2025), with the goal of relating model results to human behavioral and fMRI data. In Song et al. (2025), 36 fMRI participants were assigned to one of three groups where they watched the same events but in a different sequence. Each event was positioned at either the beginning, middle, or end of the sequence in the three groups, so that the order effect could be counterbalanced when averaging across groups. Episodes 2-18 were used in training without any manipulation to the video. Yet, the director has originally filmed the narrative to jump back and forth in time, switching between past, present, and future narratives of multiple characters.

To finely segment the episodes into multiple scenes with minimal autocorrelation between consecutive scenes, we applied an automated scene segmentation algorithm (PySceneDetect) that cuts the episode into multiple scenes based on automatically-detected camera shot changes. This resulted in on average 614 $\pm$ 68 scenes per episode (ranging from 523 to 751), with each scene lasting for 4.07 $\pm$ 4.78s (ranging from 0.13s to 182.68s).

We applied the CLIP model to the frames of each scene, extracting 512-dimensional embeddings representing the visual and semantic contents of the visual frame (Radford et al., 2021). Multiple frames' embedding vectors were averaged to represent one scene, and the resulting embedding time series of hundreds of scenes represented the unfolding narrative of an episode. Upon concatenating the train data, we *z*-normalized each embedding feature across time, of which mean and standard deviation were applied to normalize the test data.

To constrain the number of model parameters, we applied a principal component analysis to the training data to reduce the number of dimensions to 50, which were again *z*-normalized across time per feature. We transformed the test data to the trained principal component space, where we again applied the mean and standard deviation to normalize the test data. These dimensionality-reduced embedding time series were used as inputs to the model, where the input at time t (x_t) corresponded to the embedding vector of a scene.

To create an event-by-event input pattern similarity matrix, we averaged CLIP embedding vectors of all scenes that correspond to each of the 48 events to have one input pattern representing each event. Pearson's correlation was calculated for all pairs of input patterns.

Model

The core architecture of the model is a GRU RNN (Cho et al., 2014). At each time step t, the CLIP embedding $x_t \in \mathbb{R}^d$ ($d = 50$) was transformed into a $6d$-dimensional vector, decomposed into $i_r, i_z, i_n \in \mathbb{R}^{2d}$, via the parameters $\boldsymbol{W_x} \in \mathbb{R}^{6d \times d}$ and $\boldsymbol{b_x} \in \mathbb{R}^{6d}$. The previous hidden state $h_{t-1} \in \mathbb{R}^{2d}$ was likewise transformed into a $6d$-dimensional vector, decomposed into $h_r, h_z, h_n \in \mathbb{R}^{2d}$, using parameters $\boldsymbol{W_h} \in \mathbb{R}^{6d \times 2d}$ and $\boldsymbol{b_h} \in \mathbb{R}^{6d}$. The reset gate r_t and update gate z_t were then computed to control how much of the previous hidden state was carried forward versus updated. The new hidden state h_t was obtained as a transformed, weighted combination of x_t and h_{t-1}.

$$
\begin{aligned}
[i_r, i_z, i_n] &= \boldsymbol{W_x} x_t + \boldsymbol{b_x} \\
[h_r, h_z, h_n] &= \boldsymbol{W_h} h_{t-1} + \boldsymbol{b_h} \\
r_t &= \sigma(i_r + h_r) \\
z_t &= \sigma(i_z + h_z) \\
\tilde{h}_t &= \tanh(i_n + r_t \odot h_n) \\
h_t &= \tilde{h}_t + z_t \odot \left(h_{t-1} - \tilde{h}_t\right)
\end{aligned}
$$

We augmented this RNN with an EM buffer, where an RNN can write to (i.e., encoding) and read from (i.e., retrieval) an external memory storage. Inputs are linearly transformed into keys ($k_t \in \mathbb{R}^{2d}$) that represent memory addresses and are encoded in the EM, and queries ($q_t \in \mathbb{R}^{2d}$) that match the keys to calculate self-attention based on similarity and retrieve the associated values stored in the EM. The hidden states (h_t) generated from the RNN serve as values ($v_t = h_t$), representing memory contents and are encoded in an EM buffer at every time step. Linear transformations are denoted with $\boldsymbol{W_k}$ and $\boldsymbol{W_q} \in \mathbb{R}^{2d \times d}$.

$$
\begin{aligned}
k_t &= \boldsymbol{W_k} x_t \\
q_t &= \boldsymbol{W_q} x_t
\end{aligned}
$$

The key and value at every time step were encoded in the EM, with the maximum number of memory (M) set as the number of time steps in test data.

$$
\begin{aligned}
EM(K) &= [k_t, k_{t-1}, k_{t-2}, \dots, k_1] \\
EM(V) &= [v_t, v_{t-1}, v_{t-2}, \dots, v_1]
\end{aligned}
$$

The query surveys all keys stored in EM with the following equation. The equation assigns high attention weights to keys that are similar in pattern with the query, with the softmax function making the attention weights sum to 1.

$$
\mathrm{attn}_t = \mathrm{softmax}\left(\frac{q_t K^T}{\tau \cdot \sqrt{D}}\right), \qquad \text{where } K = [k_{t-1}, k_{t-2}, \dots, k_1]
$$

τ was set to 0.1 to enforce sparsity in memory selection—meaning, higher attention weights were given to a small number of past scenes while most of the memories were given near zero weights. Because the dimension of the hidden state (D) was 100, the attention calculation essentially equaled to: $\mathrm{attn}_t = \mathrm{softmax}(qK^T)$. Attention is calculated from the match between key and query, but the model retrieves values stored in the EM that are associated with the attended keys. The retrieved memory at time t is denoted as $m_t \in \mathbb{R}^{2d}$, which is the summation of past values, weighted by the model-learned self-attention.

$$
m_t = V \cdot \mathrm{attn}_t{}^T, \qquad \text{where } V = [v_{t-1}, v_{t-2}, \dots, v_1]
$$

The outcomes of the illustrated steps are h_t representing the current scene and m_t representing the retrieved and integrated past memories. We integrated the two representations by a weighted average of the two vectors: $h_t \times \alpha + m_t \times (1 - \alpha)$. Here, α was set to 0.5, which means the two vectors were simply averaged in memory integration. Given this integrated vector, model parameters $\boldsymbol{W_y} \in \mathbb{R}^{d \times 2d}$ and $\boldsymbol{b_y} \in \mathbb{R}^d$ were fitted to predict the embedding of the next time step x_{t+1}.

$$
\hat{x}_{t+1} = \boldsymbol{W_y} \cdot (h_t \times \alpha + m_t \times (1 - \alpha)) + \boldsymbol{b_y}
$$

The model parameters were initialized as follows. Weight parameters $\boldsymbol{W_x}$, $\boldsymbol{W_h}$, and $\boldsymbol{W_y}$ were initialized using orthogonal initialization, and bias parameters $\boldsymbol{b_x}$, $\boldsymbol{b_h}$, and $\boldsymbol{b_y}$ were initialized with a constant value of 0.1. For the input-to-key and input-to-query transformation matrices, $\boldsymbol{W_k}$ and $\boldsymbol{W_q}$, we used trainable parameters initialized with Gaussian noise scaled by a factor of 0.1.

The model was trained using an Adam optimizer (torch.optim.Adam) with a learning rate of 0.001. The training objective was to minimize the mean squared error between the predicted and observed vectors (nn.MSELoss).

The model was trained 50 times, with each iteration comprising weight optimizations with 17 episodes' embedding time series. In total, this amounted to 850 weight updates (17 episodes × 50 iterations). 17 episodes were presented at a random order at every iteration. We repeated this entire training and testing steps 20 times with different random seeds.

Model parameters were fixed at every round of training iteration to be applied to test data. Test was conducted on three different scrambling orders of episode 1. The results for these three test runs were averaged.

Performance of the EM-GRU model was compared to four other comparison models. First, in *shuffled memory* model, everything was the same as the original EM-GRU except that the attention weights (attn_t) calculated at every time step were randomly shuffled when they were applied to the stored values in the EM. This means randomly selected memories were retrieved as m_t. Second, in *no key* model, the query directly searched through the stored hidden states, thus the self-attention calculation changed as follows.

$$\mathrm{attn}_t = \mathrm{softmax}(\frac{q_t V^T}{\tau \cdot \sqrt{D}})$$

Third, in *no key-query* model, neither key nor query existed, thus the self-attention calculation was based on the following equation.

$$\mathrm{attn}_t = \mathrm{softmax}(\frac{h_t V^T}{\tau \cdot \sqrt{D}})$$

Here, if the V were to include the nearby time steps (e.g., h_{t-1}, h_{t-2}), their pattern similarity with h_t was bound to be high because h_t is dependent on the recurrence from the nearby past time steps. Therefore, we only included memories of 10 time steps apart so that the model could focus on retrieving memories from distant past.

$$V = [v_{t-10}, v_{t-11}, v_{t-12}, \dots, v_1]$$

Fourth, the *no EM (GRU)* model was implemented as a traditional GRU RNN that does not have the EM buffer. From the input x_t, the model estimates the hidden representation h_t with recurrence, and from this hidden state, directly predicts the next input x_{t+1}.

Human behavioral and fMRI data

Detailed descriptions of the experiments and data are provided in Song et al. (2025). Each fMRI session consisted of 10 runs. During each run, participants watched several events while pressing an "aha" button whenever they understood something new about the story or characters. Afterward, they were shown

screenshots of when they pressed the button and participants verbally explained why they pressed in those moments.

To create a memory retrieval matrix, we analyzed people's verbal responses explaining their insights. More than 40% of verbal responses included mentions of past events. For those cases, we noted an event index of the mentioned past event, an event index of the current moment when they had an "aha," and recorded +1 for that event pair. This resulted in one memory retrieval matrix per person. Upon unscrambling the event sequence to its original order, we averaged all 32 participants' data (4 participants' data were missing due to errors in data collection) to create one memory retrieval matrix that was used in this study. Memory retrieval at events 46 to 48 were omitted during coding, so they were discarded when calculating correlations.

The causal relationship matrix was not collected from this fMRI sample but was rated by three authors of the Song et al. (2025) study. Each event pair was assigned a causal relationship score from 0 to 4. Because the scores were highly reliable across the raters, the three ratings were averaged to create one causal relationship matrix that was used in this study.

Event-by-event representation pattern similarity

The representation patterns of hidden states, keys, and queries estimated at every time step (100 dimensions × time) were summarized to a matrix sized (100 dimensions × 48 events) by averaging patterns within each event. Representation pattern similarity analyses were conducted by taking the Pearson's correlation between events (48 events × 48 events), which were Fisher's z-transformed. Event orders were unscrambled to follow the original order of the episode, and the RSMs of the three scrambled order groups were averaged.

Following the same procedure used for the model, we created brain RSMs from the voxel activity pattern time series. The BOLD activity time series were extracted from all voxels in each of the 200 cortical (Schaefer et al., 2018) and 32 subcortical (Tian et al., 2020) parcels as participants watched 48 events in scrambled orders. The BOLD time series were averaged across time for 48 events respectively, resulting in a (number of voxels × 48 events) representation pattern matrix per region. We created 33 participants' RSMs (3 participants' data were omitted due to head motion) for each of the 232 brain regions, and unscrambled the event orders. Fisher's z-transformed correlation values between these 33 RSMs and the model's RSM were used in paired *t*-tests comparing EM-GRU model results to the GRU model results.

References

Anderson JR, Reder LM (1999) The fan effect: New results and new theories. Journal of Experimental Psychology: General 128:186–197.

Antony J, Lozano A, Dhoat P, Chen J, Bennion K (2024) Causal and Chronological Relationships Predict Memory Organization for Nonlinear Narratives. Journal of Cognitive Neuroscience 36:2368–2385.

Chandra S, Sharma S, Chaudhuri R, Fiete I (2025) Episodic and associative memory from spatial scaffolds in the hippocampus. Nature 638:739–751.

Chang CHC, Lazaridi C, Yeshurun Y, Norman KA, Hasson U (2021) Relating the Past with the Present: Information Integration and Segregation during Ongoing Narrative Processing. Journal of Cognitive Neuroscience 33:1106–1128.

Cho K, van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y (2014) Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp 1724–1734.

Cohn-Sheehy BI, Delarazan AI, Crivelli-Decker JE, Reagh ZM, Mundada NS, Yonelinas AP, Zacks JM, Ranganath C (2022) Narratives bridge the divide between distant events in episodic memory. Mem Cogn 50:478–494.

Dong CV, Lu Q, Norman KA, Michelmann S (2025) Towards large language models with human-like episodic memory. Trends in Cognitive Sciences.

Elman JL, McRae K (2019) A model of event knowledge. Psychological Review 126:252–291.

Fang C, Rajan K (2025) From Memories to Maps: Mechanisms of In-Context Reinforcement Learning in Transformers. arXiv.

Franklin NT, Norman KA, Ranganath C, Zacks JM, Gershman SJ (2020) Structured Event Memory: A Neuro-Symbolic Model of Event Cognition. Psychological Review 127:327–361.

Gershman SJ, Fiete I, Irie K (2025) Key-value memory in the brain. Neuron 113:1694-1707.e1.

Graesser AC, Singer M, Trabasso T (1994) Constructing inferences during narrative text comprehension. Psychological Review 101:371–395.

Griffiths TL, Sobel DM, Tenenbaum JB, Gopnik A (2011) Bayes and Blickets: Effects of Knowledge on Causal Induction in Children and Adults. Cognitive Science 35:1407–1455.

Griffiths TL, Tenenbaum JB (2009) Theory-based causal induction. Psychological Review 116:661–716.

Gumbsch C, Adam M, Elsner B, Martius G, Butz MV (2022) Developing hierarchical anticipations via neural network-based event segmentation. IEEE International Conference on Development and Learning (ICDL).

Hahamy A, Dubossarsky H, Behrens TEJ (2023) The human brain reactivates context-specific past information at event boundaries of naturalistic experiences. Nat Neurosci 26:1080–1089.

Hopfield JJ (1982) Neural networks and physical systems with emergent collective computational abilities. Proceedings of the National Academy of Sciences 79:2554–2558.

Howard MW, Kahana MJ (2002) A Distributed Representation of Temporal Context. Journal of Mathematical Psychology 46:269–299.

Kanerva P (2003) Sparse Distributed Memory. Cambridge, MA, USA: MIT Press.

Krotov D, Hopfield JJ (2021) Large Associative Memory Problem in Neurobiology and Machine Learning. International Conference on Learning Representations (ICLR) 2021 Conference.

Langston M, Trabasso T (1999) Modeling causal integration and availability of information during comprehension of narrative texts. In: The construction of mental representations during reading, pp 29–69. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Lee H, Chen J (2022) Predicting memory from the network structure of naturalistic events. Nat Commun 13:4235.

Li M, Jensen KT, Mattar MG (2024) A neural network model trained on free recall learns the method of loci. Proceedings of the Annual Meeting of the Cognitive Science Society 46.

Lu Q, Hasson U, Norman KA (2022) A neural network model of when to retrieve and encode episodic memories. eLife 11:e74445.

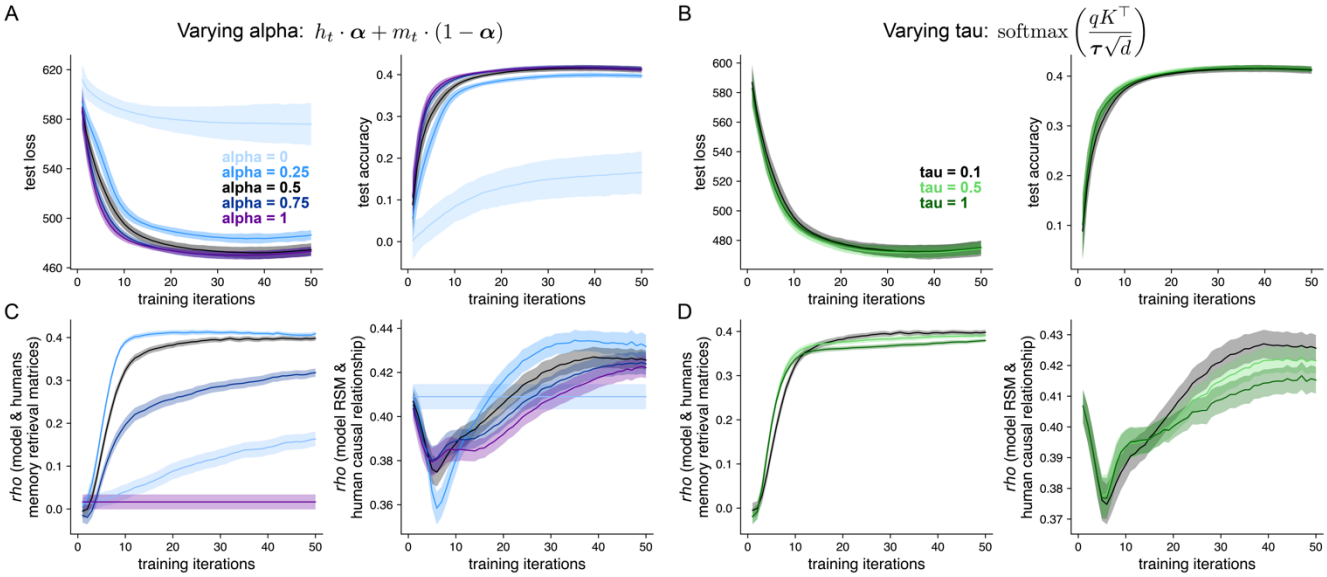
Lu Q, Hummos A, Norman KA (2024a) Episodic memory supports the acquisition of structured task representations. bioRxiv.

Lu Q, Nguyen TT, Zhang Q, Hasson U, Griffiths TL, Zacks JM, Gershman SJ, Norman KA (2024b) Reconciling shared versus context-specific information in a neural network model of latent causes. Sci Rep 14:16782.
Manning JR (2021) Episodic memory: Mental time travel or a quantum “memory wave” function? Psychological Review 128:711–725.
McNamara DS, Magliano J (2009) Toward a Comprehensive Model of Comprehension. In: Psychology of Learning and Motivation, pp 297–384. Elsevier.
Neunuebel JP, Knierim JJ (2014) CA3 Retrieves Coherent Representations from Degraded Input: Direct Evidence for CA3 Pattern Completion and Dentate Gyrus Pattern Separation. Neuron 81:416–427.
Norman KA (2010) How hippocampus and cortex contribute to recognition memory: Revisiting the Complementary Learning Systems model. Hippocampus 20:1217–1227.
Norman KA, O’Reilly RC (2003) Modeling hippocampal and neocortical contributions to recognition memory: A complementary-learning-systems approach. Psychological Review 110:611–646.
O’Reilly RC, McClelland JL (1994) Hippocampal conjunctive encoding, storage, and recall: Avoiding a trade-off. Hippocampus 4:661–682.
Polyn SM, Norman KA, Kahana MJ (2009) A context maintenance and retrieval model of organizational processes in free recall. Psychological Review 116:129–156.
Pritzel A, Uria B, Srinivasan S, Badia AP, Vinyals O, Hassabis D, Wierstra D, Blundell C (2017) Neural Episodic Control. Proceedings of the 34th International Conference on Machine Learning, pp 2827–2836.
Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G, Sutskever I (2021) Learning Transferable Visual Models From Natural Language Supervision. arXiv.
Ritter S, Wang J, Kurth-Nelson Z, Jayakumar S, Blundell C, Pascanu R, Botvinick M (2018) Been There, Done That: Meta-Learning with Episodic Recall. Proceedings of the 35th International Conference on Machine Learning, pp 4354–4363.
Rolls ET (2016) Pattern Completion and Pattern Separation Mechanisms in the Hippocampus. In: The Neurobiological Basis of Memory: A System, Attribute, and Process Analysis, pp 77–113. Cham: Springer International Publishing.
Santoro A, Bartunov S, Botvinick M, Wierstra D, Lillicrap T (2016) Meta-Learning with Memory-Augmented Neural Networks. Proceedings of the 33rd International Conference on Machine Learning (ICML) 48:1842–1850.
Schaefer A, Kong R, Gordon EM, Laumann TO, Zuo X-N, Holmes AJ, Eickhoff SB, Yeo BTT (2018) Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. Cerebral Cortex 28:3095–3114.
Smith KA, Hamrick JB, Sanborn AN, Battaglia PW, Gerstenberg T, Ullman TD, Tenenbaum JB (2024) Intuitive physics as probabilistic inference. In Bayesian Models of Cognition: Reverse Engineering the Mind. Boston, MA: MIT Press.
Song H, Ke J, Madhogarhia R, Leong YC, Rosenberg MD (2025) Cortical reinstatement of causally related events sparks narrative insights by updating neural representation patterns. bioRxiv.
Suh S, Trabasso T (1993) Inferences during Reading: Converging Evidence from Discourse Analysis, Talk-Aloud Protocols, and Recognition Priming. Journal of Memory and Language 32:279–300.
Teyler TJ, DiScenna P (1986) The hippocampal memory indexing theory. Behavioral Neuroscience 100:147–154.
Tian Y, Margulies DS, Breakspear M, Zalesky A (2020) Topographic organization of the human subcortex unveiled with functional connectivity gradients. Nat Neurosci 23:1421–1432.
Trabasso T, Sperry LL (1985) Causal relatedness and importance of story events. Journal of Memory and Language 24:595–611.
Trabasso T, van den Broek P (1985) Causal thinking and the representation of narrative events. Journal of Memory and Language 24:612–630.
Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł ukasz, Polosukhin I (2017) Attention is All you Need. Advances in Neural Information Processing Systems.
Whittington JCR, Dorrell W, Behrens TEJ, Ganguli S, El-Gaby M (2024) A tale of two algorithms: Structured slots explain prefrontal sequence memory and are unified with hippocampal cognitive maps. Neuron 113(2): 312-333.e6.

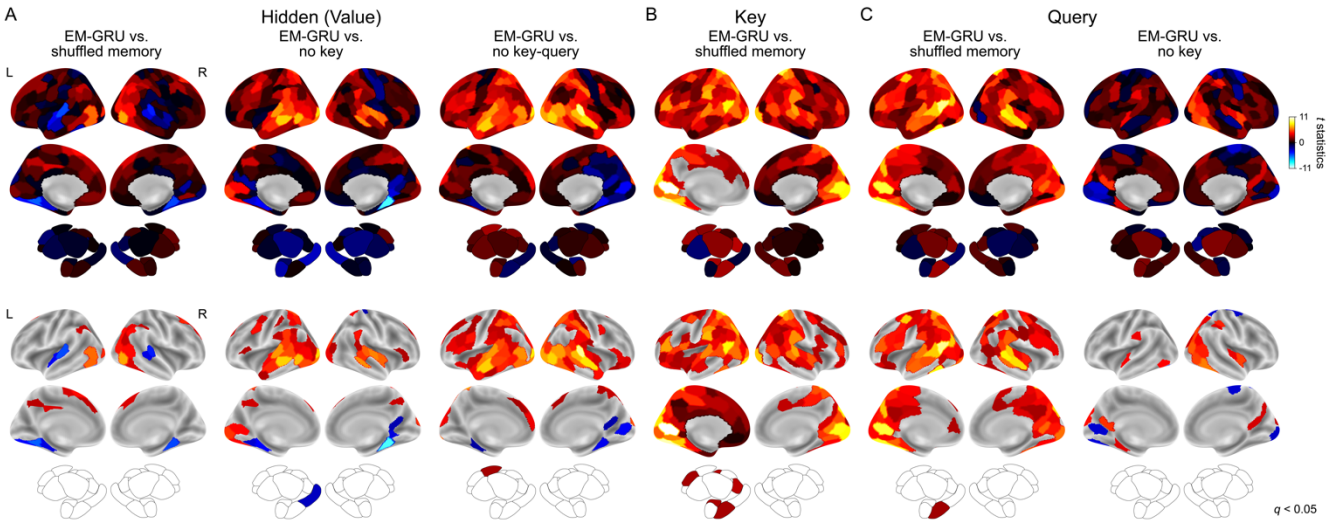
Whittington JCR, Muller TH, Mark S, Chen G, Barry C, Burgess N, Behrens TEJ (2020) The Tolman-Eichenbaum Machine: Unifying Space and Relational Memory through Generalization in the Hippocampal Formation. Cell 183:1249–1263.
Xu K, Srivastava A, Gutfreund D, Sosa F, Ullman T, Tenenbaum J, Sutton C (2021) A Bayesian-Symbolic Approach to Reasoning and Learning in Intuitive Physics. Advances in Neural Information Processing Systems, pp 2478–2490.
Zheng Y, Wolf N, Ranganath C, O'Reilly RC, McKee KL (2025) Flexible Prefrontal Control over Hippocampal Episodic Memory for Goal-Directed Generalization. arXiv.

Supplementary Information

Song *et al.* A neural network with episodic memory learns causal relationships between narrative events



Supplementary Figure S1. EM-GRU results across varying hyperparameters. **(A-B)** Loss and accuracy of the model at test, across varying α and τ. In addition to variations in hyperparameters α and τ, everything else remained the same as the original EM-GRU model. Black lines correspond to **Figure 3B**, with α = 0.5 and τ = 0.1. **(A)** The model performs better with increasing α. This indicates that the next-scene prediction is more successful if the prediction is purely based on the ongoing scene rather than the retrieved memory. **(B)** The model's performance is comparable across different choices of τ. **(C-D)** Correlations between the model and humans' memory retrieval matrices (*Left*) and correlations between the model's hidden state representation similarity matrices (RSMs) and the human causal relationships (*Right*), which complement **Figure 4C** and **Figure 5C** *Left*, respectively. **(C)** The model is more comparable to the human behavioral data on the order of α = 0.25, 0.5, 0.75, 0, and 1. This indicates that the integration of the ongoing event representation and the retrieved memory is critical to the model's resemblance to human behavior. **(D)** The model's resemblance to the human behavioral data is highest at τ = 0.1, moderate at τ = 0.5, and lowest at τ = 1. This indicates that attending to selective memories by giving sparse attention weights is more human-like than broadly attending to a large number of memories.



Supplementary Figure S2. Comparisons between the EM-GRU and control models in how much model RSMs resemble brain region RSMs. The figure complements **Figure 8**, which reports paired *t*-test results comparing the correspondence between the brain RSMs and EM-GRU vs. GRU hidden state RSMs. Here, we show results for not only **(A)** hidden states but also **(B)** keys and **(C)** queries, and for control models including the *shuffled memory*, *no key*, and *no key-query* models. In nearly all cases, EM-GRU representations show higher correspondence with the human brain representations (indicated in red-yellow). The top panels display raw *t* statistics and the bottom panels show the same results after thresholding at FDR-corrected $p < .05$.