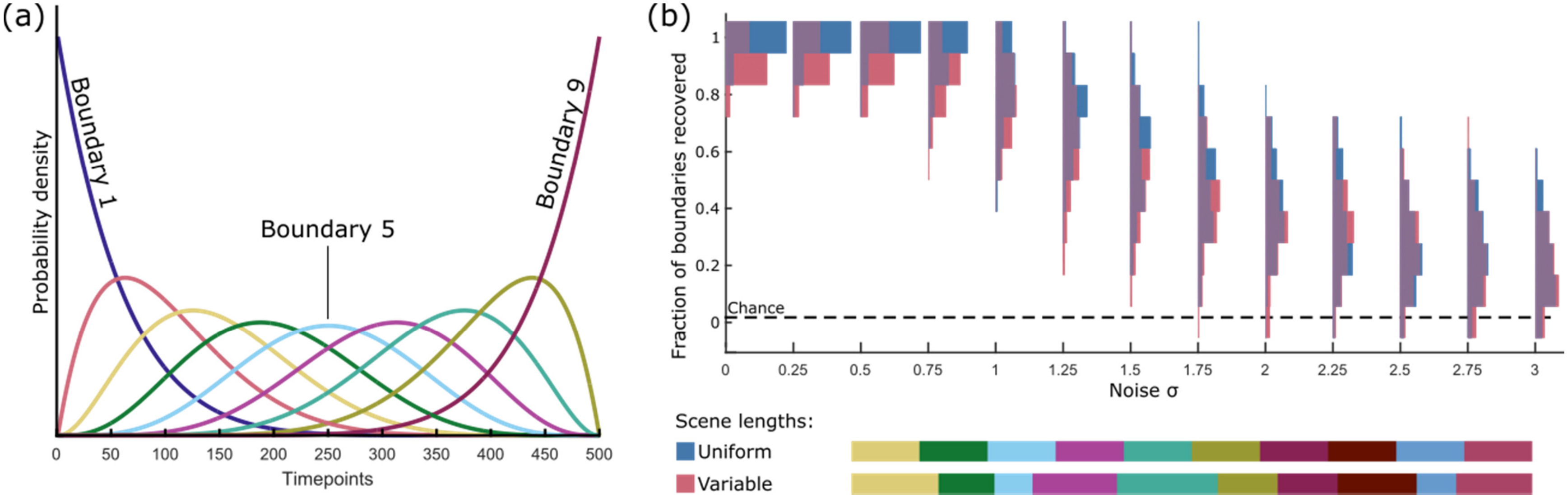

Figure S1: **Properties of the event segmentation model (related to Fig. 2).** (a) Our event segmentation model defines a uniform prior over all possible event segmentations in which every event occurs for at least one timepoint and all events occur in order. This induces a prior distribution over event boundaries, which depends on the number of timepoints T and the number of events K (T=500, K=10 in this figure). During the annealing process, the distribution of boundaries starts at this prior, which allows for a (highly uncertain) first estimate of the signature neural pattern for each event. Based on these patterns, the latent events for all timepoints are refit, and then the patterns are recalculated. The process continues, with the pattern variance slowly decreasing, until the log likelihood reaches a peak. (b) Simulated data with a discrete event structure obscured by varying levels of noise was input to the segmentation model, with T=500, K=10, and V=10. The model successfully recovers a majority of the underlying event boundaries at low noise levels, and can still identify an above-chance fraction of boundaries even at high noise levels that are as large as the differences between the event patterns. Having variable event lengths leads to only a small loss in performance, and does not change the overall performance curve.

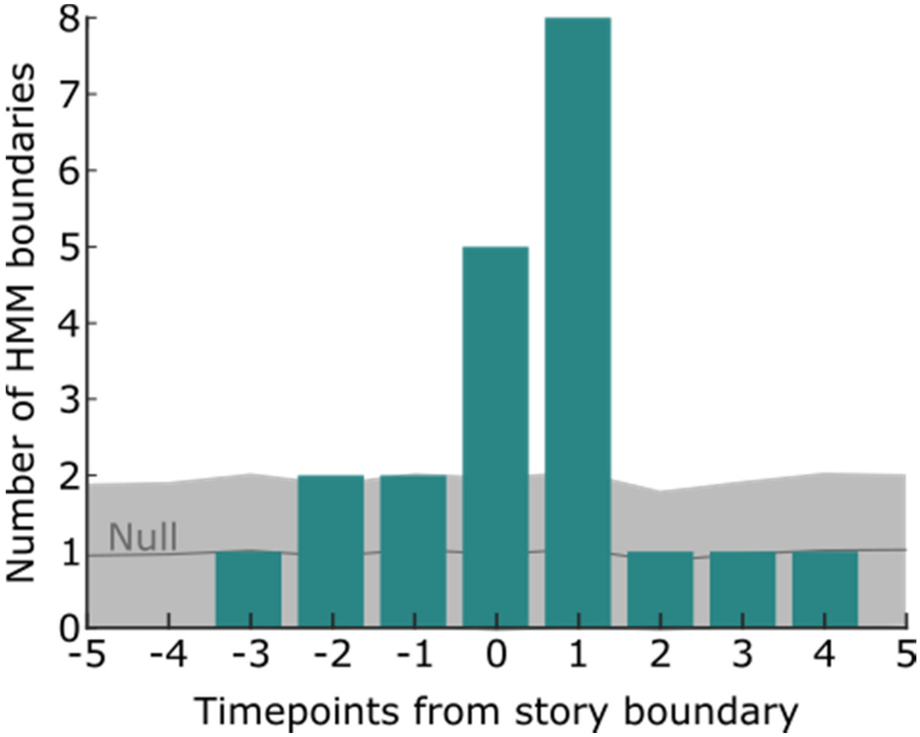

Figure S2: **The event segmentation model successfully identifies switches between stories (related to Fig 2).** Subjects listened to two stories, which were interleaved such that they alternated back and forth about every minute. Using data from PCC, an event segmentation model with 34 event transitions showed the best fit to held-out subjects (very close to the actual number of 32). Of these 34 transitions, the majority (20) were within 3 timepoints of a story switch. A null distribution was created by permuting the order of the events (preserving event lengths); under this null distribution the chance of having this many event boundaries close to true story switches was p<0.001.

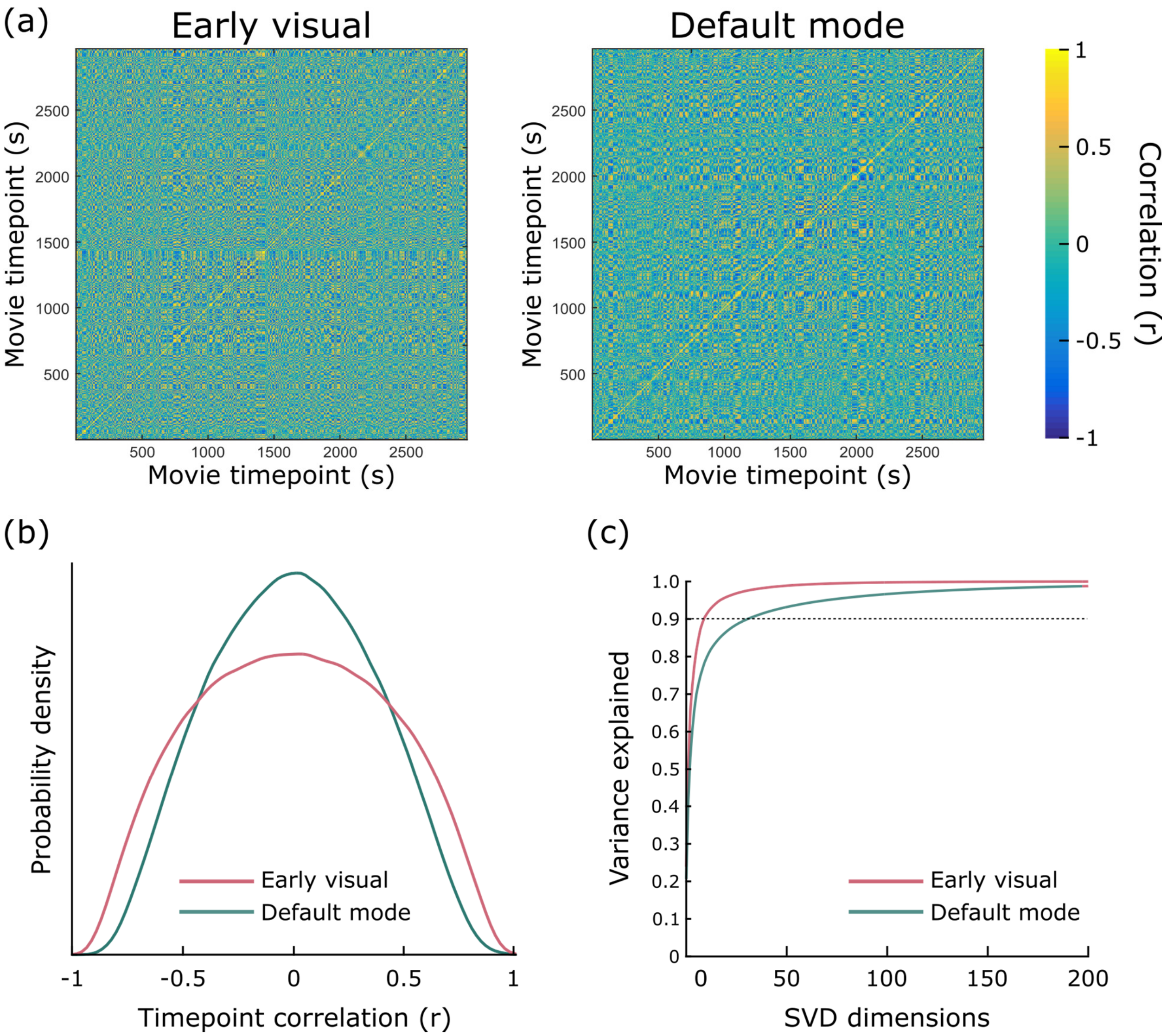

Figure S3. **Timepoint similarity matrices (related to Figs. 2, 3).** (a) For each pair of timepoints in the group average response to the 50-minute movie, similarity between activity patterns was assessed using Pearson correlation (in early visual cortex and default mode regions, angular gyrus and posterior cingulate cortex). These matrices exhibit blocks along the diagonal (characteristic of stable event patterns) as well as substantial off-diagonal structure. (b) These matrices both exhibit a range of correlation values, but the correlations are more concentrated near +/-1 in the early visual cortex. (c) Using SVD to estimate the dimensionality of the group timecourses, we find that capturing 90% of the variance requires 10 dimensions in visual cortex and 32 in default mode regions.

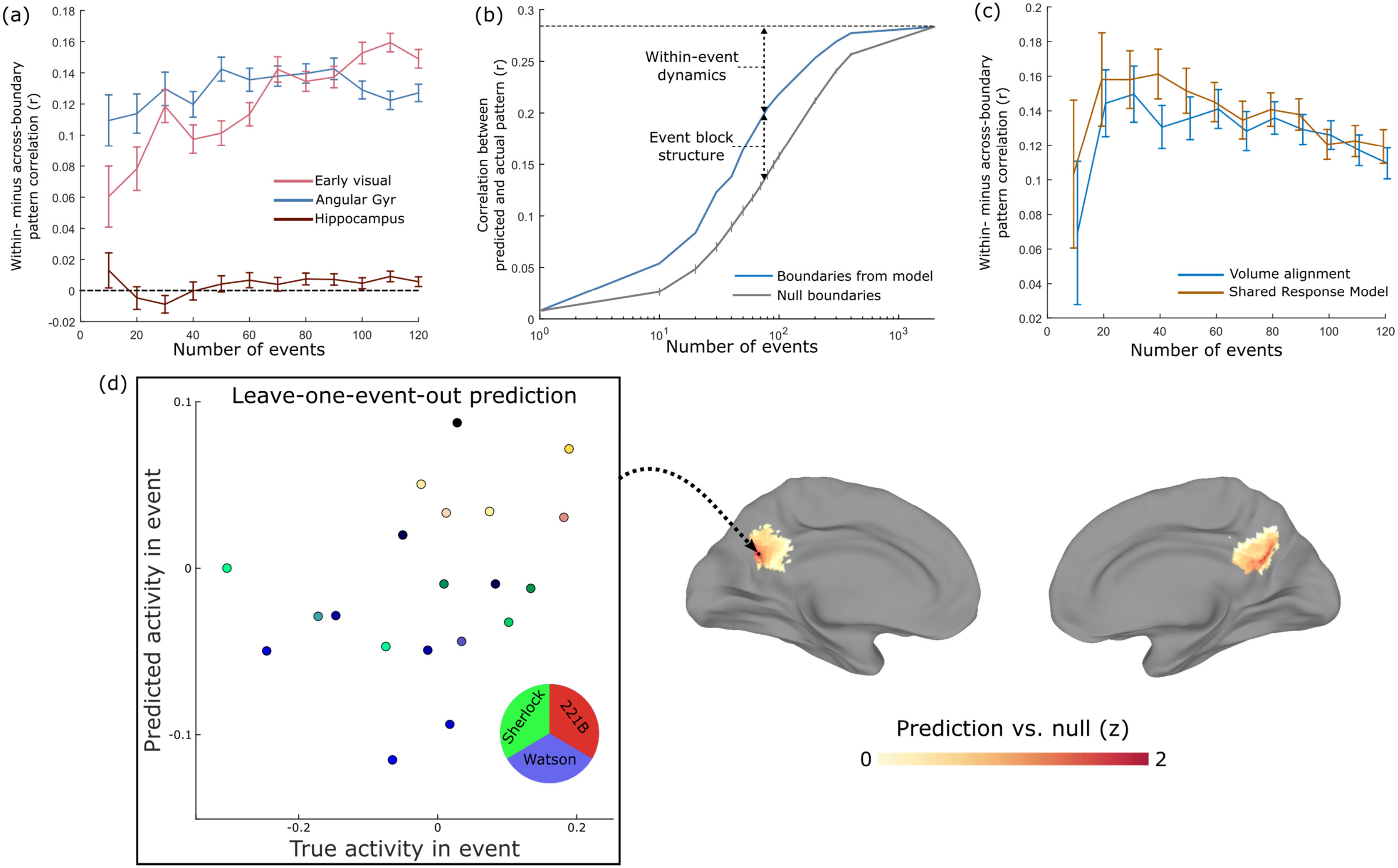
(a)
Within- minus across-boundary pattern correlation (r)
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
-0.02
0
20
40
60
80
100
120
Number of events
Early visual
Angular Gyr
Hippocampus
(b)
Correlation between predicted and actual pattern (r)
0.3
0.25
0.2
0.15
0.1
0.05
0
10^0
10^1
10^2
10^3
Number of events
Within-event dynamics
Event block structure
Boundaries from model
Null boundaries
(c)
Within- minus across-boundary pattern correlation (r)
0.2
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
20
40
60
80
100
120
Number of events
Volume alignment
Shared Response Model
(d)
Leave-one-event-out prediction
Predicted activity in event
0.1
0
-0.1
-0.2
0
0.2
True activity in event
Sherlock
221B
Watson
Prediction vs. null (z)
0
2

Figure S4. **Additional analyses of event structure evoked by the *Sherlock* movie (related to Fig 3).** (a) Early visual cortex and angular gyrus show similar goodness-of-fit (measured based on boundary prediction in a held-out subject) but have a different optimal number of events, with angular gyrus peaking around 50-90 events and visual cortex peaking around 110 events. The hippocampus did not show any significant event-structured activity patterns. Error bars indicate standard deviation of the null (scrambled-boundary) distribution. (b) To quantify the variance captured by the event segmentation model, we segmented the movie responses (in angular gyrus) from all but one of the subjects and used the average event activities to predict the responses in the held-out subject. At the optimal timescale (~75 events), this prediction is substantially better than a null model (in which the boundaries are permuted with the same event length distribution), indicating the existence of event structure at this timescale (error bars indicate 95% CI). There is also a gap between the event-structured prediction and a model-free prediction based on the continuous average of all other subjects (dashed line, equivalent to setting the number of events equal to the number of timepoints), indicating continuous dynamics on the scale of ~5 TRs that are predictable across subjects but not tied to specific boundary timepoints (since they are also captured by the null boundaries). (c) To determine whether our results were sensitive to way alignment was performed across subjects, we used the first half of the movie (from the angular gyrus) to fit a Shared Response Model (SRM) with 50 features (Chen et al., 2015). The SRM seeks to maximize functional correspondence across subjects by learning a projection matrix for each subject, from their native voxel space to a shared, low-dimensional space. We then ran the event segmentation procedure on the second half of the movie, both in the original volume space and in the shared response space (using the learned projection matrices). In both cases significant event structure was found, peaking at 20-40 events in both cases (for the last 25 minutes of the movie), with even stronger effects using SRM. Error bars indicate standard deviation of the null (scrambled-boundary) distribution. (d) To investigate whether the model-identified event patterns were related to event content, we fit a linear model to predict voxel activation for each event based on human annotations. *Sherlock* was divided into 1000 fine-grained segments (approximately 3 seconds each), and each was annotated with its location and the characters that were visible in the shot. The three annotations with the highest variance were Sherlock Holmes, John Watson, and 221B Baker Street (Sherlock and John's apartment). Each of 20 model-identified events in posterior cingulate cortex was assigned a value between 0 and 1 for each of these three annotations, calculated as the fraction of segments in which each of these three annotations was present. Using linear ridge regression, we learned to predict the activity of each voxel from these annotations using all but one of the events, and then tested our regression model on a held-out event. A prediction for an example voxel is shown in the box on the left, where each point indicates one event and has an RGB color corresponding to its three annotations (note that this voxel was selected post-hoc for illustration). Prediction performance showed a gradient throughout the region; splitting along the long axis of PCC, we found above-chance prediction performance (coefficient of determination greater than zero for held-out events) for 92.9% of voxels in the posterior/inferior half (significantly better than the permuted model, which showed 26.0% predictive voxels on average, p=0.023) and 46.2% of voxels in the anterior/superior half (not significantly different from the permuted model, which showed 25.1% predictive voxels on average, p=0.23). For a more detailed encoding-model analysis of information represented in this region, see J. Chen et al. (2017, Figure S6).

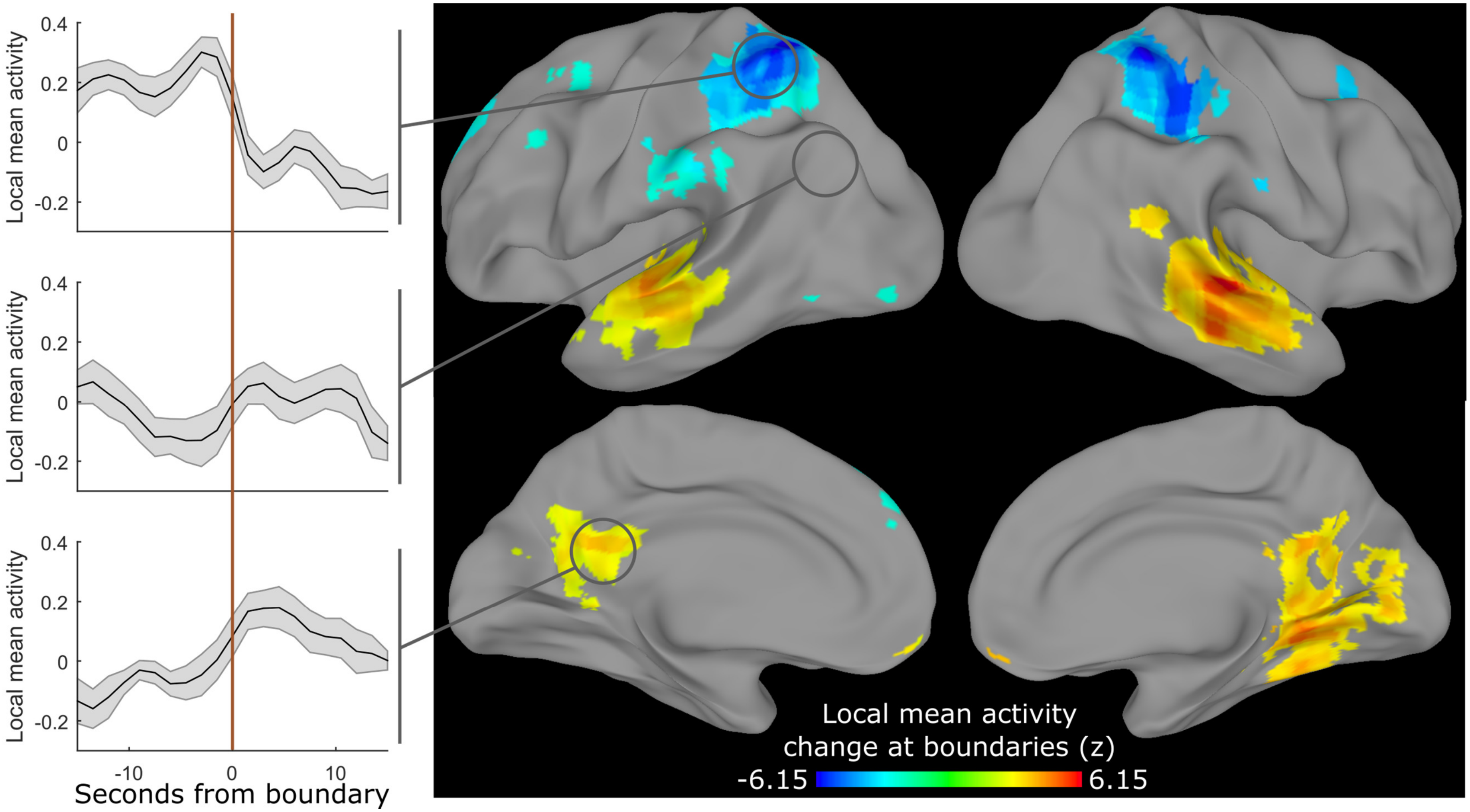

Figure S5. **Relationship between changes in activity patterns and local univariate activity (related to Fig. 6).** In Fig. 6, we identified regions whose event boundaries were related to increases in hippocampal activity. To determine if this effect could be driven by a relationship between multivariate event boundaries and *local* univariate activity, here we identified event boundaries for each searchlight using all but one subject (at that searchlight's optimal timescale) and then measured changes in the mean univariate activity of this searchlight around these boundaries in the held-out subject. Interestingly, there was no simple relationship between multivariate changes and local univariate activity, with some regions such as posterior cingulate cortex (bottom inset) showing *increases* in activity after boundaries, others such as the superior parietal lobule (top inset) showing *decreases* in activity, and others such as angular gyrus (middle inset) showing no significant change. Note that the searchlight insets were selected post-hoc for illustration.

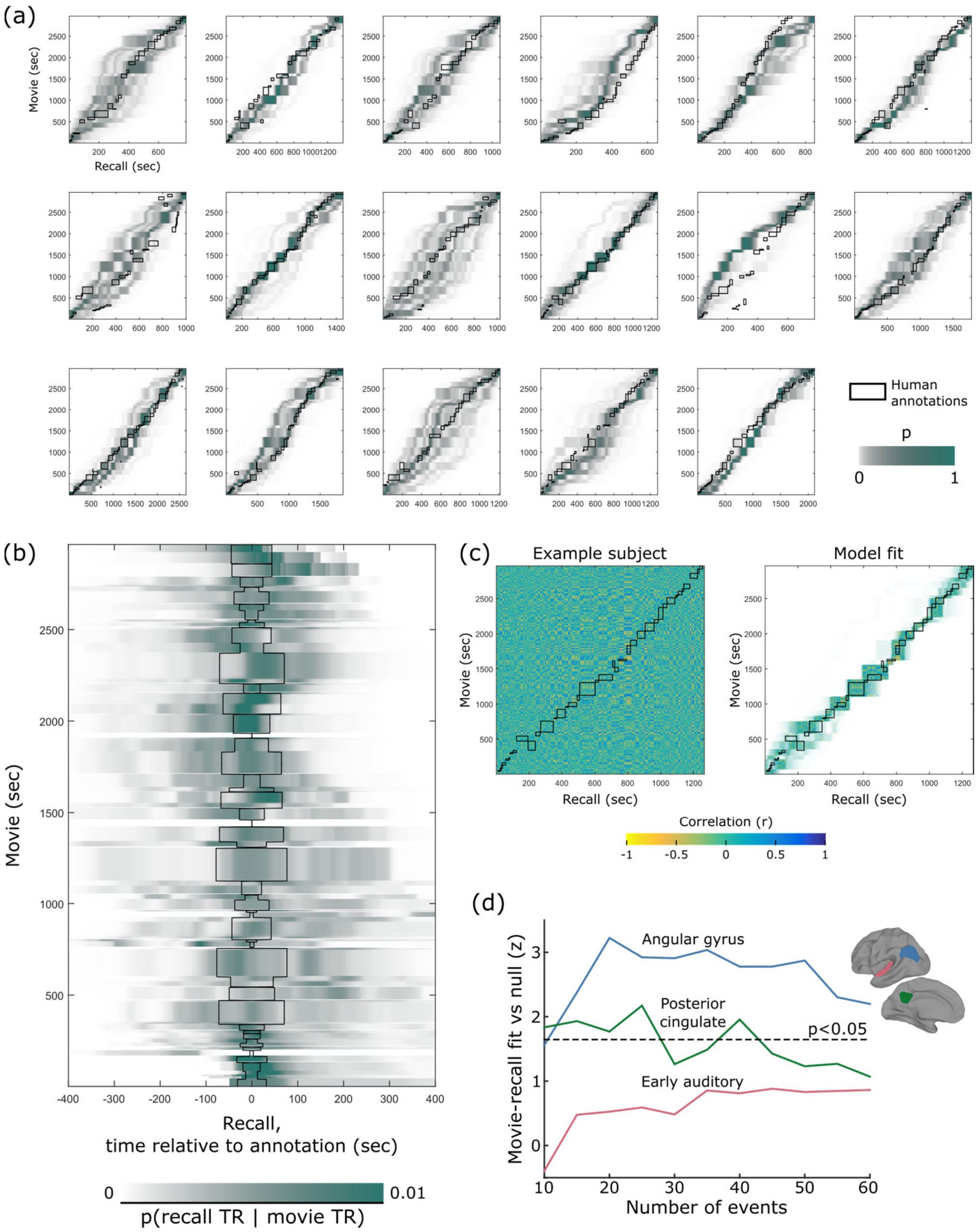

Figure S6: **Additional analyses of movie-recall correspondence (related to Fig. 7).** (a) The correspondence between movie-watching and free recall for each subject was estimated using the event segmentation model (using a combined ROI of angular gyrus and posterior cingulate cortex). Black boxes indicate the correspondence annotated by human observers. (b) To combine results across subjects, we re-centered each row of these correspondences relative to the human annotations and then averaged across subjects (for each movie timepoint, only subjects that recalled that timepoint of the movie were included in the average). Black boxes indicate the maximum duration of the recall for each movie scene, across subjects. This group correspondence shows that the model places the bulk of its probability mass in or near the human annotations, despite not using annotations during the fitting process. (c) When the timepoint-timepoint correlation matrix (left) is masked by the model-identified correspondence (right), we observe that the model selects a path with positive correlation values. (d) The results shown in main Fig. 6b hold for most choices of the number of latent events between 10 and 40, with decreasing goodness-of-fit for larger numbers of events. Note that the best fits were achieved with models having approximately 20-25 events, similar to the minimum number of human-labeled events recalled by the subjects (24, table S1 in J. Chen et al., 2017).

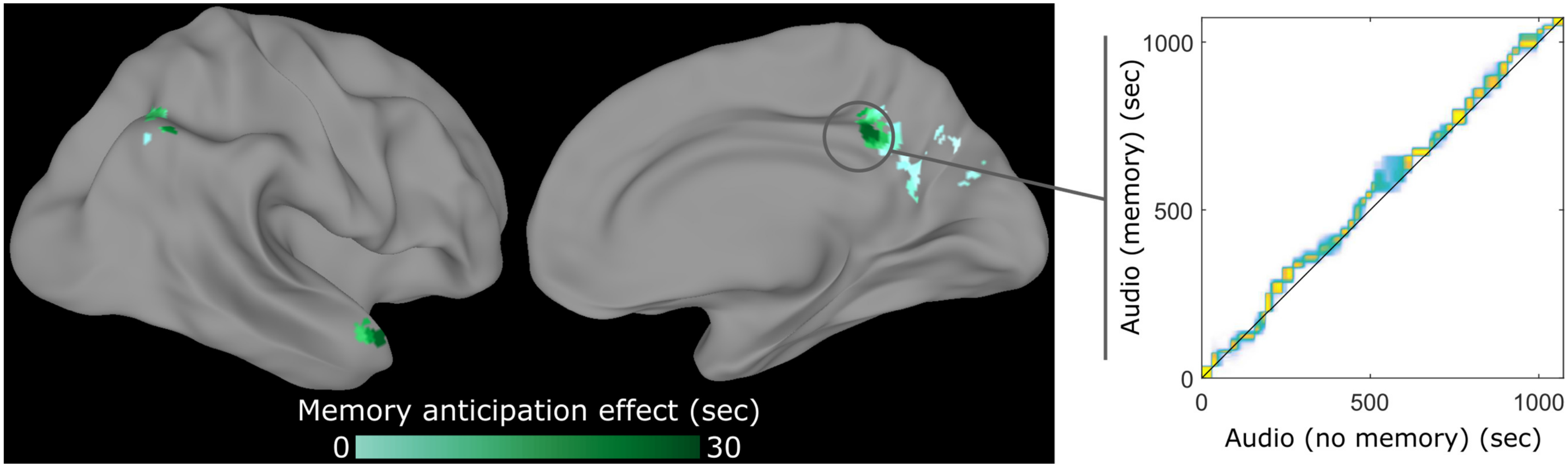

Figure S7. **Measuring anticipation effects without reference to the movie data (related to Fig. 8).** Rather than aligning the movie-watching, listening with memory, and listening without memory datasets as in Fig. 8, we can instead align just the two listening conditions and measure the deviation of the model correspondence from a straight diagonal line. For a given probabilistic correspondence, we computed the expected distance from the diagonal and compared to a null model in which subjects were randomly permuted between the memory and no-memory conditions. The probability that the memory group significantly led the no-memory group was converted to a false discovery rate and thresholded at $q<0.05$. Since this analysis uses less data, the results are noisier, but still show the same basic result, with the strongest effects in posterior cingulate cortex, and smaller effects on the angular gyrus and anterior temporal lobe (the medial frontal lobe also showed an effect at a weaker threshold). Note that the searchlight inset was selected post-hoc for illustration.

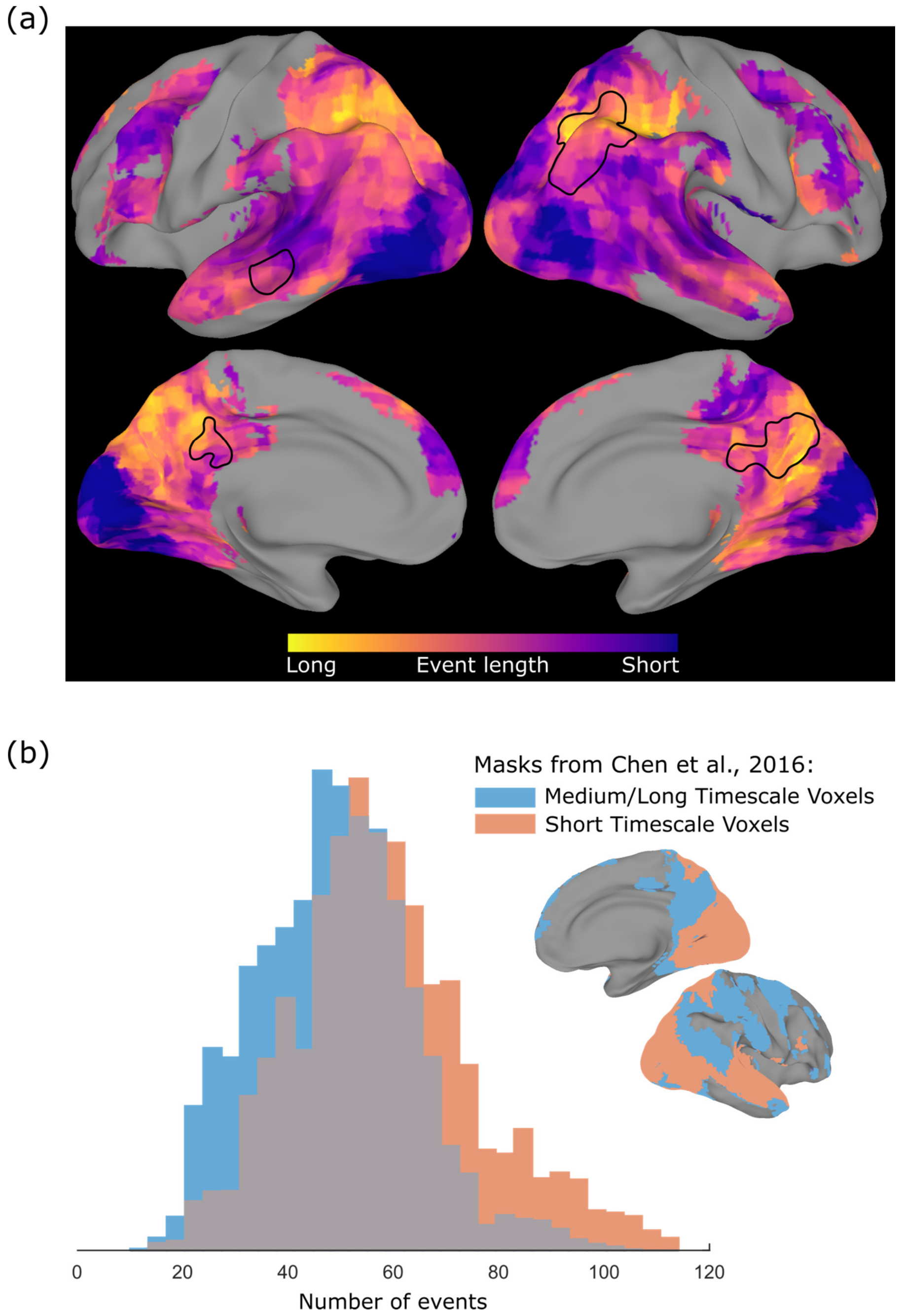

Figure S8: **Comparison of cortical topographies across analyses and with previous work (related to Figs. 3, 5, 6, 8).** (a) The optimal number of events during movie watching (from Fig. 3) is overlaid with the regions (black outlines) the show effects in all three maps in Fig. 5 (regions showing across-modality correspondence), Fig. 6 (regions with event boundaries that predict hippocampal activity increases), and Fig. 8 (regions showing anticipatory reinstatement). The primary regions of overlap are the angular gyrus and posterior medial cortex (and inferior temporal cortex), all of which exhibit long event segments. (b) The optimal number of events during movie watching was compared to a map of voxel timescales (J. Chen et al., 2016), which was defined based on sensitivity to temporal scrambling of a movie. Although derived from very different types of experimental data, these two approaches yield similar topographies. The majority of regions with a small number of (long) events have medium/long temporal receptive windows (blue), while the majority of regions with a large number of (short) events have short temporal receptive windows (orange).