Response to *Nature Communications* Reviewers

Solving the Narrative Puzzle: Hippocampal Systems for Event Encoding and Sequencing during Ongoing Comprehension

Jiwoong Park, Hayoung Song, Won Mok Shim

We sincerely appreciate the constructive feedback from the reviewers, which has led to substantial revisions to our manuscript. A key revision is the shift in focus of the results: **1) we have removed the neural pattern similarity analyses,** which originally supported our claim about memory retrieval of past events during narrative processing. We agree with the reviewers that we did not directly ask participants to perform retrieval during narrative processing and that the results are open to alternative explanations. **2) In this revision, we have put more emphasis on and expanded the results of predictive modeling of narrative sequencing abilities**—the process of integrating related events that are separated in time into a coherent structure by linking past and present information within an unfolding narrative. This refinement enables us to concentrate on the novel findings of our study: **how the brain integrates temporally discontiguous events into a coherent narrative**, a critical question that has not been comprehensively addressed in previous research.

To strengthen this new focus, we have revised our methods, combining the conditions of forward and backward retrieval moments into a single condition of sequencing moments. This approach allows us to avoid comparisons based on unbalanced temporal characteristics between the conditions and ensures a balanced duration and offset across trials, improving the generalization of the results. Additionally, by building predictive models of event encoding and sequencing abilities based on a comparable number of timepoints, we provide more rigorous functional connectivity-based predictive modeling results. Finally, we have expanded our analysis of LLM-based moments of narrative coherence for predictive modeling of event sequencing, presenting substantially novel findings.

Together, these revisions comprehensively address the key concerns raised by the reviewers and enhance the clarity, focus, and interpretability of our findings, supporting a distinction between the two memory processes: hippocampus-mPFC based event sequencing and hippocampus-PMC based event encoding during ongoing narrative comprehension.

Please see below for a point-by-point response to the reviewers' comments.

Reviewer #1

Major Comments:

1. The main analyses are based on a comparison of seven forward sequencing boundaries to nine backward sequencing ones, and since a fixed scrambled order was used, the boundaries were not counterbalanced (ideally different orders would have been used such that the same boundary could be either forward or backward) and they appear to be imbalanced in terms of distance from the related event, when in the movie they were presented, etc. In this situation, it is not unlikely some of the results could arise from item effects. Theoretically a mixed-modeling approach could help alleviate some of these concerns, but I am not sure it would help given the very small number of trials. I appreciate the use of a large number of participants, but it does not fully compensate for the small number of events.

Response: We fully acknowledge the concern regarding potential item effects due to imbalances in the original forward vs. backward sequencing conditions. To address this issue, we have **removed the forward vs. backward distinction** and instead defined a **single, unified condition of sequencing moments**, which includes all instances where participants were expected to integrate past and current events into a coherent narrative. These moments have an equal duration and offset for all trials (Supplementary Figure 1A), ensuring consistency across conditions and eliminating potential confounds related to item effects.

---

2. The emphasis of the manuscript is on linking reactivation at boundaries to memory performance. Yet the demonstrated link is quite tentative, as it is only at the participant level, and it is not clear how the different levels used for analysis of behavior/brain data are related to one another. The ordering score seems to be at the level of meaningful events annotated by raters (~35 events), the neural pattern similarity at the level of segments (17, based on a subset of 9 backward sequencing events), and the NSP likelihood at the level of 2s segments (top 50). It would have been more convincing if a more direct link could be drawn, for example to demonstrate higher similarity for segments recalled in the correct order, at the same level.

Response: We appreciate the reviewer's concern regarding the indirect relationship between neural reactivation, sequencing, and memory performance. Given our revised focus on **narrative sequencing rather than memory retrieval**, we have **removed all neural pattern similarity analyses** from the manuscript. Instead, we emphasize the role of hippocampus-centered functional connectivity in predicting participants' ability to sequence events correctly. This change provides a **more robust and direct link** between neural measures and behavioral performance by using **functional connectivity-based predictive modeling approaches** (Shen et al., 2017) that are directly related to event sequencing ability.

---

3. The increased similarity between activity patterns at expected retrieval moments and related events (p. 6) is interpreted as reflecting retrieval. And as a control there is a comparison to unrelated events. As these are moments that were presented in immediate succession in the original film, and potentially visually and semantically similar (see comment 4), this could be due to similarity of the stimulus rather than retrieval.

Response: We agree with the reviewer that the previous neural pattern similarity results may have been due to semantic or perceptual similarities of the scenes rather than due to memory reinstatement. Therefore, we have removed the reactivation analyses in the revised manuscript. **Our study now primarily focuses on functional connectivity patterns during narrative sequencing moments**, rather than neural reactivation of past events. Importantly, the functional connectivity patterns were constructed using only the sequencing moments, not the previous events themselves. As a result, our findings are not influenced by low-level visual or semantic similarity between past and current events.

Minor Comments:

4. One of the basic assumptions of the manuscript is that out-of-place segments will elicit retrieval of the adjacent part of the preceding related segment. This could or could not be the case depending on the content of the film and the division to segments. Providing the segmented movie would be very helpful, and if it is not possible due to copyright issues – to elaborate on the transitions, and why retrieval is expected to be specifically of the beginning/end of the preceding segment for forward/backward sequencing.

Response: To clarify the rationale behind our event segmentation and sequencing moments, we have provided additional details in the Methods section (under "Stimulus" and "Experimental procedure"). Despite being temporally scrambled, the overall movie retained a **clear coherent underlying narrative structure with identifiable causal relationships**, requiring participants to reconstruct the original sequence during subsequent free recall. Given this task demand, we hypothesized that when encountering an out-of-place segment, participants would naturally link it to its previously viewed related event in an effort to integrate fragmented narrative elements into a coherent sequence. To further address this concern, we **can provide the scrambled video as supplementary materials** for clarification regarding the segmentation and transitions between events in our stimuli if requested.

---

5. How were the time-courses in Figure 1B calculated? Perhaps I missed this, but I could not find indication of whether this was percent signal change, z-scored signal and whether it was relative to a specific baseline.

Response: We have added details specifying that the time courses in **Supplementary Figure 1B** (previously **Figure 2B**) represent **z-scored signal change**, computed relative to a baseline defined as the **mean signal across the entire experiment**. We also acknowledge that the reviewer's comment may have been referring to Figure 2B in the original manuscript. In the **revised analysis**, hippocampal time courses are now used **exclusively for the purpose of identifying post-event boundary moments**, rather than for determining retrieval moments.

---

6. The conclusion that there is a higher response for backward sequencing before boundaries and for forward sequencing following boundaries does not seem to be

supported by the time-course in Figure 2B. It is true that while focusing on the specific time-points chosen that is the case, this appears to be due to an earlier response for forward sequencing rather than a stronger one. I am mentioning this as a minor comment since this is not one of the main conclusions of the paper, but it is quite critical for this specific result

Response: In the revised manuscript, we have removed **forward vs. backward** comparisons. Instead, we have consolidated these two conditions into a **single condition of sequencing moments**, ensuring a **balanced duration and offset across trials**. Our revised analyses now focus on **contrasting sequencing and encoding moments, which are closely matched in terms of their temporal profiles**.

---

7. It is not clear how the connectivity analysis was limited to specific timepoints. Were these timepoints extracted and then concatenated? Was the connectivity calculated within each window and then averaged? This would affect interpretation.

Response: We appreciate the reviewer's clarification on this point. In our analysis, we constructed a hippocampo-cortical FC model using the hippocampus as a seed and compared it to a cortico-cortical FC model without the hippocampus. FC was computed separately for three time periods: (1) **sequencing moments** (64 TRs across 16 sequencing pairs), (2) **post-event boundary moments** (68 TRs across 17 event boundaries); and (3) **all time points** (600 TRs from the movie plus 10 TRs of blank screen). For sequencing and post-event boundary moments, BOLD signals were first **concatenated across all relevant events**, and functional connectivity was then calculated **using Pearson's correlation on this single concatenated time series**.

This approach follows prior studies demonstrating that **concatenating time series prior to computing FC yields comparable or slightly improved reliability compared to averaging FC matrices calculated separately for individual windows** (Cho et al., 2021). Notably, Cho et al. found that the time-series concatenation strategy (Strategy-T) performed better than averaging separate FC matrices, particularly when using a smaller number of parcels, as in our study.

---

8. It would be helpful to provide details regarding distributions of the different segmentations, for example in the LLM-generated coherence which segments were chosen and how distant were they from their paired segment? Could there be consecutive segments? How much overlap was there between the NSP defined retrieval moments and the predefined ones?

Response: We appreciate the reviewer's request for further details on the distribution and characteristics of LLM-generated sequencing moments. In response, we have conducted **additional analyses to compare their temporal properties with those of pre-defined sequencing moments.** As shown in Supplementary Figure 7A, LLM-generated sequencing

moments had a comparable duration to pre-defined sequencing moments but exhibited slightly longer temporal gaps between consecutive sequencing moments and shorter temporal distances to their linked prior events. Additionally, we provide a quantitative breakdown of the overlap between LLM-generated and pre-defined sequencing moments (**approximately 15%, Supplementary Fig. 7B, left panel**), demonstrating that while the two approaches capture distinct moments, they converge on a subset of critical sequencing moments. Furthermore, LLM-generated sequencing moments were associated with significantly higher narrative coherence compared to other time points, reinforcing their relevance for identifying key moments of event sequencing (Supplementary Fig. 7B, right panel). Finally, as shown in Supplementary Figure 7C, hippocampus-vmPFC functional connectivity remained consistently predictive of ordering scores across varying numbers of LLM-generated sequencing moments, supporting the robustness of these results regardless of how sequencing moments were defined.

---

9. Given the significant discrepancy between the raters performing event segmentation (46 vs. 25), it would be preferable to add raters to reach a reliable segmentation, instead of averaging the two.

Response: We appreciate the reviewer's concern regarding the discrepancy in the number of event segmentations reported by the two raters. However, it is important to clarify that we did not directly use the event boundaries or the number of events determined by human raters in our analysis. Instead, **their evaluations were used to derive ordering and content scores**, which quantify participants' memory performance. Notably, **the two independent raters showed strong consistency in their scoring**, as evidenced by significant correlations between their ratings for both content and ordering scores (Ordering score: $r = 0.768$, $p < 0.001$; Content score: $r = 0.899$, $p < 0.001$, Supplementary Figure 2B). Given this high inter-rater reliability, we **averaged the ordering and content scores from the two raters** to obtain stable and representative measures for subsequent analysis.

Reviewer #2

Major Comments:

1. The paper introduces many novel analyses, each of which is quite complex. Although this is in some ways a strength of the paper, the authors could consider whether all of the different analyses are providing unique insights about the primary scientific questions being asked. Specifically, the language model analyses (Figure 6) are interesting since they can predict a continuous behavioral measure of narrative comprehension, but this seems unnecessary for this stimulus which is already structured to have easily-identifiable connections between related events. The neural results in this section appear to all be described as just comparable to / consistent with the results in the other figures; although this is an interesting validation of the language model approach, this line of work might be better served by being in a separate paper that directly looks at moment-by-moment narrative coherence (perhaps in a less-structured stimulus).

Response: We appreciate the reviewer's comment regarding the complexity of the analyses and whether all methods provide unique insights into the primary research questions. To streamline the manuscript and focus on event sequencing and functional connectivity-based predictive modeling, we have removed the neural reinstatement analysis. This revision ensures a clear focus on the primary question of our study, that is how hippocampus-centered systems support narrative sequencing.

Regarding the LLM-based approach, we have clarified its unique contribution in the revised Discussion section. While our stimulus follows a structured narrative, **our findings from LLM-generated sequencing moments reveal that periods of high narrative coherence**—rather than pre-defined structures—**can reliably identify moments of event sequencing mediated by hippocampus-vmPFC interactions, a method that extends to unmanipulated narratives**. This data-driven approach identified key moments of narrative construction that were not captured by pre-defined sequencing moments, reinforcing its utility in detecting critical time points for event sequencing. Notably, only approximately 15% of LLM-generated sequencing moments overlapped with pre-defined moments (Supplementary Fig. 7B), demonstrating that **the LLM successfully identified previously unidentified time points that were equally predictive of event sequencing performance.** These findings highlight the robustness and generalizability of hippocampus-vmPFC connectivity in supporting sequencing across different independent definitions of sequencing moments. While we acknowledge that LLM-based approaches could be further explored in a diverse set of structured and less-structured stimuli in future research, we believe that their application in the current study provides a critical validation of the role of narrative coherence in hippocampus-vmPFC interactions and strengthens our broader findings on memory sequencing.

---

2. The way that the "content score" is defined for participant recalls is difficult to interpret and not well-motivated. My understanding (from Supp Fig 4) is that a "movie content distribution" template is created that reflects, for each timepoint in the movie, the fraction of other timepoints that share this timepoint's topic. Then for each participant, a "recall content distribution" is defined as, for each timepoint in the movie, the fraction of recall timepoints that share this timepoint's topic. The similarity between these two distributions is computed using the inverse KL divergence. The content score therefore appears to reflect whether the proportions of topics in the recall reflect the true proportions of topics in the movie, not the "amount of correct semantic information retained from the movie" as the authors describe. A participant who accurately recalled every topic in the movie could receive a poor score on this metric, if their recall time for each topic was not proportional to its presentation time in the movie. I could not find this metric defined in one of the previous papers the authors cited; one of the cited papers, Heusser et al. (2021), instead computes a "precision score" by finding the maximum recall correlation for each movie topic and then averaging across topics, which to me is a much more intuitive definition of whether semantic content was successfully recalled.

Response: We appreciate the reviewer's thorough read of our method. To address the concern regarding the content score measurement, we have expanded our description in the Methods section. In our study, we measured the amount of semantic content retained in free recall using

the inverse of KL-divergence between the topic distributions of the movie and participants' recall. While Heusser et al. (2021) used an explicit event-matching approach (segmenting free recall using a Hidden Markov Model, then counting the number of recalled events), this method is not suitable for our study due to the higher difficulty of our recall task caused by a scrambled narrative. Specifically, participants were required to recall the movie in its original order, which resulted in shorter recall durations and fewer sentences per participant, making event segmentation challenging. Furthermore, in our data, single recall sentences often contained information spanning multiple events, preventing a clear one-to-one event mapping between recall and movie segments. Although KL-divergence assesses the proportions of recalled topics relative to those in the full movie, a lack of recall for specific events lowers the distribution, altering the KL-divergence value. Thus, **our approach captures how well participants retained the overall semantic structure of the movie**, rather than one-on-one precise retrieval of each event, and whether their recall contained a balanced representation of each narrative element. To further validate our approach, we compared the content scores derived from our model to human raters' assessments, where raters counted the number of correctly recalled events included in each participant's recall. We found a strong correlation ($r = 0.9$, $p<0.001$, Supplementary Figure 4B) between these two measures, demonstrating that **our content score reliably reflects the amount of event-related information present in participants' recall**.

---

3. The authors first examine the reactivation of related events around boundaries and then the connectivity between brain regions around boundaries, and find that the regions predictive of recall ordering are "predominately" the same. In the primary ROIs of interest, is there a quantitative relationship between the degree of reactivation and the degree of connectivity? Are these processes generally cooperative, as the authors suggest? In hippocampus and mPFC it appears that connectivity might be competing with reactivation, since participants with high ordering scores have high hippocampal and mPFC reactivation but low hippocampal-mPFC connectivity; could this competition be an alternative hypotheses for why this low connectivity is predictive of ordered behavior?

Response: We appreciate the reviewer's question regarding the relationship between reactivation and connectivity. In our previous analysis, defining neural patterns for reactivation required tailoring the reference patterns for previous events differently across ROIs due to their intrinsic timescales. This methodological complexity limited our ability to assess the relationship between reactivation and functional connectivity in a systematic and interpretable way. To address this **we have removed the reactivation analyses in the revised manuscript and instead focused on predictive modeling of functional connectivity patterns** during narrative comprehension. Our updated analysis directly examines how hippocampus-centered functional connectivity patterns predict individual differences in event encoding and sequencing ability, without relying on assumptions about underlying reactivation mechanisms. This approach allows us to focus on the dynamic coordination between the hippocampus and cortical regions, independent of neural similarity measures used in the previous version. In the revised manuscript, we have also expanded our discussion on why reduced hippocampal-mPFC connectivity predicts enhanced event sequencing abilities, which aligns with previous findings from both animal research and human neuroimaging studies.

4. The full cortico-cortical connectivity model in Fig 5A seems like a poor (and unnecessary) baseline for the hippocampo-cortical model. Because the full FC model uses 19,900 edges to predict ~60 recall scores, it is likely to be highly overfit, and it is not surprising that it does not perform well on a held-out subject. If the authors are intending to show that hippocampal connectivity is uniquely strong for predicting recall, an alternative approach would be to test against models that measure connectivity e.g. between vmPFC and all other regions, which would have the same complexity as the hippocampal connectivity model.

Response: We appreciate the reviewer's concern regarding the potential overfitting of the full cortico-cortical connectivity model. However, our predictive modeling approach incorporates a feature selection procedure before regression, ensuring that only functionally relevant edges are included in the final model. Specifically, we first select edges that are significantly correlated with behavioral scores ($p < 0.05$) and then compute two summary metrics by separately summing positive and negative FC strengths. This results in only two free parameters in the regression model, regardless of the initial number of edges. Consequently, **although the cortico-cortical FC model starts with a larger set of edges, the number of free parameters remains consistent across models, enabling a fair comparison without overfitting concerns.** This approach follows the cross-validated functional connectivity-based prediction framework established by Finn et al. (2015) and further validated in subsequent studies (Rosenberg et al., 2016; Shen et al., 2017), ensuring the robustness and generalizability of our findings.

To further validate the specificity of the hippocampus, we tested alternative models using other brain regions as seeds, matching the number of edges across all comparisons. These analyses confirmed that **hippocampal connectivity provided significantly superior predictive performance for both ordering and content scores compared to models using other seed regions.** Moreover, models excluding the hippocampus consistently exhibited lower performance across varying numbers of cortical ROIs. These findings, detailed in Supplementary Figure 6, highlight the distinct role of hippocampal connectivity in supporting narrative memory processes.

Minor Comments:

5. In figure one it seems like there are only 5 events in the story, and 2 instances of forward and backward sequencing. Looking at the supplementary materials made it clear that this is not the case, but it could be helpful to show how many events there are in the first figure.

Response: We appreciate this feedback and have revised **Figure 1's caption to clearly indicate that it illustrates only five event segments for visualization purposes**, while the full narrative contains seventeen segments. The caption now writes: "The figure shows five event segments for illustration; the complete narrative contained seventeen segments." To

provide further context, we have also directed readers to the Methods section and Supplementary Figure 1 for a detailed explanation of the scrambling procedure and sequencing moment specifications.

---

6. The framing of the introduction seems to focus on what is required to understand a narrative, but a lot of the processes described are relevant more generally for real-life event perception and experience, which could be discussed a little more.

Response: We appreciate the reviewer's suggestion to broaden the framing of our introduction. As the reviewer pointed out, while our study specifically investigates narrative comprehension, the cognitive mechanisms we examine—event encoding and sequencing—are fundamental to real-life event perception and experience more broadly. To clarify this, we have revised the Introduction and Discussion section to highlight the broader relevance of these processes beyond structured narratives. Specifically, **we now emphasize how natural experiences unfold as open-ended and uncertain sequences, requiring continuous updating and integration of newly encountered information into coherent representations**. Our findings demonstrate how distinct memory systems dynamically support the integration of complex, naturalistic information into meaningful narratives, providing insights into the neural mechanisms underlying real-world memory processes.

---

7. The way that the fMRI representation of previous events is determined (Supp Fig 2) seems very complex and is difficult to follow. Does "offset from event boundary" refer to the start of the previous event for backward sequencing and the end of the previous event for forward sequencing (as implied by Supp Fig 2A)? Is there a simple explanation for why the start of the prior event is not a good template for the backward sequencing analysis (it seems like it should be the best template, since this is the part of the event that directly connects to the new event)?

Response: Since we have removed neural pattern similarity analyses in the revised manuscript, this concern is no longer applicable. However, we have ensured that our updated analysis clearly defines how functional connectivity (FC) patterns were constructed. Specifically, in our revised approach, **sequencing moments were selected with equal duration and offset across all trials**. By aligning sequencing moments with the temporal structure of post-event boundaries, we have ensured that the **comparison between sequencing and encoding moments is balanced.**

---

8. Subpanel Fig 4D seems like an unintuitive place to put this analysis - this is a purely behavioral measure that could have been presented earlier, rather than being included in a figure about neural representations.

Response: Since we have removed the analysis comparing forward vs. backward retrieval in the revised manuscript, this concern is no longer applicable. Additionally, **we have reorganized the figures to present behavioral analysis results first, followed by neural**

analyses, ensuring a logical progression from cognitive measures of narrative understanding to their corresponding neural mechanisms.

---

9. What is the statistical test in Supp Fig 6B meant to show? My understanding is that this test is showing that edges that tend to get selected for inclusion in the model due to their relationship with the ordering score are, on average, more related to the ordering score - is this surprising?

Response: In Supplementary Figure 6 of the original manuscript, we aimed to demonstrate that regions with stronger neural pattern similarity correlations with ordering scores were more likely to be selected as predictive edges in the functional connectivity-based model, highlighting a potential link between neural reactivation and functional connectivity in narrative sequencing. Since we have removed the neural pattern similarity analysis in the revised manuscript, this concern is no longer applicable. Accordingly, Supplementary Figure 6B has been removed, and our updated analyses now focus exclusively on functional connectivity-based predictive models.

---

10. Code is provided for plotting the figures, but in some cases (e.g. computing the Next Sentence Probability values) this code is reading from datafiles that contain the results of analyses, and neither the datafiles nor the code to produce them are provided.

Response: We appreciate the reviewer's comment regarding code availability. To improve transparency and reproducibility, we have now included data files and additional scripts that generate the NSP values from the raw text data using the large language model (LLM). Additionally, we have provided detailed documentation on how these analyses were performed, including code to reproduce the key computational steps leading to the results. Furthermore, **we have verified that our analysis code runs successfully in an independent environment** with all dependencies specified and all necessary resources available in the code and data repository (https://github.com/jwparks/NarrativePuzzle).

Reviewer #3

Major Comments:

1. There is insufficient operationalization and limited ability to directly measure the targeted memory processes in this study. The issue starts in the introduction in which the authors propose two processes, encoding and "establishing relationships." It is unclear what is meant by the latter process. The authors then mention retrieval and integration processes. Encoding, retrieval, and integration are three separate processes (e.g. Richter et al., 2016 NeuroImage), thus they cannot be used interchangeably as they seem to be here. Even if this issue were rectified, however, the issue remains that the authors are not directly

measuring any of these processes. Throughout the manuscript, the authors assume that their findings -- hippocampal activity, similarity measures, connectivity, etc. -- reflect retrieval, but this is not necessarily the case. All of the reported effects may instead reflect encoding rather than retrieval. Thus it is not clear that the objective of the proposal - to identify distinct processes that support narrative comprehension - can be achieved in the current study. Nor is it clear that the challenge of "precisely identifying the specific moments when memory retrieval and integration processes occur" (page 19) has been or can be addressed (especially considering the temporal limitations of fMRI). Participants are assumed to retrieve and integrate information in order to perform the ordering task, but this need not happen during the movie itself. Inferences could be made during the test phase. Relatedly, the authors quantify "integration difficulty" as "the difference in the amount of recalled content between the two events"; however, this operationalization is not well justified.

Response: We appreciate the reviewer's thoughtful comments regarding the clarity of the memory processes investigated in our study. We recognize the importance of distinguishing between encoding, retrieval, and integration, and in response, we have refined our theoretical framing to improve clarity. Specifically, **we have revised the manuscript to use "event sequencing" as the primary term, rather than separately discussing retrieval and integration**. By event sequencing, we refer to the process by which past and current events are linked in a structured order to form a coherent narrative representation. While event sequencing may involve retrieval of past information, our study does not assume that retrieval is the sole mechanism at play. Instead, we focus on how the hippocampus and cortical regions dynamically interact to sequence events during ongoing comprehension. This construct more accurately reflects the experimental design, which examines how neural systems support the organization of temporally fragmented events into a coherent sequence.

Additionally, we acknowledge the reviewer's concern regarding the operationalization of memory processes. While it is challenging to isolate retrieval, integration, or encoding within the constraints of fMRI, **we incorporate two complementary behavioral metrics that capture distinct aspects of narrative memory: ordering score and content score.** The ordering score reflects participants' ability to recall events in the correct temporal sequence, serving as a direct measure for event sequencing ability. The content score quantifies the amount of semantic information retained from the movie, providing a robust assessment of event encoding. Both scores vary across individuals, allowing us to assess how hippocampal functional connectivity predicts differences in narrative comprehension. We have now expanded our discussion of these behavioral measures and their relevance to distinguishing encoding and sequencing processes in memory.

Furthermore, we have removed the concept of "integration difficulty," which was previously assessed through memory reinstatement analysis based on neural pattern similarity. Instead, our revised manuscript focuses on predictive modeling using distinct moments of expected sequencing and post-event boundary moments. By employing ordering and content scores as behavioral measures, we directly evaluate how hippocampal functional connectivity supports event sequencing and encoding processes.

Finally, we appreciate the reviewer's concern regarding the timing of retrieval and integration processes. While it is possible that some inferences about event order occur during the recall phase, prior research suggests that retrieval and integration processes can also occur during movie viewing. Notably, Chen et al. (2016) demonstrated that hippocampal activity becomes more similar across participants when they encounter distant events that relate to previously viewed content, suggesting a shared process of memory retrieval and integration. Also, Hahamy et al. (2023) suggested that humans reactivate neural representations of past events while understanding a narrative. **These findings support the hypothesis that participants actively link related but temporally separated events during narrative perception, rather than relying solely on post-hoc reconstruction during recall.** In our study, we identified expected sequencing moments—time points when past and current events are likely integrated—and showed that hippocampus-vmPFC connectivity during these moments predicted individual differences in ordering performance. These results suggest that memory sequencing occurs dynamically during movie viewing as participants continuously construct and update their evolving narrative models.

---

2. The rationale for the vmPFC/PMC connectivity analyses is straightforward, yet (1) there are additional analyses that are not part of these predictions, the reasoning for which is not clear and (2) there are no direct tests to support the repeated claims of having found "distinct" systems. First, much is made regarding functional connectome models with and without the hippocampus. That a model with the hippocampus performs better than a model without is reassuring, but does not seem related to the specific predictions regarding hippocampal connectivity with vmPFC or PMC. Second, the authors claim that functional connectivity between hippocampus and PMC "played a key role in predicting content scores" (page 13) yet there are no direct tests and no evidence is provided that connectivity between hippocampus and vmPFC/PMC is specifically related to later memory. The only tests reported are those that confirm the authors predictions (vmPFC and order, PMC and content) with no direct tests across memory performance metrics (e.g. vmPFC order vs. content) or across regions (e.g. hippocampal connectivity with another candidate region). There are additionally some confusing interpretations made regarding the strength of connectivity. The authors find a negative correlation between hippocampus and vmPFC which they describe as functional "decoupling" and then state that this is consistent with prior work demonstrating "strengthened" hippocampus-vmPFC interactions during encoding (page 14). Decoupling does not seem to be consistent with strengthened interactions.

Response: We appreciate the reviewer's concerns and have conducted **additional statistical tests to directly compare hippocampus-vmPFC and hippocampus-PMC functional connectivity in predicting ordering and content scores**. We demonstrated that hippocampus-vmPFC connectivity was exclusively associated with ordering scores (100% of cross-validation folds; $p<0.001$) and not content scores (0%), whereas hippocampus-PMC connectivity was predominantly linked to content scores (87.6%, $p<0.001$) and not ordering scores (0%).

Furthermore, hippocampus-vmPFC connectivity negatively correlated with ordering scores ($r=-0.37$, $p<0.01$) but was unrelated to content scores ($r=0.03$, $p=0.77$), whereas

hippocampus-PMC connectivity showed the opposite pattern, correlating with content scores ($r=0.26$, $p<0.05$) but not ordering scores ($r=0.03$, $p=0.81$). Using Steiger's Z-test to directly compare correlation strengths, we confirmed that hippocampus-vmPFC FC was significantly more associated with ordering than content scores ($z(62)=2.614$, $p<0.01$), while no significant difference was observed for hippocampus-PMC FC ($z(62)=1.436$, $p=0.15$). Additionally, we clarified that hippocampus-vmPFC decoupling aligns with prior studies suggesting reduced synchronization when integrating new events into existing schemas rather than encoding novel information. Together, **these results provide strong evidence for functionally distinct hippocampus-centered systems supporting event sequencing and encoding**. These results, now included in Figure 5C, provide statistical evidence supporting distinct memory systems.

---

3. The presumed superiority of LLMs for assessing narrative coherence is not suitably justified. First, it is not clear why pre-defined narrative structures are a limitation. Second, it is not clear how the LLMs performance was validated if not based on these pre-defined structures. Third, it is not evident that the LLM outperforms when compared to predictions based on pre-defined narrative structures. Finally, it is not clear how or why LLM-generated narrative coherence is related to retrieval.

Response: We appreciate the reviewer's request for providing justification regarding the use of large language models (LLMs) in assessing narrative coherence. Below, we provide point-by-point responses to each concern:

1) **Why are pre-defined narrative structures a limitation?**
Pre-defined sequencing moments are based on the original chronological order of the narrative, assuming that event integration occurs at these moments due to the explicitly segmented and scrambled structure. However, **narrative comprehension is a dynamic process, and participants may naturally link events based on their temporal or causal relationships.** For example, in our narrative stimulus, scenes involving a thief and a police officer are presented separately, yet participants may infer their connection based on the common expectation that the police will eventually catch the thief. By using an LLM to identify sequencing moments based on narrative coherence rather than pre-defined moments based on experimental manipulations, we implement a data-driven approach that captures naturally occurring moments of event integration beyond artificially imposed boundaries, potentially offering greater ecological validity for studying natural narratives without experimental manipulations

2) **How was LLM performance validated if not based on pre-defined structures?**
To ensure the validity of LLM-generated sequencing moments, we compared them to pre-defined sequencing moments and examined their predictive power in modeling narrative memory. Despite only 15% overlap between the two sets of moments (Supplementary Fig. 7B), both sets yielded comparable prediction accuracy for ordering scores, demonstrating that our LLM approach successfully identifies critical event sequencing moments, which were not identified based on pre-defined structures. Additionally, we validated the LLM's ability to detect high-coherence moments by showing that hippocampus-vmPFC functional connectivity

decreases at these moments, a neural signature robustly associated with memory integration in previous studies (CITE).

3) **Does the LLM outperform pre-defined narrative structures?**
Our results demonstrated that **LLM-generated sequencing moments outperformed pre-defined sequencing moments when a smaller number of high-narrative-coherence time points were used** (e.g., Top 20 moments, Figure 5). This suggests that event sequencing occurs selectively at key narrative junctures rather than uniformly across all transition points with the highest concentration at moments where narrative coherence peaks. The LLM effectively identifies these critical moments, providing a precise selection of event sequencing moments. Additionally, **LLM-generated sequencing moments yielded more robust predictive accuracy across varying cortical parcellations**, highlighting their stability compared to pre-defined sequencing moments (Supplementary Figure 5).

4) **How is LLM-generated narrative coherence related to retrieval?**
We appreciate the reviewer's question regarding the relationship between LLM-generated narrative coherence and memory retrieval. Large language models (LLMs) trained with the Next Sentence Prediction (NSP) task are designed to process long-range contextual dependencies (Devlin et al., 2018), allowing them to predict whether one sentence naturally follows another. This ability enables **the model to identify moments in a narrative where event transitions are highly predictable based on prior context, aligning with the cognitive processes underlying event sequencing**. Furthermore, previous research has demonstrated that encountering semantically or causally coherent events increases the likelihood of narrative comprehension (Song et al., 2021), memory retrieval and integration (Cohn-sheehy et al., 2021). Studies have also shown that structured narratives facilitate hippocampal pattern reinstatement (Chen et al., 2016) and that schema-congruent events promote retrieval-based integration (van Kesteren et al., 2012; Schlichting & Preston, 2015). Based on these findings, we propose that moments with high narrative coherence, as identified by the LLM trained with the NSP task, correspond to time points when event sequencing is most likely to occur.

We have clarified these points in the revised manuscript, particularly in the Discussion section, to better contextualize the role of LLMs in our study.

---

4. The selection procedure for some parameters is underspecified and as written seem arbitrary. For instance, the decision to select 6s for backward sequencing events vs. 4s for forward sequencing events or the decision to select the time period with the "second highest" similarity. Relatedly, the selection of "all events preceding the current backward or forward sequencing even in the scrambled movie" will artificially decrease pattern similarity for unrelated stimuli given both temporal autocorrelation and scanner drift.

Response: We appreciate the reviewer's concern regarding the selection of time windows for different conditions. In the original manuscript, forward and backward retrieval moments were separate conditions. In the revised manuscript, we no longer differentiate between these conditions and instead use a single, unified condition of sequencing moments with temporal

parameters established through a consistent selection procedure across trials. **The duration of sequencing moments was specifically chosen to match post-event boundary moments,** providing a balanced number of time points for functional connectivity analysis. Additionally, we have clarified our selection criteria in the Methods section: Post-event boundary moments were determined based on the period when significantly elevated hippocampal activity was shown following event boundaries, and sequencing moments were aligned with this temporal structure (maintaining identical duration) to maintain consistency (Supplementary Fig. 1A, 1B).

Conclusion

We are grateful to the reviewers for their insightful comments, which have significantly improved the manuscript. By shifting our focus to narrative sequencing, refining our functional connectivity-based predictive models, and removing analyses with potential confounds, we provide a more coherent and rigorous framework for understanding the distinct hippocampal-cortical memory systems underlying event encoding and sequencing. We believe these substantial revisions have strengthened the manuscript's contribution and improved its clarity, interpretation, and potential impact.

1. Shen, X. et al. Using connectome-based predictive modeling to predict individual behavior from brain connectivity. Nat Protoc 12, 506–518 (2017).
2. Cho, J. W., Korchmaros, A., Vogelstein, J. T., Milham, M. P. & Xu, T. Impact of concatenating fMRI data on reliability for functional connectomics. Neuroimage 226, 117549 (2021).
3. Heusser, A. C., Fitzpatrick, P. C. & Manning, J. R. Geometric models reveal behavioural and neural signatures of transforming experiences into memories. Nat Hum Behav 5, 905–919 (2021).
4. Finn, E. S. et al. Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. Nat. Neurosci. 18, 1664–1671 (2015).
5. Rosenberg, M. D. et al. A neuromarker of sustained attention from whole-brain functional connectivity. Nat. Neurosci. 19, 165–171 (2016).
6. Chen, J. et al. Accessing Real-Life Episodic Information from Minutes versus Hours Earlier Modulates Hippocampal and High-Order Cortical Dynamics. Cereb. Cortex 26, 3428–3441 (2016).
7. Hahamy, A., Dubossarsky, H. & Behrens, T. E. J. The human brain reactivates context-specific past information at event boundaries of naturalistic experiences. Nat. Neurosci. 26, 1080–1089 (2023).
8. Song, H., Park, B.-Y., Park, H. & Shim, W. M. Cognitive and Neural State Dynamics of Narrative Comprehension. *J Neurosci Official J Soc Neurosci* **41**, 8972–8990 (2021).
9. Cohn-Sheehy, B. I. *et al.* The hippocampus constructs narrative memories across distant events. *Curr Biol* **31**, 4935-4945.e7 (2021).
10. van Kesteren, M. T. R., Ruiter, D. J., Fernández, G. & Henson, R. N. How schema and novelty augment memory formation. *Trends Neurosci* **35**, 211–219 (2012).
11. 1.Schlichting, M. L. & Preston, A. R. Memory integration: neural mechanisms and implications for behavior. *Curr Opin Behav Sci* **1**, 1–8 (2015).

Sincerely,
Jiwoong Park (jiwoongpark@skku.edu)
Hayoung Song (hayoung@wustl.edu)
Won Mok Shim (wonmokshim@skku.edu)