# The Neural Basis of Event Segmentation: Stable Features in the Environment are Reflected by Neural States

**Djamari Oetringer[1,✉], Dora Gözükara[1], Umut Güçlü[1], and Linda Geerligs[1]**

[1]Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, the Netherlands

**Our senses receive a continuous stream of complex information. Parsing this information into meaningful events allows us to extract relevant information, remember it, and act upon it. Previous research has related these events to so-called 'neural states': temporally and regionally specific stable patterns of brain activity, which tend to coincide with events in the stimulus. Neural states show a temporal cortical hierarchy: short states are present in early sensory areas, while longer states can be found in higher-level areas. Here we investigated what these neural states represent. We hypothesized that states at different levels of the cortical hierarchy are shaped by aspects of the stimulus to which these brain areas are responsive. To test this hypothesis, we analyzed fMRI data of participants watching a movie, using a data-driven method to identify the neural states. We found support for the aforementioned hypothesis: specifically the parahippocampal place area and retrosplenial cortex, known to be sensitive to places, showed an alignment between neural state boundaries and moments in the movie with a change in location, independent of changes in visual features. These findings suggest that neural states reflect stable features in the (internal model of) the external environment, and that the cortical temporal hierarchy partly reflects the temporal scales at which representations of the environment evolve.**

**Event segmentation | neural states | naturalistic viewing | fMRI | greedy state boundary search | StudyForrest | stable brain patterns**

**Correspondence:** ***djamari.oetringer@donders.ru.nl***

## Introduction

The continuous stream of sensory input that we receive in our daily lives contains a large amount of information. Despite this complexity, we are seemingly easily able to process this information and use it appropriately. A theory that can partially explain this skill is Event Segmentation Theory, which entails that we automatically segment incoming information into distinct events (Zacks et al., 2007). Here, events are "segments of time at a given location that is conceived by an observer to have a beginning and an end" (Zacks and Tversky (2001), p. 3). The process of event segmentation can support the processing of complex information by creating smaller chunks of meaningful information (Kurby and Zacks, 2008). It can additionally aid memory encoding by updating representations of the current event at event boundaries, freeing information from working memory (Kurby and Zacks, 2008; Zacks et al., 2007), and by increasing access to relevant information and reducing interference (Shin and DuBrow, 2020). Various behavioral studies suggest that events can be segmented at different levels of granularity and that coarse-grained event boundaries tend to co-occur with the fine-grained boundaries (Kurby and Zacks, 2008; Newtson, 1973; Zacks et al., 2001b). These results suggest that events are structured as a nested hierarchy, where big events contain multiple smaller ones.

While the behavioral properties and consequences of event segmentation have been studied extensively, much less is known about the neural mechanisms underlying event segmentation. Initially, the neural underpinnings were studied by looking at evoked responses at the moment of an event boundary, while participants experienced a naturalistic and dynamic stimulus. Event boundaries have been associated with an increased neural response in a wide network of brain regions, including the precuneus, fusiform gyrus, anterior and posterior cingulate gyrus, temporal gyrus, and middle frontal gyrus (Speer et al., 2007; Zacks et al., 2001a, 2010). A number of studies have particularly looked into the event boundary response in the hippocampus, which is larger after events that are later remembered compared to those that are later forgotten (Ben-Yakov and Dudai, 2011; Ben-Yakov et al., 2013; Ben-Yakov and Henson, 2018), which further supports the importance of event segmentation for memory. These studies show where in the brain the neural activity is associated with the occurrence of event boundaries but they do not reveal how segmentation might be implemented at different levels of the cortical hierarchy. This gap can be addressed by studying so-called neural states, which are relatively stable patterns of brain activity within a specific region (Geerligs et al., 2021).

In their groundbreaking study on neural mechanisms of event segmentation, Baldassano et al. (2017) proposed that each brain area along the cortical hierarchy segments information, starting with short neural states in low-level sensory areas and long states in high-level multimodal areas. In support of this theory, it has now been shown that neural states form a partially nested hierarchy across a set of regions (Baldassano et al., 2017; Geerligs et al., 2022), similar to how fine- and coarse-grained events are structured in a nested hierarchy. Neural state boundaries have been shown to coincide with event boundaries, suggesting that neural states are driven by changes in the mental model or the external environment of the participant (Baldassano et al., 2017; Geerligs et al., 2022). Indeed, neural state durations can vary substantially over time

within a particular brain region, suggesting that the durations do not reflect a fixed temporal scale of information processing but instead change dynamically based on temporal fluctuations in the features of the stimulus.

Even though it is clear that neural states are associated with the external environment, it remains unclear what a neural state actually represents. In this study, we tested the hypothesis initially presented by Baldassano et al. (2017), that neural state boundaries in a given brain region reflect a change in the environment that is relevant for that brain region. For example, the fusiform face area (FFA) would have a neural state boundary when you move your head to look from one person to another. If this is the case, neural states might represent stable features in the external environment or in the mental representation that a person has about that environment. Since aspects that are relevant for higher-level regions also tend to evolve at a slower timescale, this could also explain the observed temporal hierarchy in neural state durations.

To investigate our hypothesis, we used an open fMRI dataset in which subjects were watching a movie. We investigated the neural states in a small set of regions of interest (ROIs) for which we could make clear predictions about the features in the external environment that they are responsive to. These were the parahippocampal place area (PPA) and retrosplenial cortex (RSC) that are both known to play an important role in visual scene recognition (Epstein and Kanwisher, 1998; Epstein and Higgins, 2007; Kriegeskorte and Kievit, 2013). If neural state boundaries in a particular region reflect a change in the environment that is relevant for that region, we expect that the RSC and PPA would show an alignment between their neural state boundaries and moments in the movie at which there was a change in location. Importantly, we expected this alignment between neural states and locations to be specific to these location-sensitive areas, and to be independent from co-varying changes in the stimulus such as event boundaries and changes in low-level visual features. Therefore, we included a correction for covariates, and additionally investigated the neural state boundaries in an early visual ROI, where we expected to see an alignment between state boundaries and changes in low-level visual features, but not with location changes. The specificity of these associations was further investigated in an exploratory whole-brain analysis. We indeed observed a reliable and regionally specific alignment between neural state boundaries and changes in the stimulus that are associated with the region, independent from a multitude of covariates in the stimulus. These results indicate that neural states reflect stable features in the (internal model of) the external environment. They also suggest that the cortical temporal hierarchy in terms of neural state durations partly reflects the temporal scales at which different aspects of the environment naturally evolve.

## Methods

**fMRI data.** To investigate neural states during complex and dynamic input with some available stimulus annotations, we used an open dataset called StudyForrest (`www.studyforrest.org/`). We specifically used the fMRI dataset in which 15 participants (age 19–30, mean 22.4, 10 female, 5 male) watched a slightly cut-down version of the movie Forrest Gump dubbed in German (Hanke et al., 2016). The 2-hour movie was divided into 8 segments that were presented across two sessions on the same day. In total there were 3599 volumes of data acquired with a 32-channel head coil using a whole-body 3 Tesla Philips Achieva dStream MRI scanner. The acquired images were T2*-weighted echo-planar images (gradient-echo, 2 s repetition time (TR), 30 ms echo time, 90° flip angle, 1943 Hz/Px bandwidth, parallel acquisition with sensitivity encoding (SENSE) reduction factor 2). There were 35 axial slices (thickness 3.0 mm) with 80 × 80 voxels (3.0 × 3.0 mm) of in-plane resolution, 240 mm field-of-view (FoV), and an anterior-to-posterior phase encoding direction with a 10% inter-slice gap, recorded in ascending order. For anatomical alignment, we used the T1-weighted high-resolution structural scans as provided by Hanke et al. (2014).

To functionally define our regions of interests, we used the functional localizer data that is additionally part of StudyForrest (Sengupta et al., 2016). The MRI acquisition setup of this localizer was the same as for the movie data.

**Annotations.** In order to properly investigate the relationship between neural states and the stimulus, a multitude of stimulus annotation categories were gathered: shots, locations, events, low-level visual features, and visual features of different levels of complexity using a convolutional neural network. Annotations of shots and locations were taken from Häusler and Hanke (2016), who manually annotated the data by determining when there was a new shot on a frame-by-frame basis, and noting down at which location each shot took place. They defined three levels of locations: 'locale' (e.g. 'living room' or 'corridor downstairs'), 'setting' (e.g. 'Gump house' or 'Jenny's house'), and 'major location' (e.g. 'Greenbow Alabama' or 'California'). In our analysis, we only used 'locale' (from this point on referred to as 'small-scale locations') and 'setting' (from this point on referred to as 'large-scale locations'). The level 'major location' was not included as it reflects conceptual changes that are probably more relevant to other areas, such as the anterior temporal lobe (Bonner and Price, 2013). Further, this influence, if present, would depend on the subjects' knowledge of the geography of the United States of America, while the subjects were German. Finally, given the scale of this annotation, the number of changes in such large-scale locations was relatively low throughout the movie, making it more difficult to detect any significant effect (Hamilton and Huth, 2020).

To ensure that associations between neural state boundaries and location changes were not caused by changes in other aspects of the stimulus, we also acquired annotations of low-level visual features and event boundaries that could be included in the analysis as covariates. The annotations of low-level visual features were based on perceptual hashes of the various frames and taken from the StudyForrest GitHub pages[1]. Perceptual hashing is an algorithm that takes an im-

[1] github.com/psychoinformatics-de/studyforrest-data-confoundsannotation

age as input, and outputs a string of characters that can be interpreted as the fingerprint of that image. This algorithm is designed such that the fingerprints of images are close to each other when they contain similar features, and is therefore insensitive to cropping or a change in resolution. Specifically relevant in a movie, this measure is insensitive to ongoing movement. For example, two consecutive frames in which a character is running towards the left would be seen as very similar to the human eye, as well as to perceptual hashing, while a pixel-based measurement would be very sensitive to such movement. The difference in low-level visual features over time were defined on a frame-by-frame basis: after computing each frame's perceptual hash, the Hamming distances between the perceptual hash values of consecutive frames were measured, and consequently normalized in order to have them range between 0 and 1.

The perceptual hashes described above only account for the visual features that are of one specific level of complexity, while visual features of other levels of complexity also have the potential to drive associations between for example changes in location and neural state boundaries. Therefore, visual features at various levels of complexity were additionally extracted from the movie by using Alexnet (Krizhevsky et al., 2012), a convolutional neural network trained for object recognition. Each stimulus frame, excluding the gray bars at the bottom and top of the screen, was given as input to Alexnet with pretrained weights. The outputs of all layers except the output layer were extracted, giving us an approximation of visual features across 8 layers, which can be seen as 8 different levels of complexity (Zeiler and Fergus, 2014). As we were interested in the frame as a whole, we averaged the outputs over the spatial dimensions to discard the spatial information. Then, the frame-to-frame differences were computed by taking the cosine distance between the resulting vectors of two consecutive frames.

Finally, event boundaries of the movie stimulus were defined by Ben-Yakov and Henson (2018), whose subjects watched the Forrest Gump movie stimulus and indicated with a button press when they thought a new event had started. These subjects were native English and watched the movie in either German or English, though the button presses were not significantly different between the original English movie and the German dubbed version. 0.9 seconds were subtracted from each button press to account for response time, and then button presses of different subjects were taken together when they were close in time. In order to exclude accidental presses and event boundaries with low between-subject reliability, only moments at which at least 5 out of 16 subjects had pressed a button around roughly the same time were considered an event boundary.

Overall, the correlations between the different annotations were high (see Supplementary section A). This is due to the nature of movies and the annotation categories. For example, a change in location can only occur when there is also a new shot. Although the event boundaries were acquired with a more subjective method, the moments at which observers press the button generally tends to coincide with changes in location (Magliano et al., 2001).

**Preprocessing.** In this study we used the movie fMRI data that had previously been partially preprocessed and denoised by Liu et al. (2019). They performed motion correction, slice timing correction, brain extraction and high-pass temporal filtering (cutoff of 200 s). They then denoised the data by using spatial ICA decomposition and manual inspection of the resulting components to remove artifacts as a results of hardware, head motion, physiology, and unclear sources. As the computation of neural state boundaries requires spatial information, we specifically used the unsmoothed version of this published dataset.

We further preprocessed the data using SPM12 software (http://www.fil.ion.ucl.ac.uk/spm) by aligning the functional data between the eight runs, performing coregistration with the T1 anatomical image of each subject, followed by segmentation, and finally normalization (Ashburner and Friston, 2005) to MNI space. Only gray matter voxels were included, which were defined as voxels that survive the threshold of 0.35 in the gray matter probability map (TPM) as provided by the SPM software. The final gray matter mask only contained voxels that were measured for all subjects in all runs. Here, a voxel was labeled as 'measured' during one run when its value during the first timepoint of that run was higher than 0.8 times the mean of all voxels at that same timepoint.

After normalization and gray matter masking, the data was z-scored per voxel per run. To functionally align the subjects' data and thereby improving inter-subject correlation, hyperalignment was performed on the whole brain, using the implementation of the PyMVPA toolbox (Guntupalli et al., 2016; Haxby et al., 2020). This hyperalignment was computed on the fMRI data of run 4 (488 volumes) using searchlight-hyperalignment. The searchlights had a radius of 3 voxels and all voxels within the gray matter mask were used as the center of a searchlight. The computed hyperalignment parameters were then applied to all remaining runs. Since run 4 was used to compute the transformation parameters, as well as the optimal delay between annotations and neural states (see below), this run was excluded from any further analysis. Run 4 was chosen as it contained a relatively low number of changes in location in the movie, and because it was roughly in the middle of the movie, making it more representative than for example the first or last run.

**ROI definitions.** To functionally define the ROIs, we used data from a functional localizer experiment that was part of the StudyForrest dataset (Sengupta et al., 2016). In this experiment, the StudyForrest participants that had also participated in the audio-visual movie part of the dataset were presented with faces, objects, houses, landscapes, bodies and scrambled images while fMRI data was collected. After data preprocessing, Sengupta et al. (2016) performed a GLM analysis with one regressor per image category. Next, they computed multiple contrasts per subject to find subject-specific voxels that are sensitive to one or multiple of the aforementioned image categories. We used some of these contrasts to define our ROIs. For the left and right PPA and the RSC, we

specifically used the contrast of houses and landscapes with all other image categories. For left and right early visual ROIs we took the contrast of scrambled images with all other image categories, and we excluded the midline, defined as a sagittal plane with a thickness of 2 voxels surrounding the exact middle of the left-right axis, to be able to separate a left and right cluster. To obtain group-level ROIs, required for obtaining reliable neural states, we took the subject-level contrast values, normalized them to MNI space using SPM12, and applied a between-subject t-test per voxel. We then applied the gray matter mask and thresholded the resulting t-values such that one cluster of about 100 voxels was left at roughly the expected anatomical location (Nasr et al., 2011). Descriptives and visualizations of the resulting ROIs can be found in Supplementary Section B.

**Analysis.** To reliably identify neural states using Greedy State Boundary Search (GSBS; Geerligs et al. 2021), the data must have a high enough signal-to-noise ratio (SNR). Since the stimulus is only presented once, data from a single subject has low SNR, which is why analyses were performed on the group-level data by averaging the data across participants. To obtain the neural states per ROI, we applied GSBS to the group-averaged ROI data in each run. GSBS is a data-driven method to extract the neural states from fMRI data in an iterative manner. The algorithm determines the optimal number of states by maximizing within-state correlations and minimizing between-state correlations. Parameters were set to the values as recommended by Geerligs et al. (2021): the maximum number of neural states was set to half the number of time points in a given run, and finetuning was set to 1, meaning that each neural state boundary could be shifted by 1 TR after each iteration, in case that would give a better fit. We specifically used the states-GSBS algorithm, which is an improved and more reliable version of GSBS, introduced by Geerligs et al. (2022), that potentially places two boundaries in a given iteration rather than always placing one boundary at a time.

GSBS gave us a timeline of neural state boundaries per run per ROI, including the strength of each boundary. The strength of a boundary is defined as 1 minus the Pearson correlation of the average neural activity patterns of the two neural states that surround the boundary. Thus, a higher strength implies that the two consecutive neural states are more dissimilar. To compare the neural state boundaries to the various annotations, all annotations were first delayed by 4.5 seconds to account for the hemodynamic response, and then downsampled from seconds to fMRI volumes. The delay of 4.5 seconds was based on an independent analysis with the data from the unused run 4, as shown in Supplementary Section C. Finally, all values per run were concatenated, creating one vector per annotation for the whole movie, as well as one neural state boundary vector per ROI.

To measure the overlap between the neural state boundaries and various annotations, we used partial Pearson correlation. Given the high correlations between the various annotation categories, partial correlation was necessary to minimize the possibility that a significant correlation could be explained by a covariate instead of the annotation of interest. For example, a significant correlation with changes in location could actually be due to changes in low-level visual features or vice versa. By using partial correlation, any association we observe cannot be explained by any of the covariates. Computing partial correlations is essentially a two-step procedure. First, two residuals are computed, which are the values of a given annotation and the neural state boundary timeline while correcting for all potential covariates. The second step is to compute Pearson's correlation between the two resulting residuals.

**Table 1.** Overview of the covariates that were selected for each annotation category that was analyzed.

| Annotation of interest | Covariates |
|---|---|
| Low-level visual features | Large-scale locations<br>Small-scale locations<br>Events |
| Small-scale locations | Alexnet layers<br>Low-level visual features<br>Shots<br>Events |
| Large-scale locations | Alexnet layers<br>Low-level visual features<br>Shots<br>Events |
| Events | Alexnet layers<br>Low-level visual features<br>Shots<br>Large-scale locations<br>Small-scale locations |

To be sure that we only correct for relevant covariates, a different set of covariates was used for each annotation of interest. Which covariates were chosen for which annotation of interest can be seen in Table 1. For example, the residual of 'small-scale location changes' describes the moments at which there was a small-scale change in location, while correcting for all 8 Alexnet layers, low-level visual features, shots, and events.

The p-value that results from Pearson's correlation between the residuals could not be used here as the annotation values did not meet the assumption of normality. We therefore used permutation testing to compute p-values. In each permutation, the order of neural states were shuffled within each run, while keeping the number of neural states and their duration intact. Then, partial correlation was again computed on the concatenated runs. In total there were 10,000 permutations to compute a null distribution per ROI and annotation category combination. The p-value was computed by counting how many values in the null distribution were above or equal to the actual measurement, and thus the test was one-tailed. All tests were corrected for multiple comparisons using Bonferroni correction, correcting for the number of ROIs.

The resulting correlation coefficients would partially depend on the number of changes in a given annotation and the number of neural state boundaries in a given ROI. Therefore, the correlations of different ROIs and annotations cannot directly be compared to each other. Thus, for visualization purposes,

the relative correlation was computed. This is defined as the actual correlation coefficient divided by the maximum possible correlation coefficient. The maximum possible correlation coefficient was computed by first computing the residuals, then sorting all values per residual per run from low to high, and then computing the correlation between the sorted residuals.

As part of a more exploratory whole-brain analysis, we created 3461 partially overlapping spherical searchlights across the gray matter masked cortex. The centers of the searchlights were spaced 2 voxels apart, and the radius of each searchlight was set to 3.5 voxels. Only searchlights overlapping with at least 20 voxels were included, and the cerebellum and sub-cortical areas were excluded, giving us searchlights of up to 178 voxels (4806 $mm3$), with an average size of 144 voxels (3888 $mm3$). Then, the same analysis was applied to each searchlight as was done for the previously mentioned ROIs, except for the number of permutations, which was set to 1,000 per searchlight. To correct for multiple comparisons, we applied a false discovery rate (FDR) correction over all searchlights (Benjamini and Hochberg, 1995). A voxel was only interpreted as significant when the average p-value of all searchlights overlapping with that voxel was smaller than the highest p-value that survived the FDR-correction.

## Results

**Location changes are selectively associated with neural state boundaries in PPA and RSC.** If neural states are driven by aspects in the environment to which a brain area is responsive, then the state boundaries in location-sensitive areas should align with moments in the movie at which there is a change in location. This association should remain present even when correcting for visual features at different levels of complexity, shots, and event boundaries. Indeed, the neural states boundaries in the right PPA showed a significant association with both small- and large-scale location changes ($p < 0.001$; see Figure 1A). For small-scale location changes we additionally found a significant alignment with neural states boundaries in the left RSC ($p < 0.005$), while the alignments in the left PPA ($p = 0.009$) and the right RSC ($p = 0.016$) did not survive the Bonferroni correction. The number of neural state boundaries was much higher in both PPA (right: 373, left: 367) and RSC (right: 412, left: 416) than the number of small-scale location changes (231 changes; see Figure S4), indicating that there is no one-to-one correspondence between neural states and locations. For large-scale location changes, the alignment with neural state boundaries in the left RSC ($p = 0.016$) did not survive Bonferroni correction. The differences between the left and right ROIs could be related to the low number of subjects and consequently the low signal-to-noise ratio. Adding more subjects may alleviate the lateralization.

Location changes are naturally accompanied by changes in visual features and also tend to occur when there is a new shot. To test whether our corrections for changes in low-level visual features were sufficient, we also investigated the left

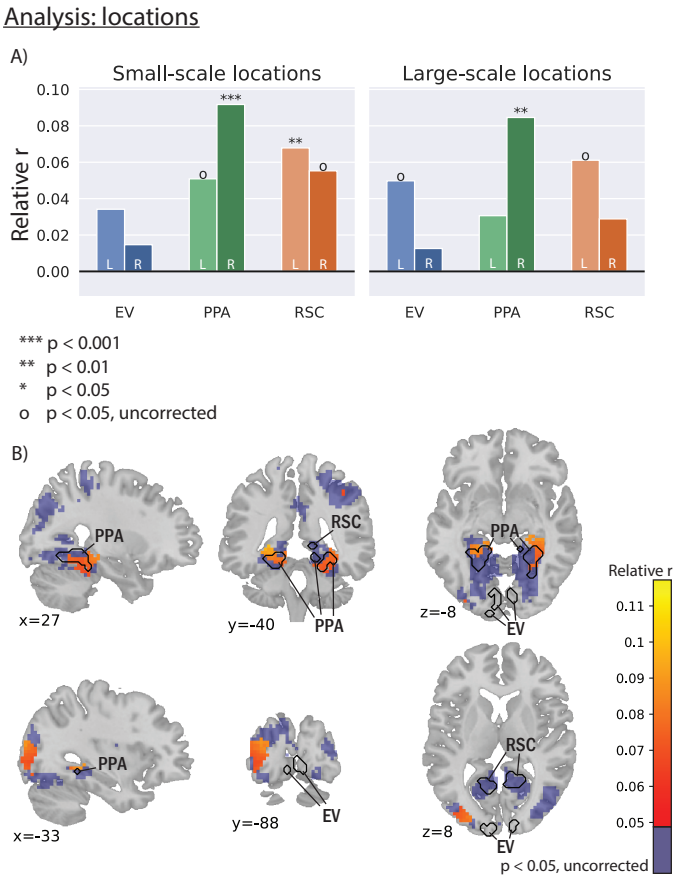


**Fig. 1.** Neural state boundaries in the location-sensitive PPA and RSC showed a significant alignment with location changes, especially for the small-scale locations. Importantly, the alignment was absent in early visual areas. The values represent the relative r between neural state boundaries and changes in location, while correcting for shots, events, low-level visual features, and features from 8 Alexnet layers. The relative r is defined as the actual Pearson's correlation coefficient divided by the Pearson's correlation coefficient that can maximally be expected given the number and strengths of the neural state boundaries and the annotations. EV: early visual. PPA: parahippocampal place area. RSC: retrosplenial cortex. A) ROI analysis for both small-scale and large-scale locations. B) Thresholded results of the whole-brain analysis for small-scale locations. The results of the whole-brain analysis for large-scale locations can be found in Figure S5.

and right early visual areas, which should not show a significant alignment between their neural state boundaries and changes in location. In those ROIs, we indeed only observed a partial correlation with large-scale location changes in the left early visual ($p = 0.031$), which did not survive Bonferroni correction. These results suggest that the association we observed in the PPA and RSC cannot be explained by residual effects of changes in visual features, and that the association with locations is specific to the location-sensitive ROIs.

The specificity of the association between location changes and neural state boundaries was also supported by the exploratory whole-brain analysis. Two clusters with a significant alignment between the neural state boundaries and small-scale location changes were overlapping with both the left and right functionally defined PPA (Figure 1B, top). These clusters were also present for large-scale locations changes, though much smaller (Figure S5). A third small-scale location cluster was detected in left V2 (Figure 1B,

bottom). Contrary to the ROI analysis, no location clusters overlapped with the functionally defined RSC. This may be because the correction for multiple comparisons was stricter in the whole-brain analysis than in the ROI analysis. Taken together, these results support the hypothesis that neural state boundaries overlap with location changes in location-sensitive regions specifically.

**Changes in visual features align with neural state boundaries in visual areas.** If changes in the relevant features in the environment are driving neural state boundaries, we should see this effect not only for location changes but also for other features. Therefore, we also looked at how changes in low-level visual features are associated with neural state boundaries in the early visual areas, while correcting for locations changes and event boundaries. Indeed, we observed that both the left and right early visual areas showed a significant alignment between neural state boundaries and changes in low-level visual features ($p < 0.001$; see Figure 2A). This alignment was also observed in left and right PPA ($p < 0.005$), but the correlations in RSC did not survive Bonferroni correction (left: $p = 0.011$, right: $p = 0.034$). The clusters found in the exploratory whole-brain analysis suggested a more regionally specific association between changes in visual features and neural state boundaries. A big cluster was present in medial V1 (Figure 2B, top), overlapping with the functionally defined early visual areas. A second cluster was present in the left frontal eye fields (FEF; Figure 2B, bottom). Contrary to the ROI analysis, no cluster overlapped with the functionally defined PPA or RSC, which may be because the correction for multiple comparisons was stricter in the whole-brain analysis than in the ROI analysis. We did indeed observe alignment between neural state boundaries and changes in low-level visual features in the PPA (but not the RSC) in the whole-brain analysis when no correction for multiple comparisons was performed. Another possibility is that the difference in shape is detrimental: both left and right functionally defined PPA and RSC had an elongated shape (Figure S2), while the searchlights used in the whole-brain analysis are round and defined by a center and a radius, possibly making the whole-brain analysis unable to capture the full characteristics of the PPA and RSC.

**Event boundaries align with neural states in higher-level cortical areas.** Alignments between neural state boundaries and event boundaries have been investigated before in previous literature, but not yet while correcting for visual covariates in the stimulus. To investigate to which extent these alignments could be driven by visual aspects rather than event boundaries, we performed our whole-brain analysis on event boundaries as well, while correcting for visual features at various levels of complexity, as well as shots and location changes. The clusters found to show a significant alignment with event boundaries were numerous and spread throughout the cortex (Figure 3), and could specifically be found in higher-level areas such as the angular gyrus, auditory cortex, and medial prefrontal cortex, but not in low-level visual areas.

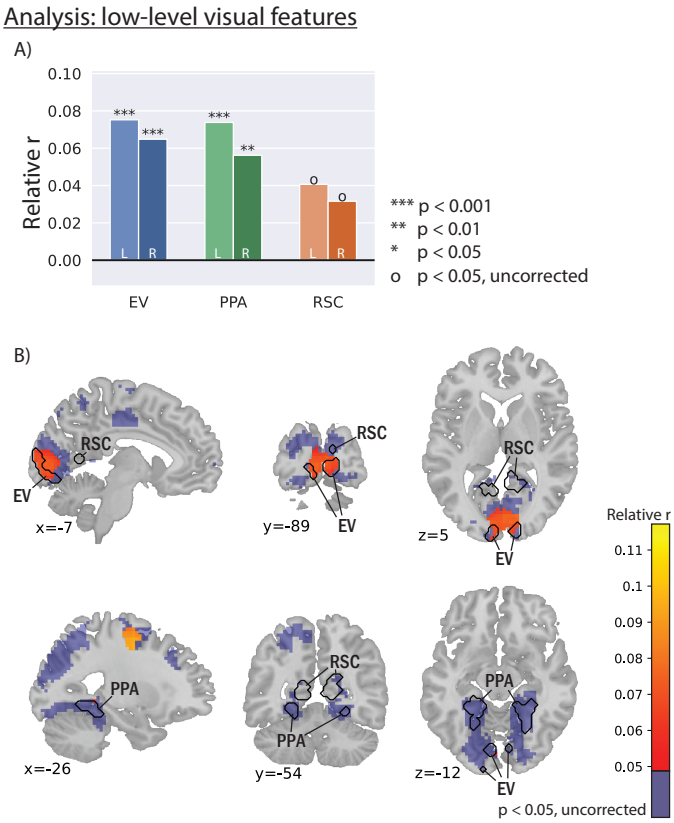


**Fig. 2.** Neural states in early visual areas showed a significant alignment with low-level visual features, in both the ROI and the whole-brain analysis. In the ROI analysis, this alignment was also present in the PPA. The values represent the relative r between neural state boundaries and changes in low-level visual features, while correcting for locations and events. The relative r is defined as the actual Pearson's correlation coefficient divided by the Pearson's correlation coefficient that can maximally be expected given the number and strengths of the neural state boundaries and the annotations. EV: early visual. PPA: parahippocampal place area. RSC: retrosplenial cortex. A) ROI analysis. B) Thresholded results of the whole-brain analysis.

**Correcting for covariates is crucial to isolate feature specific associations.** Because of the high correlations in the various annotation categories, we used partial correlation in our analysis to minimize the possibility that a significant correlation could be explained by a covariate rather than the annotation of interest. To investigate the consequences of this approach, we also performed the same analysis in Supplementary Section F but with Pearson's correlation rather than partial correlation. In summary, we found that many areas now showed a significant alignment with more annotation categories than in the partial correlation analysis, in both the ROI analysis and the whole-brain analysis. In particular for events, the vast majority of the cortex now showed a significant alignment. This suggests that the correction for covariates allowed us to better isolate feature specific associations between neural state boundaries and changes in the stimulus.

**Shots show no unique association with neural state boundaries.** The presence of shots in the stimulus could be interesting to investigate, as these do not happen in real life while we do consider movie stimuli to be naturalistic. Although such a discontinuity in visual information can still be interpreted as a continuous action depending on how the

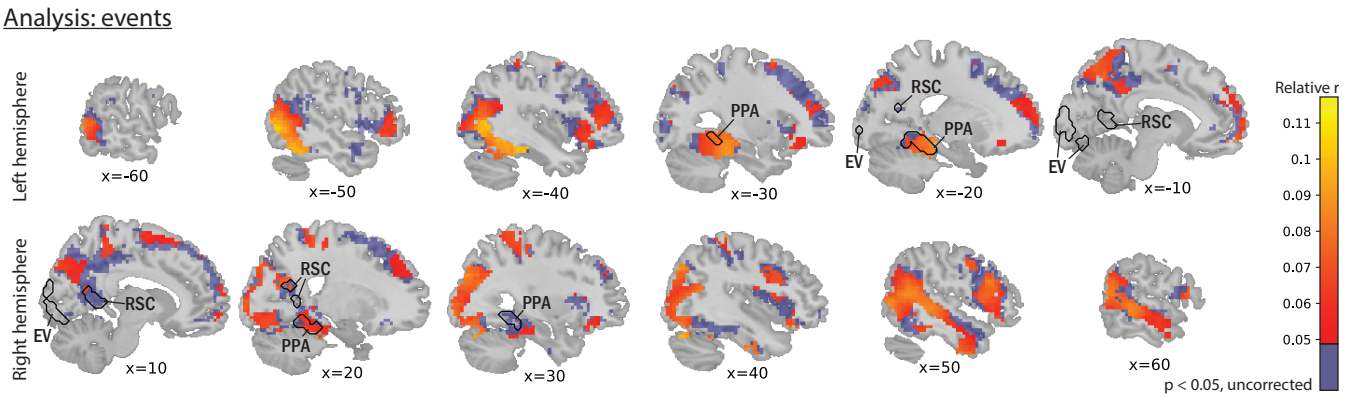


**Fig. 3.** Voxels with a significant alignment between neural state boundaries and event boundaries were found throughout the brain, including the angular gyrus, auditory cortex, and medial prefrontal cortex, but excluding low-level visual areas. Values represent relative r values between neural state boundaries and event boundaries across the brain, while correcting for low-level visual features, shots, locations, and 8 layers of Alexnet. The relative r is defined as the actual Pearson's correlation coefficient divided by the Pearson's correlation coefficient that can maximally be expected given the number and strengths of the neural state boundaries and the annotations. EV: early visual. PPA: parahippocampal place area. RSC: retrosplenial cortex.

shots are edited together (Magliano and Zacks, 2011), one may wonder whether shots lead to any particular changes in neural states that would not appear in an actual real environment. Sanz-Aznar et al. (2021) for example found modulations in various frequency bands that are possibly related to processes of memory encoding and access, as well as to the default mode network. To study the effects of shots on neural state boundaries, we ran another extra analysis, both within ROIs and across the whole brain, with shots as our annotation of interest (see Supplementary Section G). When performing a simple correlation analysis without correcting for covariates, a significant alignment was present between the neural states and shots in all of our ROIs, as well as across a big part of the cortex and in particular the visual areas. However, when correcting for low-level visual features, changes in location, events, and 8 layers of Alexnet, this alignment was absent in all ROIs and all voxels. Thus, in terms of alignment between neural states and shots, there does not seem to be anything relevant about shots on top of the changes in visual features. This is in line with Zacks et al. (2010), who found that shots did not produce increases in segmentation in and of themselves, but were instead related to other transitions in the stimulus such as changes location, object, or characters.

## Discussion

Understanding how the brain segments information as it receives continuous and dynamically varying sensory input is of critical importance for moving to a more ecologically valid understanding of brain function (Lee et al., 2020; Willems et al., 2020). Recent work has established that event segmentation could be underpinned by neural states, which form a temporal hierarchy across the cortex. The aim of this study was to test whether these neural states reflect meaningful temporal units in the environment. Our findings provide support for this hypothesis and suggest that the cortical temporal hierarchy found in previous literature partly reflects the temporal scales at which representations in the environment evolve. In the following paragraphs we will first discuss the evidence for this conclusion in more detail before looking at broader implications of this research for our understanding of neural state segmentation in the brain.

**Neural states at different levels of the cortical hierarchy are shaped by aspects of the environment to which these brain areas are responsive.** It has previously been hypothesized that neural states should reflect meaningful units in the environment (Baldassano et al., 2017). However up to now, neural state boundaries have only been linked to event boundaries specifically (Baldassano et al., 2017; Geerligs et al., 2022; Mariola et al., 2022) and it has not been investigated if neural states at different levels of the temporal hierarchy indeed represent different meaningful units. In line with this hypothesis, we found that there is a significant alignment between changes in location and neural state boundaries in the place-sensitive PPA and RSC. The whole brain analysis showed that this alignment was highly specific to such place-sensitive areas. Importantly, these significant alignments could not be explained by changes in visual features of various levels of complexity, nor by event boundaries. Additional support was found by investigating neural states in early visual areas, which did not show this association with locations, but did have a significant alignment with low-level visual features. Again, the whole brain analysis revealed that this association between low-level features and neural states was highly specific to early visual areas. Together these results provide strong support for our hypothesis that neural states in location sensitive areas reflect the stable representations of the locations present in the external environment.

Although we found alignments between the ends of neural states and the ends of meaningful units in the stimulus, neural states and stable features in the environment do not have a one-to-one correspondence. In general, brain regions are not necessarily uniquely responsive to one specific measurable aspect in the environment (Fedorenko et al., 2013; Woolgar et al., 2016). In our study in particular, there were many more neural states boundaries in both the PPA and the RSC than the number of location changes throughout the stimulus, indicating that one neural state does not correspond to one lo-

cation on the screen. In the PPA specifically, a significant association between neural states and low-level visual features was found in addition to the association with changes in location. Previous studies have found that the PPA is sensitive to high spatial frequencies (Rajimehr et al., 2011) in addition to places (Seeliger et al., 2021), and texture (Cant and Goodale, 2011). Thus, neural states within a particular brain area could pertain to multiple aspects of the environment, making it difficult to relate a specific neural state to one specific feature in the stimulus.

**Neural states are more than passive representations**

In this study, we focused on the idea that neural states are a reflection of the environment. However, they are likely to be more than just a passive representation of the current sensory input. Instead they could also be associated with how we act on that environment, what we know about the environment from previous experience, and what we are attending to (Baldassano et al., 2017; Zacks and Tversky, 2001). For example, our exploratory whole-brain analysis revealed a cluster in the FEF that had a significant alignment between neural states and low-level visual features. An increase in FEF activity has previously been related to saccades around event boundaries (Rosano et al., 2002; Speer et al., 2003). These saccades happen because event boundaries tend to co-occur with changes in low-level visual features and eye movements are more likely to happen when there are visual changes (Swallow et al., 2004). Thus, our cluster in the FEF probably demonstrates an association between neural state boundaries and eye movements, indicating that active processes, such as eye movements, could also be associated with neural states.

Evidence for the idea that neural states might reflect what we know about the environment from previous experience comes from previous literature showing that neural states may contain information that is currently not accessible in the environment, while still being relevant. For example, Zadbood et al. (2017) showed that the same event-specific activity pattern occurs while experiencing a stimulus, while remembering that same stimulus, and while listening to another person describing it. Thus, activity patterns can be specific to a stimulus even in its absence. Additionally, Ezzyat and Davachi (2021) found object-specific activity to be persistent even after the object had changed, but only as long as the event had not ended yet. This persistent activity pattern seemed to reflect a stable mental model within the event. Thus, neural states could reflect the current model or representation of an event or context (DuBrow et al., 2017; Richmond and Zacks, 2017).

Some brain areas, such as early sensory areas, may be more sensitive to the information that is currently present in the environment, while others may rely more on previous input (Hasson et al., 2008). This possibility is partially reflected in our results, where the low-level visual features only reflect information was available on the screen, with a significant alignment in early visual areas. In contrast, event boundaries are much more conceptual and require knowledge about the situation beyond the current sensory information, and these event boundaries aligned with neural state boundaries in higher-level areas. By having neural states across the cortex, it is possible to have many such representations with different levels of abstraction across many different time scales. These multi-scale representations could aid information processing by reducing complexity at each scale, and together representing a multi-scale environment (Quax et al., 2020). Given the high number of areas that have been reported to be related to events or event boundaries, different parts of the brain may be representing different aspects of event information.

In line with the notion of neural states, previous literature suggests that there might be a 'gating' mechanism that separates distinct mental representations (Chien and Honey, 2020). Information from the recent past is either integrated or not, making it possible to separate currently relevant information from interfering information (Ezzyat and Davachi, 2011; Kurby and Zacks, 2008; Shin and DuBrow, 2020; Zacks et al., 2007). Here we showed that the timing of neural state boundaries is meaningful in multiple brain areas, providing additional support that the gate closes at appropriate moments. Which moments are considered to be appropriate, could be influenced by top-down mechanisms such as attention, as neural states are affected by the task at hand (De Soares et al., 2023).

**The temporal cortical hierarchy is partly driven by the temporal evolution of various features in the environment.** A multitude of temporal hierarchies across the brain have previously been reported. Using fMRI and a movie stimulus, Hasson et al. (2008) introduced the concept of temporal receptive windows (TRWs): the duration of the period in the past in which information affects the current response. This time window was longer in higher-level areas, showing a temporal hierarchy across the visual cortex. These results have been replicated in the auditory domain (Lerner et al., 2011), and in electrophysiological measurements (Honey et al., 2012). In line with this temporal cortical hierarchy, the duration of neural states shows a nested temporal hierarchy (Baldassano et al., 2017; Geerligs et al., 2022). The finding that these neural states align with meaningful units in the environment, suggests that the neural state hierarchy could naturally arise from how fast various aspects of information from the environment tend to evolve: low-level visual features can change very quickly, while a change in location occurs much less often. Previously reported variations in the duration of neural states could be explained by a variation in duration of the relevant aspect in the external environment, such as a movie scene at one location taking longer than a scene at another location (Geerligs et al., 2022). Moreover, a nested hierarchy of information is additionally present in the environment and could be reflected in nested neural states: when there is a change in location, there will always be a change in low-level visual features, but not vice versa. Taken together, we show that the hierarchical information that is present in the natural environment is at least one possible component that contributes to the temporal cortical hierarchies that have been reported in previous literature. However, it is likely that this is just one of multiple aspects

that contribute to the hierarchies, as intrinsic timescale and power spectra gradients have been reported in the absence of a stimulus as well (Honey et al., 2012; Raut et al., 2020; Stephens et al., 2013).

**The merits and drawbacks of naturalistic stimuli.** Naturalistic stimuli such as movies and audio books contain a lot of complex information, with different features simultaneously developing over time at multiple scales. Together, these evolving features form a coherent narrative. This contributes to their ecologically validity, as the same type of information is present in the real-world environment (Willems et al., 2020). Studying the brain in an ecologically valid environment facilitates the possibility to study the brain during an engaging and realistic experience (Lee et al., 2020), and is essential to prevent oversimplified theories (Cantlon, 2020; Nastase et al., 2020).

Multi-scale information that is present in both the real-world environment and naturalistic stimuli often forms a hierarchy, with for example a new person entering the room coinciding with a major change in visual and/or auditory information. Therefore, when annotating multiple aspects in a naturalistic stimulus, such aspects are likely to be highly correlated (Hamilton and Huth, 2020; Nastase et al., 2020). To deal with this issue, we corrected for a multitude of covariates in our analyses, including visual features (to study location-specific states) and event boundaries. We observed much more regionally specific associations between neural states and different annotation categories when a correction for covariates was performed, compared to when no corrections were applied. This can be illustrated by looking at the association between event boundaries and neural states. Some of the regions that showed associations between neural states and events in previous work (Geerligs et al., 2022), including early visual regions, anterior insula, and anterior cingulate cortex, only showed a significant alignment between neural state boundaries and event boundaries in the absence of any correction for covariates. This suggests that this association may be because of the co-occurrence between event boundaries and visual changes, rather than essentially being associated with events. However, most higher-level regions that were observed before, also showed an association between neural states and events with corrections for covariates. These regions included the cingulate cortex, the dorsal medial prefrontal cortex, superior frontal gyrus, middle frontal gyrus, precuneus, middle temporal gyrus and superior temporal gyrus, suggesting that these areas contain representations of abstract and gist-level descriptions of an event (Bird, 2020).

Even though our correction for covarying visual features was effective, it was probably not perfect. For example, when looking at the association between location changes and neural states in the whole-brain analysis, we observed an unexpected cluster in left V2. While not overlapping with our functionally defined early visual areas, V2 could still be considered too low-level to be sensitive to changes in location. These results highlight that even though we carefully selected a broad set of covariates that should capture both low and high-level visual features in the external environment, we can never be completely confident that we were able to account for all visual changes. Indeed, since location changes can only be observed by combining many different visual features in the environment, a full disentanglement may be impossible. In addition, there may be other covariates in the movie that we did not correct for. For example, the auditory cortex showed a significant partial alignment between neural state boundaries and event boundaries. This may have been because we did not correct for any auditory features in the movie, while it is plausible that music and conversations often end or start at new scenes, and thus could be correlated with event boundaries.

In fact, taking into account all possible features is infeasible. It is impossible to fully isolate one aspect of a movie stimulus. Although this sounds like a drawback of naturalistic stimuli, it is also a reflection of their ecological validity, as those correlations are inherently present in the real world as well. For example, event boundaries reflect a change in one or multiple dimensions of the stimulus. If we would account for all of these dimensions as covariates, there would be nothing left of the event. Therefore, the covariation of various aspects of the same stimulus is an important aspect to keep in mind when analyzing naturalistic data, and including covariates in the analyses is one way to disentangle some of these effects.

**Conclusion.** In conclusion, we found support for the hypothesis that neural state boundaries in a specific area occur when an aspect in the environment that is relevant to that area changes. This also suggests that neural states reflect stable features in the environment and provides further support for the idea that neural state segmentation at different levels of the cortical hierarchy underlies event segmentation, aiding information processing as well as memory. Finally, our results indicate that the cortical temporal hierarchy in terms of neural state durations partly reflects the temporal scales at which different aspects of the environment naturally evolve.

**ACKNOWLEDGEMENTS**
We would like to thank Aya Ben-Yakov for providing us with the event segmentation results of the StudyForrest movie stimulus. Linda Geerligs was supported by a Vidi grant (VI.Vidi.201.150) from the Netherlands Organization for Scientific Research.

## References

Ashburner, J. and Friston, K. J. (2005). Unified segmentation. *Neuroimage*, 26(3):839–851.

Baldassano, C., Chen, J., Zadbood, A., Pillow, J. W., Hasson, U., and Norman, K. A. (2017). Discovering Event Structure in Continuous Narrative Perception and Memory. *Neuron*, 95(3):709–721.e5.

Ben-Yakov, A. and Dudai, Y. (2011). Constructing realistic engrams: Poststimulus activity of hippocampus and dorsal striatum predicts subsequent episodic memory. *Journal of Neuroscience*, 31(24):9032–9042.

Ben-Yakov, A., Eshel, N., and Dudai, Y. (2013). Hippocampal immediate poststimulus activity in the encoding of consecutive naturalistic episodes. *Journal of Experimental Psychology: General*, 142(4):1255–1263.

Ben-Yakov, A. and Henson, R. N. (2018). The hippocampal film editor: Sensitivity and specificity to event boundaries in continuous experience. *Journal of Neuroscience*, 38(47):10057–10068.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300.

Bird, C. M. (2020). How do we remember events? *Current Opinion in Behavioral Sciences*, 32:120–125.

Bonner, M. F. and Price, A. R. (2013). Where is the anterior temporal lobe and what does it do? *Journal of Neuroscience*, 33(10):4213–4215.

Cant, J. S. and Goodale, M. A. (2011). Scratching beneath the surface: New insights into the functional properties of the lateral occipital area and parahippocampal place area. *Journal of Neuroscience*, 31(22):8248–8258.

Cantlon, J. F. (2020). The balance of rigor and reality in developmental neuroscience. *NeuroImage*, 216:116464.

Chien, H. Y. S. and Honey, C. J. (2020). Constructing and Forgetting Temporal Context in the Human Cerebral Cortex. *Neuron*, 106(4):675–686.e11.

De Soares, A., Kim, T., Mugisho, F., Zhu, E., Lin, A., Zheng, C., and Baldassano, C. (2023). Top-down attention shifts behavioral and neural event boundaries in narratives with overlapping event scripts. *bioRxiv*.

DuBrow, S., Rouhani, N., Niv, Y., and Norman, K. A. (2017). Does mental context drift or shift? *Current opinion in behavioral sciences*, 17:141–146.

Epstein, R. and Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature 1998 392:6676*, 392(6676):598–601.

Epstein, R. A. and Higgins, J. S. (2007). Differential Parahippocampal and Retrosplenial Involvement in Three Types of Visual Scene Recognition. *Cerebral Cortex*, 17(7):1680–1693.

Ezzyat, Y. and Davachi, L. (2011). What constitutes an episode in episodic memory? *Psychological science*, 22(2):243–252.

Ezzyat, Y. and Davachi, L. (2021). Neural Evidence for Representational Persistence Within Events. *Journal of Neuroscience*.

Fedorenko, E., Duncan, J., and Kanwisher, N. (2013). Broad domain generality in focal regions of frontal and parietal cortex. *Proceedings of the National Academy of Sciences*, 110(41):16616–16621.

Geerligs, L., Gözükara, D., Oetringer, D., Campbell, K. L., van Gerven, M., and Güçlü, U. (2022). A partially nested cortical hierarchy of neural states underlies event segmentation in the human brain. *ELife*, 11:e77430.

Geerligs, L., van Gerven, M., and Güçlü, U. (2021). Detecting neural state transitions underlying event segmentation. *NeuroImage*, 236:118085.

Guntupalli, J. S., Hanke, M., Halchenko, Y. O., Connolly, A. C., Ramadge, P. J., and Haxby, J. V. (2016). A Model of Representational Spaces in Human Cortex. *Cerebral Cortex*, 26(6):2919–2934.

Hamilton, L. S. and Huth, A. G. (2020). The revolution will not be controlled: natural stimuli in speech neuroscience. *Language, cognition and neuroscience*, 35(5):573–582.

Hanke, M., Adelhöfer, N., Kottke, D., Iacovella, V., Sengupta, A., Kaule, F. R., Nigbur, R., Waite, A. Q., Baumgartner, F., and Stadler, J. (2016). A studyforrest extension, simultaneous fMRI and eye gaze recordings during prolonged natural stimulation. *Scientific Data*, 3(1):1–15.

Hanke, M., Baumgartner, F. J., Ibe, P., Kaule, F. R., Pollmann, S., Speck, O., Zinke, W., and Stadler, J. (2014). A high-resolution 7-Tesla fMRI dataset from complex natural stimulation with an audio movie. *Scientific Data 2014 1:1*, 1(1):1–18.

Hasson, U., Yang, E., Vallines, I., Heeger, D. J., and Rubin, N. (2008). A hierarchy of temporal receptive windows in human cortex. *Journal of Neuroscience*, 28(10):2539–2550.

Häusler, C. O. and Hanke, M. (2016). An annotation of cuts, depicted locations, and temporal progression in the motion picture "Forrest Gump" [version 1; referees: 3 approved]. *F1000Research*, 5:2273.

Haxby, J. V., Guntupalli, J. S., Nastase, S. A., and Feilong, M. (2020). Hyperalignment: Modeling shared information encoded in idiosyncratic cortical topographies. *elife*, 9:e56601.

Honey, C. J., Thesen, T., Donner, T. H., Silbert, L. J., Carlson, C. E., Devinsky, O., Doyle, W. K., Rubin, N., Heeger, D. J., and Hasson, U. (2012). Slow Cortical Dynamics and the Accumulation of Information over Long Timescales. *Neuron*, 76(2):423–434.

Kriegeskorte, N. and Kievit, R. A. (2013). Representational geometry: integrating cognition, computation, and the brain. *Trends in cognitive sciences*, 17(8):401–412.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Pereira, F., Burges, C., Bottou, L., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

Kurby, C. A. and Zacks, J. M. (2008). Segmentation in the perception and memory of events. *Trends in cognitive sciences*, 12(2):72–79.

Lee, H., Bellana, B., and Chen, J. (2020). What can narratives tell us about the neural bases of human memory? *Current Opinion in Behavioral Sciences*, 32:111–119.

Lerner, Y., Honey, C. J., Silbert, L. J., and Hasson, U. (2011). Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *Journal of Neuroscience*, 31(8):2906–2915.

Liu, X., Zhen, Z., Yang, A., Bai, H., and Liu, J. (2019). A manually denoised audio-visual movie watching fMRI dataset for the studyforrest project. *Scientific Data*, 6(1):1–9.

Magliano, J. P., Miller, J., and Zwaan, R. A. (2001). Indexing Space and Time in Film Understanding. *Applied Cognitive Psychology*, 15(5):533–545.

Magliano, J. P. and Zacks, J. M. (2011). The impact of continuity editing in narrative film on event segmentation. *Cognitive science*, 35(8):1489–1517.

Mariola, A., Fountas, Z., Barnett, L., and Roseboom, W. (2022). Event segmentation in continuous, naturalistic videos from model-based, data-driven, and human perspectives. *PsyArXiv*.

Nasr, S., Liu, N., Devaney, K. J., Yue, X., Rajimehr, R., Ungerleider, L. G., and Tootell, R. B. (2011). Scene-selective cortical regions in human and nonhuman primates. *Journal of Neuroscience*, 31(39):13771–13785.

Nastase, S. A., Goldstein, A., and Hasson, U. (2020). Keep it real: rethinking the primacy of experimental control in cognitive neuroscience. *NeuroImage*, 222:117254.

Newtson, D. (1973). Attribution and the unit of perception of ongoing behavior. *Journal of Personality and Social Psychology*, 28(1):28–38.

Quax, S. C., D'Asaro, M., and van Gerven, M. A. (2020). Adaptive time scales in recurrent neural networks. *Scientific reports*, 10(1):11360.

Rajimehr, R., Devaney, K. J., Bilenko, N. Y., Young, J. C., and Tootell, R. B. H. (2011). The "Parahippocampal Place Area" Responds Preferentially to High Spatial Frequencies in Humans and Monkeys. *PLoS Biology*, 9(4):e1000608.

Raut, R. V., Snyder, A. Z., and Raichle, M. E. (2020). Hierarchical dynamics as a macroscopic organizing principle of the human brain. *Proceedings of the National Academy of Sciences*, 117(34):20890–20897.

Richmond, L. L. and Zacks, J. M. (2017). Constructing Experience: Event Models from Perception to Action.

Rosano, C., Krisky, C. M., Welling, J. S., Eddy, W. F., Luna, B., Thulborn, K. R., and Sweeney, J. A. (2002). Pursuit and saccadic eye movement subregions in human frontal eye field: a high-resolution fmri investigation. *Cerebral Cortex*, 12(2):107–115.

Sanz-Aznar, J., Sánchez-Gómez, L., Bruni, L. E., Aguilar-Paredes, C., and Wulff-Abramsson, A. (2021). Neural responses to shot changes by cut in cinematographic editing: An EEG (ERD/ERS) study. *PLOS ONE*, 16(10):e0258485.

Seeliger, K., Roth, J., Schmid, T., and Hebart, M. (2021). Synthesizing preferred stimuli for individual voxels in the human visual system. *Journal of Vision*, 21(9):2311–2311.

Sengupta, A., Kaule, F. R., Guntupalli, J. S., Hoffmann, M. B., Häusler, C., Stadler, J., and Hanke, M. (2016). A studyforrest extension, retinotopic mapping and localization of higher visual areas. *Scientific Data*, 3(1):1–14.

Shin, Y. S. and DuBrow, S. (2020). Structuring Memory Through Inference-Based Event Segmentation. *Topics in Cognitive Science*, page tops.12505.

Speer, N. K., Swallow, K. M., and Zacks, J. M. (2003). Activation of human motion processing areas during event perception. *Cognitive, Affective, & Behavioral Neuroscience*, 3(4):335–345.

Speer, N. K., Zacks, J. M., and Reynolds, J. R. (2007). Human brain activity time-locked to narrative event boundaries: Research article. *Psychological Science*, 18(5):449–455.

Stephens, G. J., Honey, C. J., and Hasson, U. (2013). A place for time: the spatiotemporal structure of neural dynamics during natural audition. *Journal of Neurophysiology*, 110(9):2019–2026.

Swallow, K., Zacks, J., et al. (2004). Hierarchical grouping of events revealed by eye movements. *Psychonomic Society*, 9:81.

Willems, R. M., Nastase, S. A., and Milivojevic, B. (2020). Narratives for neuroscience. *Trends in neurosciences*, 43(5):271–273.

Woolgar, A., Jackson, J., and Duncan, J. (2016). Coding of visual, auditory, rule, and response information in the brain: 10 years of multivoxel pattern analysis. *Journal of cognitive neuroscience*, 28(10):1433–1454.

Zacks, J. M., Braver, T. S., Sheridan, M. A., Donaldson, D. I., Snyder, A. Z., Ollinger, J. M., Buckner, R. L., and Raichle, M. E. (2001a). Human brain activity time-locked to perceptual event boundaries. *Nature Neuroscience 2001 4:6*, 4(6):651–655.

Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., and Reynolds, J. R. (2007). Event perception: A mind-brain perspective. *Psychological Bulletin*, 133(2):273–293.

Zacks, J. M., Speer, N. K., Swallow, K. M., and Maley, C. J. (2010). The brain's cutting-room floor: Segmentation of narrative cinema. *Frontiers in Human Neuroscience*, 4.

Zacks, J. M. and Tversky, B. (2001). Event structure in perception and conception. *Psychological bulletin*, 127(1):3.

Zacks, J. M., Tversky, B., and Iyer, G. (2001b). Perceiving, remembering, and communicating structure in events. *Journal of Experimental Psychology: General*, 130(1):29–58.

Zadbood, A., Chen, J., Leong, Y. C., Norman, K. A., and Hasson, U. (2017). How We Transmit Memories to Other Brains: Constructing Shared Neural Representations Via Communication. *Cerebral Cortex*, 27(10):4988–5000.

Zeiler, M. D. and Fergus, R. (2014). Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 818–833. Springer.

# Supplementary Material

## A Correlations between annotation categories

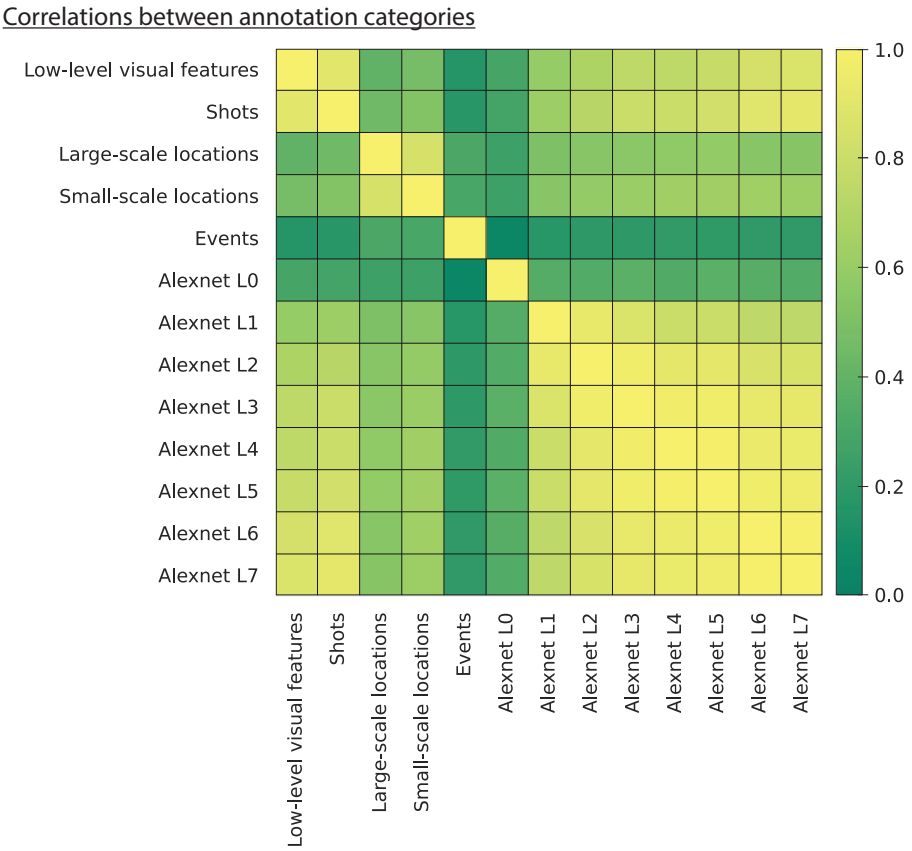


**Fig. S1.** Pearson's correlation between all the annotation categories after downsampling to fMRI volumes. The number after "Alexnet L" indicates the layer of Alexnet, where "L0" is the input layer.

Figure S1 visualizes the various correlations between all annotation correlations used in this study. Overall, event boundaries had the lowest correlation with an average correlation coefficient of 0.26. This is probably because event boundaries were identified based on button presses, thus some time passes before participants realize that a new event has started and press the button. How much time is necessary can vary between event boundaries, as some transitions are more apparent than others. Therefore, some event boundaries co-occur with for example changes in location, while others seem to be slightly delayed.

## B ROI descriptives

**Table S1.** Descriptives of the functionally defined ROIs.

| ROI | Threshold (t) | MNI center (x y z) | | | Number of voxels | Volume ($mm^3$) |
|---|---|---|---|---|---|---|
| Early visual left | 3.50 | -10.44, | -94.26 | 0.84 | 100 | 2700 |
| Early visual right | 3.39 | 9.96 | -90.33 | 3.45 | 100 | 2700 |
| PPA left | 4.50 | -22.68 | -44.31 | -11.64 | 100 | 2700 |
| PPA right | 5.79 | 25.17 | -44.22 | -12.60 | 100 | 2700 |
| RSC left | 4.36 | -13.60 | -57.78 | 10.21 | 77 | 2079 |
| RSC right | 4.55 | 14.04 | -51.78 | 9.06 | 100 | 2700 |

The left RSC only had 77 voxels rather than 100, as a slightly higher t-value gave a region of 193 voxels. This is because there were two clusters connected to each other with just one voxel. The shape and location of each ROI is visualized in Figure S2.

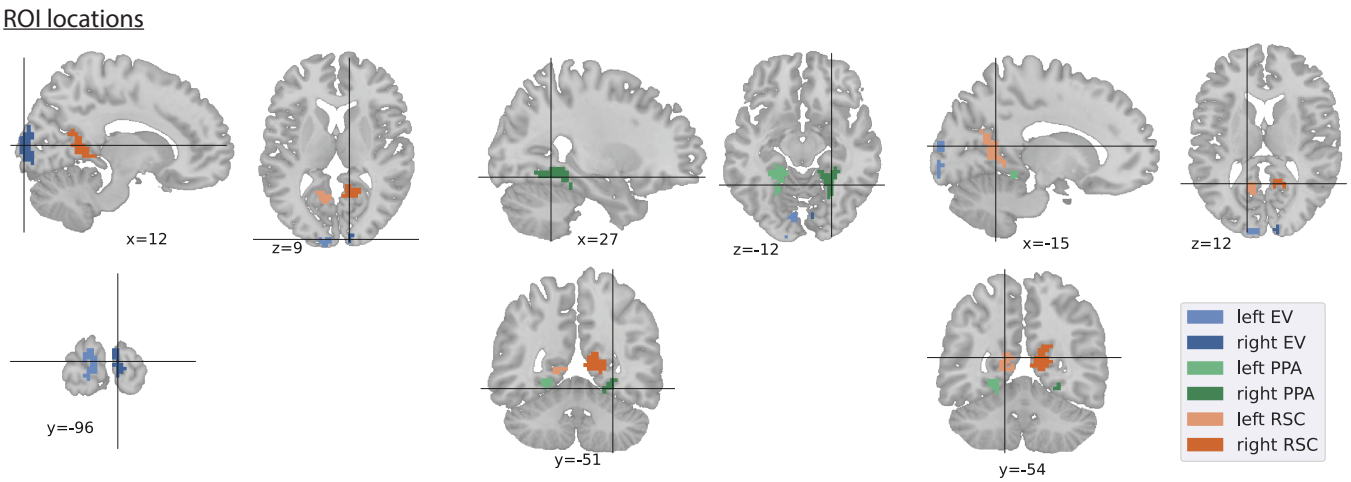


**Fig. S2.** Location of each functionally defined ROI. EV: early visual. PPA: parahippocampal place area. RSC: retrosplenial cortex.

## C Hemodynamic response delay

To compute the appropriate hemodynamic delay for aligning neural state boundaries and movie annotations, we assumed that the neural state boundaries in all ROIs are correlated with event boundaries in the stimulus, as previously found by Baldassano et al. (2017) and Geerligs et al. (2022). For each ROI, we extracted the neural states boundaries of run 4, which was not used in any other analysis, and set all timepoints with a boundary to 1 and all others to 0. We then computed the Jaccard index between these neural state boundaries and the event boundaries for a given delay. The delay was varied between 0 and 10 s with steps of 0.5 s, which was applied before downsampling the event boundaries from seconds to fMRI volumes. This analysis is in line with Geerligs et al. (2022), who showed that the optimal delays across the cortex were similar to previously found BOLD signal delays. The results of this analysis can be found in Figure S3. Based on these results, we chose to use a delay of 4.5 s, as this was the optimal delay for the left and right PPA and the right early visual, in between two optimal delays of the left RSC, and close to the optimal delay of the right RSC. Left early visual however is an outlier and had an optimal delay of 6 s.

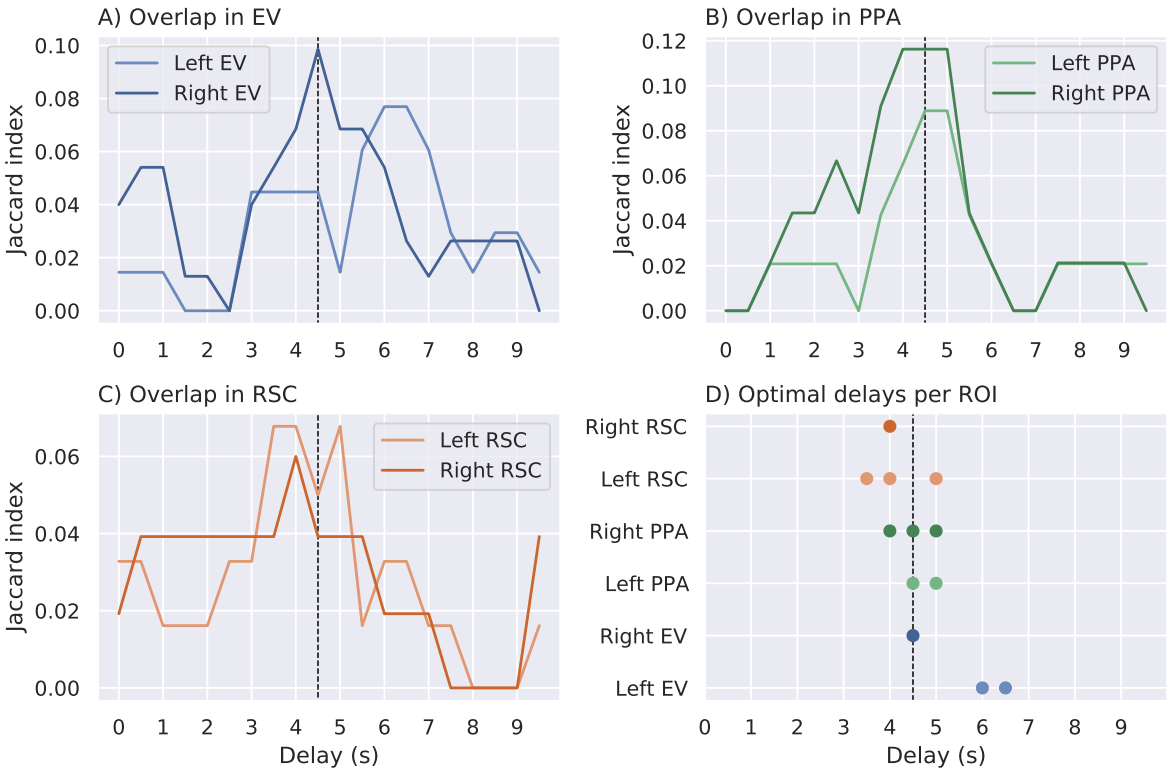


**Fig. S3.** Jaccard index between the event boundaries and the neural state boundaries of run 4 over various delays. A,B,C) the Jaccard index over delay for each ROI. D) the optimal delay per ROI. The dashed lines at 4.5 s show the delay that we used in the main analysis. EV: early visual. PPA: parahippocampal place area. RSC: retrosplenial cortex.

## D Number of neural state boundaries and annotation boundaries

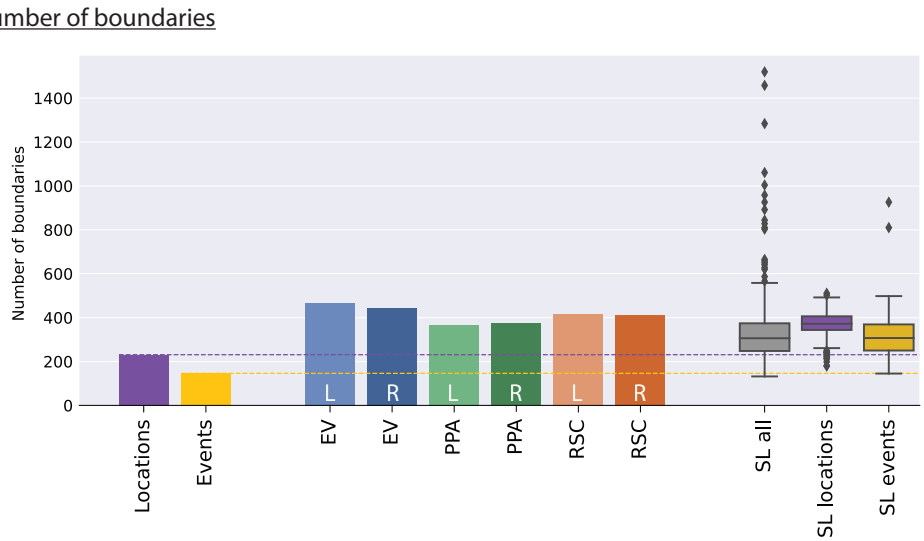


**Fig. S4.** Throughout the movie stimulus, 231 changes in small-scale locations and 146 event boundaries occurred (excluding run number 4). All ROIs and many searchlights had a higher number of neural state boundaries. Locations: small-scale locations. EV: early visual. PPA: parahippocampal place area. RSC: retrosplenial cortex. SL all: all searchlights. SL location: searchlights with a significant alignment between their neural state boundaries and changes in small-scale locations, after FDR-correction. SL events: searchlights with a significant alignment between their neural state boundaries and event boundaries, after FDR-correction.

Figure S4 summarizes the number of small-scale location changes, event boundaries, and neural state boundaries across the movie stimulus, excluding run number 4. In total, there were 231 changes in small-scale location, and 146 event boundaries. Both of these numbers are much lower than the number of neural state boundaries in each ROI (left early visual: 467, right early visual: 444, left PPA: 367, right PPA: 373, left RSC: 416, right RSC: 412). When taking all searchlights throughout the cortex, the number of neural state boundaries per searchlight ranged from 132 to 1520. When only selecting those searchlights that had a significant alignment between locations and neural states, 277 out of 284 searchlights had a higher number of neural state boundaries than the number of changes in location (231), with a minimum of 180 and a maximum of 511. When only selecting searchlights with a significant alignment with events, only one searchlight had fewer neural state boundaries (145) than number of event boundaries (146), and the remaining searchlights ranged from 152 to 926 neural state boundaries.
Although many areas show an alignment between their neural states and locations or events, there are often more neural state boundaries than moments in the movie stimulus with a relevant change. This indicates that neural states in specific areas reflect more than just locations or events.

## E Results whole-brain analysis with large-scale locations

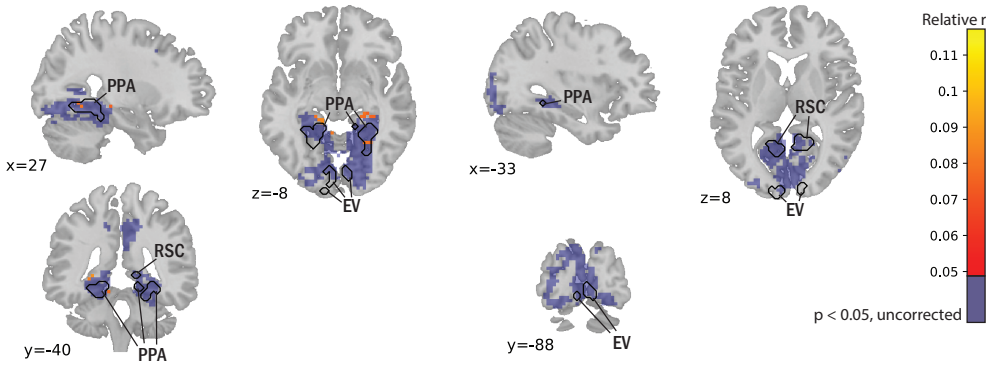


**Fig. S5.** Voxels with a significant alignment between neural state boundaries and large-scale location changes were sparse and could be found close to the PPA, but these clusters were small. Values represent relative r values between neural state boundaries and large-scale locations across the brain, while correcting for low-level visual features, shots, events, and 8 layers of Alexnet. The relative r is defined as the actual Pearson's correlation coefficient divided by the Pearson's correlation coefficient that can maximally be expected given the number and strengths of the neural state boundaries and the annotations. EV: early visual. PPA: parahippocampal place area. RSC: retrosplenial cortex.

In the main text we showed the association between neural states boundaries across the brain and small-scale location changes (see Figure 1B). Here, we also visualized the association with large-scale location changes (Figure S5). For large-scale locations, some voxels with a significant alignment could be found close to the PPA. Compared to the whole-brain analysis with the small-scale locations, the clusters in the PPA were much smaller, and the cluster in left V2 disappeared.

## F Results without correction for covariates

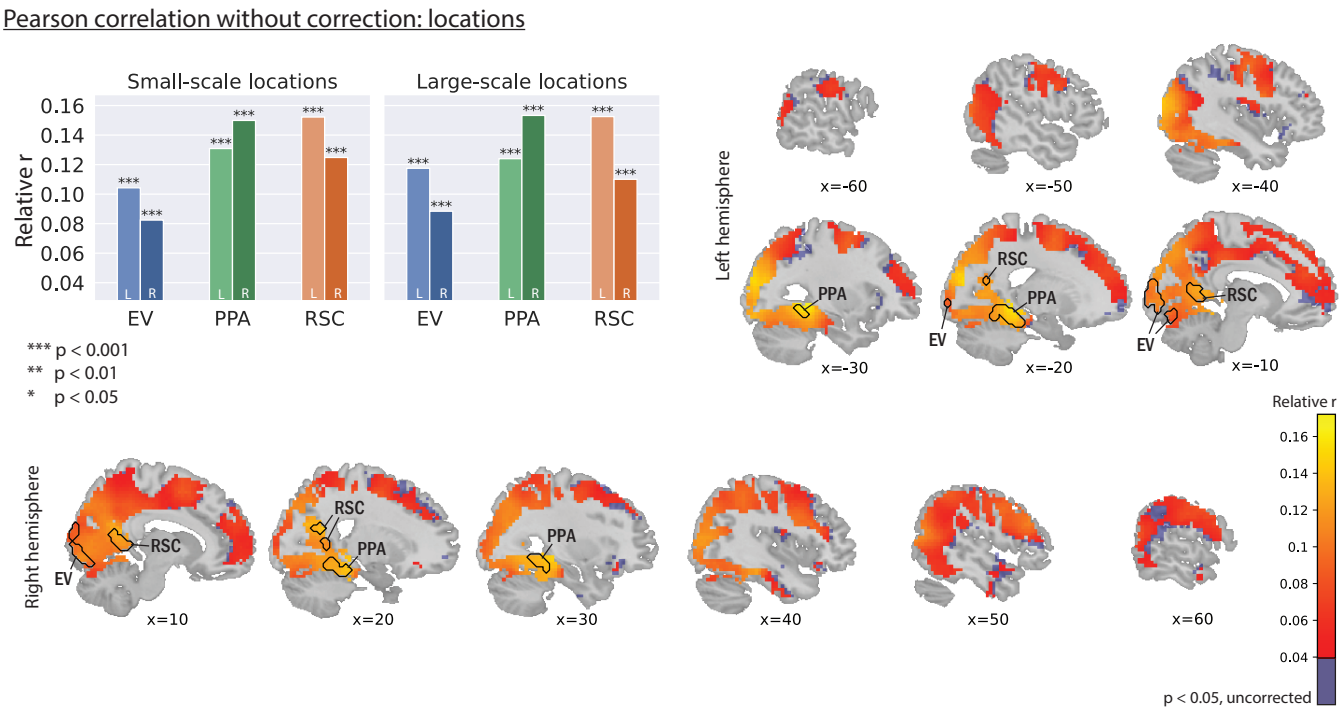


**Fig. S6.** When not correcting for covariates, neural states in all ROIs and in many areas across the cortex showed a significant alignment with locations. The values represent the relative r between neural state boundaries and changes in location. The relative r is defined as the actual Pearson's correlation coefficient divided by the Pearson's correlation coefficient that can maximally be expected given the number and strengths of the neural state boundaries and the annotations. EV: early visual. PPA: parahippocampal place area. RSC: retrosplenial cortex. The brain plots only show the results of the whole-brain analysis for small-scale locations. The results of the whole-brain analysis without covariate correction for large-scale locations are not depicted here.

To investigate the importance of correction for covariates, we also reran our main analyses with Pearson's correlation as the measurement rather than partial correlation. For changes in location, neural state boundaries now showed a significant alignment in all ROIs ($p < 0.001$), and this alignment was present across the vast majority of the cortex, including early visual areas as well as prefrontal areas (Figure S6). A similar pattern was found in the alignment between neural state boundaries and changes in low-level visual features (Figure S7): this alignment was significant in all ROIs ($p < 0.001$). Furthermore, voxels in which neural state boundaries aligned with changes in low-level visual features now extended across higher-level areas, with some clusters being found in the frontal areas and the middle cingulate cortex. The clusters of significant alignment between event boundaries and neural state boundaries were larger compared to the analysis with covariates, with significant voxels now also being present in low-level visual areas (Figure S8).

Taking all annotation categories together, without any correction for covariates many areas showed overlap between two or more categories, with many visual areas showing a significant correlation between neural state boundaries and changes in low-level visual features, location changes, and event boundaries. Overall, regions that showed significant alignment in our partial correlation analysis still had a significant alignment when not correcting for any covariates, but many additional regions now showed the same alignment. Thus, correcting for covariates is necessary to exclude the possibility that the significant alignment found in a specific region or voxel is actually related to another aspect of the movie.

## Pearson correlation without correction: low-level visual features

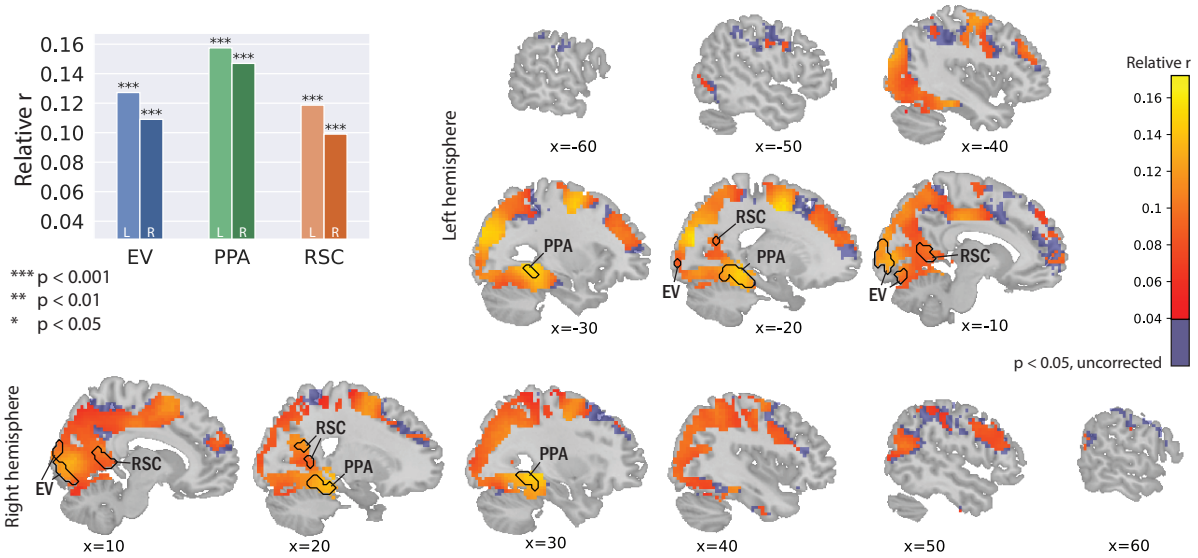


**Fig. S7.** When not correcting for covariates, neural states in all ROIs and in many areas across the cortex showed a significant alignment with low-level visual features. The values represent the relative r between neural state boundaries and changes in location. The relative r is defined as the actual Pearson's correlation coefficient divided by the Pearson's correlation coefficient that can maximally be expected given the number and strengths of the neural state boundaries and the annotations. EV: early visual. PPA: parahippocampal place area. RSC: retrosplenial cortex.

## Pearson correlation without correction: events

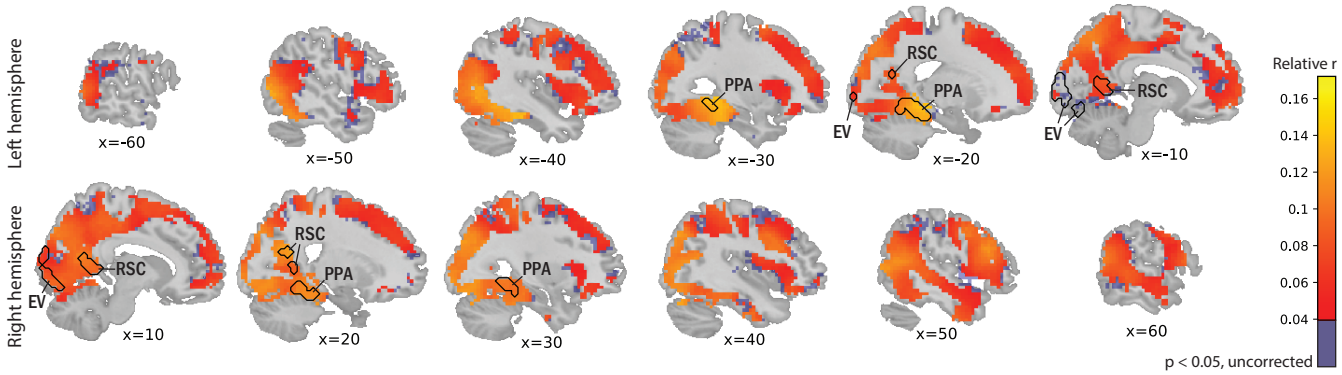


**Fig. S8.** When not correcting for covariates, neural states in all ROIs and in many areas across the cortex showed a significant alignment with events. The values represent the relative r between neural state boundaries and event boundaries. The relative r is defined as the actual Pearson's correlation coefficient divided by the Pearson's correlation coefficient that can maximally be expected given the number and strengths of the neural state boundaries and the annotations. EV: early visual. PPA: parahippocampal place area. RSC: retrosplenial cortex.

## G Shot analysis

Although not part of the main analysis, we wondered whether there was anything special about shots in movies, other than the changing visual features. Therefore, we performed the ROI and whole-brain analyses on shots as well, with low-level visual features, small- and large-scale locations, events, and 8 layers of Alexnet as covariates. We also performed the same analyses with Pearson's correlation rather than partial correlation. The results of these analyses are found in Figure S9. In the ROI analysis, all correlation coefficients as a result of partial correlation were negative and not significantly higher than the null distribution. In the whole-brain analysis with partial correlation, no voxels survived correction for multiple comparisons. When we do not correct for the covariates, the correlation coefficients were above zero and significant in all ROIs ($p = 0.000$) and across many brain areas including the visual cortex and middle cingulate cortex.

Thus, although shots showed significant alignment with neural state boundaries in multiple areas of the brain, these alignments disappeared when correcting for changes in visual features of multiple levels of complexity, event boundaries, and changes in location. We can therefore conclude that shots do not show any additional brain pattern changes in terms of neural states, on top of the changes in visual features.

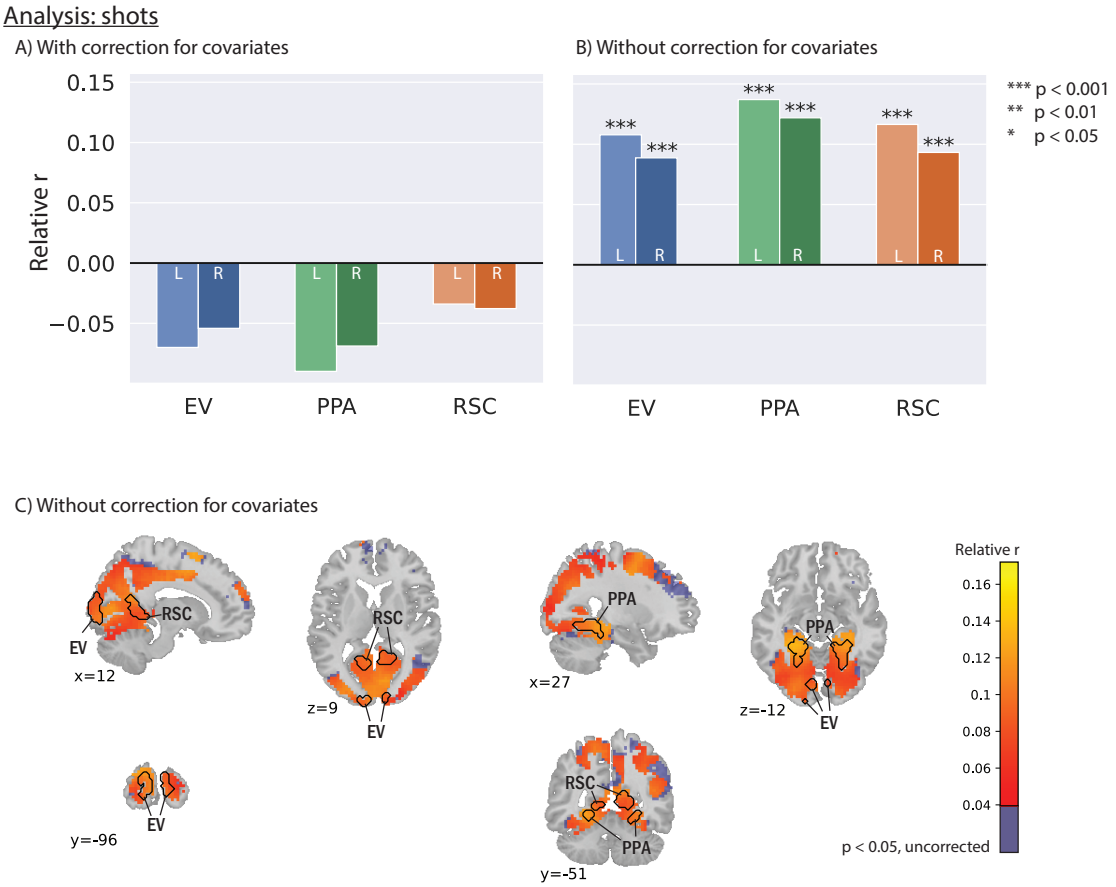


**Fig. S9.** When there was no correction for covariates, a significant alignment between neural states and shots was present in all functionally defined ROIs, as well as in many voxels of the visual cortex. However, when correcting for low-level visual features, locations, events, and 8 layers of Alexnet, all significant alignments disappeared. The values represent the relative r between neural state boundaries and event boundaries. The relative r is defined as the actual Pearson's correlation coefficient divided by the Pearson's correlation coefficient that can maximally be expected given the number and strengths of the neural state boundaries and the annotations. EV: early visual. PPA: parahippocampal place area. RSC: retrosplenial cortex. A) ROI analysis with correction for covariates. B) ROI analysis without any covariate correction. C) Thresholded whole-brain results without any covariate correction. The whole-brain results of partial correlation analysis with shots are not shown here, as no voxel survived the correction for multiple comparisons.