Neuroscience

Decoding movie content from neuronal population activity in the human medial temporal lobe

Reviewed Preprint
v1 • August 13, 2025
Not revised

Franziska Gerken, Alana Darcher, Pedro J Gonçalves, Rachel Rapp, Ismail Elezi, Johannes Niediek, Marcel S Kehl, Thomas P Reber, Stefanie Liebe, Jakob H Macke, Florian Mormann, Laura Leal-Taixé

Dynamic Vision and Learning Group, Technical University of Munich, Munich, Germany • Department of Epileptology, University Medical Center of Bonn, Bonn, Germany • Machine Learning in Science, Excellence Cluster Machine Learning and Tübingen AI Center, University of Tübingen, Tübingen, Germany • VIB-Neuroelectronics Research Flanders (NERF), Leuven, Belgium • imec, Leuven, Belgium • Machine Learning Group, Technical University of Berlin, Berlin, Germany • Department of Experimental Psychology, University of Oxford, Oxford, United Kingdom • Faculty of Psychology, UniDistance Suisse, Brig, Switzerland • Department of Epileptology and Neurology, University Hospital Tübingen, Tübingen, Germany • Empirical Inference, Max Planck Institute for Intelligent Systems, Tübingen, Germany

eLife Assessment

This **valuable** study demonstrates that it is possible to decode information about characters and locations from single-unit responses in the human brain to a narrative movie, using a **convincing** technical approach to capture information in population-level dynamics. The study introduces an exciting dataset of single-unit responses in humans during a naturalistic and dynamic movie stimulus, with recordings from multiple regions within the medial temporal lobe. Using both a traditional firing-rate analysis as well as a population decoding analysis to connect these neural responses to the visual content of the movie, they show that in this dataset, the decoding of semantic scene features (e.g., the person currently on screen), but not scene transitions, is surprisingly driven by classically non-responsive neurons. Based on these findings, the authors argue that dynamic naturalistic semantic information may be processed within the medial temporal lobe at the population level.

https://doi.org/10.7554/eLife.106758.1.sa4

Abstract

The human medial temporal lobe (MTL), a region implicated in memory and high-level cognition, contains neurons that respond selectively to stimuli belonging to specific categories, such as individual people, landmarks, or objects. However, these neurons have been largely studied via static, isolated presentations of stimuli. Therefore, it is unclear how neurons in the MTL respond to rich stimuli such as movies, and which dynamical stimulus features can be retrieved from neuronal population spiking activity. We studied single-unit responses from 2286 neurons recorded from the amygdala, hippocampus, entorhinal cortex, and parahippocampal cortex of 29 intracranially implanted patients during the presentation

of an 83-minute movie. We found only a few individual neurons that exhibited a classic selective response to semantic features. However, we successfully decoded the presence of characters, settings, and visual transitions from neuronal population activity. The information relevant for decoding varies across regions depending on the feature category, as visual transitions could be decoded from subsets of neurons with selective responses, whereas character and location features relied on distributed representations. Our results demonstrate an approach for reliably decoding movie features in the human MTL, and suggest that the brain uses a population code when representing character and location features.

Introduction

The human medial temporal lobe (MTL) plays an integral role in the representation of semantic information. Single neurons in the MTL exhibit strong and highly selective tunings to categories, such as faces or locations [1,2], and to specific concepts, such as individual celebrities or objects [3,4]. These semantically tuned neurons relate to the processing of information at a conscious, declarative level, as their activity varies depending on perception [5–7]—when images are presented but unseen, these neurons exhibit reduced and delayed spiking compared to consciously seen images [8]. Such cells also support the formation of new memories [9], and are involved in the retrieval of previously encoded experiences [10,11].

Studies investigating the representation of semantic information in the human MTL have mostly focused on characterizing these neurons individually, often without considering their population dynamics. Many of these studies have identified semantically tuned cells by screening for stimulus-selective responses to static images depicting isolated persons or objects [12,13]. While this approach is effective for probing specific functional properties of individual neurons, it limits the generalizability of findings to more complex and dynamic contexts. Naturalistic and dynamic stimuli, such as movies, provide closer approximations to real-world environments, but also pose substantial challenges, as they are presented continuously and depict complex evolving scenes. Several studies have used functional magnetic resonance imaging (fMRI) to study neural responses to natural movies. One line of work has focused on representations in early visual areas [14,15] and across the cortex [16,17]. Another line of work has identified synchrony in the brain states of separate individuals viewing a common movie [18,19]. Such states appear to be particularly well aligned in individual brains surrounding transition events within an ongoing continuous stimulus, as reported in fMRI studies [20,21], intracranial field potentials along the human ventral visual pathway [22], and in single neurons in the human parahippocampal gyrus, hippocampus, and amygdala [11]. Most single neuron studies which use movies as stimuli have been based on only short ‘snippets’ of movies, and have so far largely examined representation on the level of the individual neuron [23] with few having addressed frame-wise representations in longer movie sequences [24]. However, it remains an open question of how populations of neurons in the MTL collectively respond to naturalistic visual stimuli, such as those encountered in real-world environments, and which features of population activity encode information about specific components of such stimuli.

In this study, we investigate how information embedded in a naturalistic and dynamic stimulus is processed by neuronal populations in the human medial temporal lobe (MTL). Specifically, we asked the following questions: (1) Which aspects of a movie’s content can be decoded from neuronal activity in the MTL? (2) Which brain regions are informative for specific stimulus categories (e.g. visual transitions or characters)? (3) Is the relevant information distributed across the neuronal population? We recorded the activity of neurons from patients with intracranially implanted electrodes as each watched the full-length commercial movie “500 Days of Summer”. Our dataset is unique in both size and duration: we recorded a total of 2286 neurons across 29

patients during the complete presentation of the movie. To analyze the relationship between the neuronal activity and the film's content, we labeled the presence of main characters, whether a scene was indoors or outdoors, and visual transitions of the movie on a frame-by-frame basis.

We introduce a machine learning-based decoding pipeline that decodes a movie's visual content from population-level neuronal activity. At the single-unit level, individual neurons generally lacked reliable responses to the visual features, and we observed consistent stimulus-related changes in firing rates primarily during visual transitions. However, at the population level, we achieved strong decoding performance across all visual features. For visual transitions, neurons exhibiting consistent changes in activity played a key role in population-level decoding. In contrast, no similar pattern emerged when decoding character identities. By analyzing the contributions of individual neurons, we identified distinct subsets of neurons that influenced decoding performance, extending beyond the subsets identified by stimulus-aligned changes in firing activity. Remarkably, we found that restricting the analysis to a substantially smaller subset of these key neurons was sufficient to replicate the full population-level decoding performance. Taken together, our findings show that information about dynamic stimuli can be decoded from neuronal population activity, even in the absence of strong single-neuron selectivity.

Results

We recorded from 29 patients (17 female, ages 22 → 63) as each watched the movie *500 Days of Summer* (83 minutes) in its entirety (**Fig. 1a**). Patients were bilaterally implanted with depth electrodes for seizure monitoring, and spiking activity was recorded from a total of 2286 single- and multi-neurons across the amygdala (A; 580 neurons, 25.37%), hippocampus (H; 794, 34.73%), entorhinal cortex (EC; 440, 19.25%), parahippocampal cortex (PHC; 373, 16.32%), and piriform cortex (PIC; 99, 4.33%) (**Fig. 1b**). We pooled the neurons across patients (distribution shown in Supp. Fig. S1) and performed subsequent analyses on the resulting population. Due to the low number of neurons relative to the complete population, the PIC was excluded from subsequent region-wise analyses.

To determine which features of the movie are represented in MTL activity, we obtained frame-wise annotations of character presence, indoor/outdoor setting, and the occurrence of a visual transition, i.e. camera cuts and scene cuts (similar to "soft" and "hard" boundaries [11]) (**Fig. 1c**). All annotations concern the visual occurrence of a given feature in the frame, and do not consider feature content of the movie's audio. We used a mixture of manual and automated methods for obtaining these annotations (see Methods; sample frames in Supp. Fig. S2). For character-related content, we restricted the analysis to those characters most relevant to the movie's narrative (Tom, Summer, McKenzie) as determined by screentime. For the analysis of location, we investigate indoor and outdoor settings. Note that these two features are mutually exclusive, and the annotation of both are combined into a single label.

Stimulus-aligned responsive neurons found primarily in parahippocampal cortex

To investigate whether the visual features of the movie are encoded at the level of individual neurons, we analyzed stimulus-evoked changes in the firing rate of neurons after the onset of characters, indoor and outdoor scenes, and visual transitions, for each MTL subregion separately (**Fig. 2a**, Supp. Fig. S3 - S5). We classified each individual neuron as responsive or non-responsive according to a previously established criterion [8] which compares the spiking activity after stimulus onset to that of a baseline period, adapted to the dynamic presentation. Specifically, for each annotated feature, we identified all instances where the feature appeared following at least 1000 ms of absence and remained continuously present for at least 1000 ms. We applied a cluster permutation test [26] to find time-points in which the firing rate between the sets of

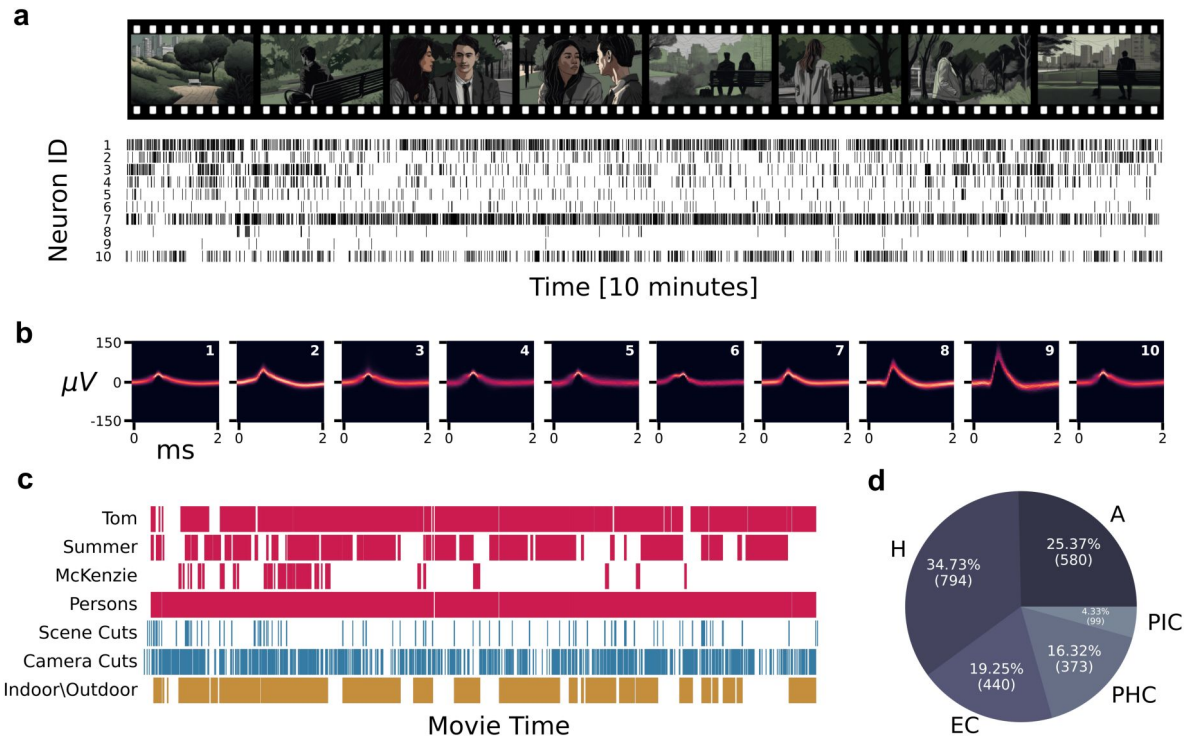

Figure 1.

Overview of dataset, features, and decoding approach.

a) Example of the recorded neuronal activity. Patients watched the complete commercial film *500 Days of Summer* while neuronal activity was recorded via depth electrodes. Top row: example movie frames. Due to copyright, the original movie frames have been replaced with images generated using stable diffusion [25]. Bottom row: spike trains from ten amygdala neurons of a single patient, where each row shows data from an individual neuron (corresponding ID number given as a label). b) Spike density plot showing the waveforms of each neuron in a (corresponding neuron ID given in top right). Neurons shown include both single- and multi-neurons. c) Distribution of labels across the entire movie (runtime: 83 minutes). Occurrences of character-related features are in magenta, visual transition events in blue, and location events in yellow. d) Distribution of the 2286 neurons across the recorded regions (A: amygdala, H: hippocampus, EC: entorhinal cortex, PHC: parahippocampal cortex, PIC: piriform cortex) for all 29 patients.

responsive and non-responsive neurons differ (see Methods for additional details). We found individual neurons with significant stimulus-evoked responses in all regions for visual transitions, Persons, and the characters Tom and Summer, including Face-only onsets (bin-wise Wilcoxon signed-rank test, Simes-corrected, $\omega = 0.001$, see Supp. Table S1). Of these stimulus-evoked changes, neurons in the parahippocampal cortex responded to the largest set of stimulus features (Summer, Persons, Camera Cuts, Scene Cuts, Outdoor; $p \uparrow 0.001$, cluster permutation test), as compared to the hippocampus (Tom, Camera Cuts, Outdoor), amygdala (Camera Cuts), and entorhinal cortex (None). Over half of the parahippocampal neurons responded to Camera Cuts during the movie (193/373, 51.74%, bin-wise Wilcoxon signed-rank test, Simes-corrected, $\omega = 0.001$) (**Fig. 2b**), and stimulus-evoked changes in firing were consistent across the set of responsive neurons ($p \uparrow 0.001$, cluster permutation test). For the remaining regions, stimulus-evoked modulations were less consistent across individual neurons (see Supp. Fig. S5a). Comparatively few parahippocampal neurons responded to Scene Cuts (25/373, 6.70%), although there was nonetheless a consistent pattern of modulation across responsive neurons ($p = 0.018$, cluster permutation test) (**Fig. 2b**). A similar pattern was observed for the onset of Outdoor scenes in the parahippocampal cortex (4/373, 1.07%). No responses were detected for the onset of indoor scenes. For characters, significant stimulus-evoked modulation in the responsive neurons were only observed for Summer (PHC, $p = 0.006$, cluster permutation test) and, to a lesser extent, Tom (H, $p = 0.022$, cluster permutation test) (**Fig. 2a**, Supp. Fig. S3). Taken together, the parahippocampal cortex contained neuronal subsets that showed a consistent pattern of increased firing after the onset of a visual transition (Camera or Scene Cut) within the movie. Characters and character faces evoked a clear change in firing activity at the subpopulation-level in the parahippocampal cortex, and sparingly in the hippocampus, but not in other tested regions.

Decoding of semantic content from population responses

The observed pattern of responses in individual neurons aligned to feature onsets suggests that these cells primarily carry information relating to visual transitions, and, to a lesser degree, the characters Summer and Tom, and Persons. However, previous work suggests that there may be differences in the coding capacity at the single-neuron and population level [27]. Could there be stimulus-related information represented at the population level that is not apparent in the responses of individual neurons? To explore this, we decoded character presence, location, and visual transitions from the aggregated activity of the entire neuronal population using a recurrent neural network designed to capture both within-neuron and between-neuron activity patterns.

Decoding from neuronal activity

We aligned the activity of neurons across patients using the movie frames as a common time reference, generating a single neuronal pseudo-population, as done in previous work decoding from populations of single neurons [28,29]. We first evaluated single-neuron decoding performance for the main character, Summer, by fine-tuning a firing activity threshold for each neuron. Using the total spike count around frame onset (spanning 800 ms before and after onset), we optimized a threshold to the validation set from each cell and subsequently predicted Summer's appearance in the test set. This established a performance baseline for decoding character presence solely from the firing of individual neurons, without machine learning-based decoding algorithms. Decoding performances obtained by simply applying a threshold to single-neuron activity did not reliably predict the content, as the majority of neurons in the population performed near chance level (**Fig. 3b**). We then extended this analysis to a population-based approach, employing a Long Short-Term Memory (LSTM) network [30], a deep neural network well-suited for processing dynamic sequential data (**Fig. 3a**). This population-based approach improved decoding performance for Summer, surpassing the performances achieved by individual neurons alone. Decoding performances for all tested labels significantly exceeded chance level (zero for Cohen's Kappa), as shown in **Fig. 3c**. The Persons label achieved the highest performance (mean ± standard error of the mean, SEM: 0.36 ± 0.05), followed by the

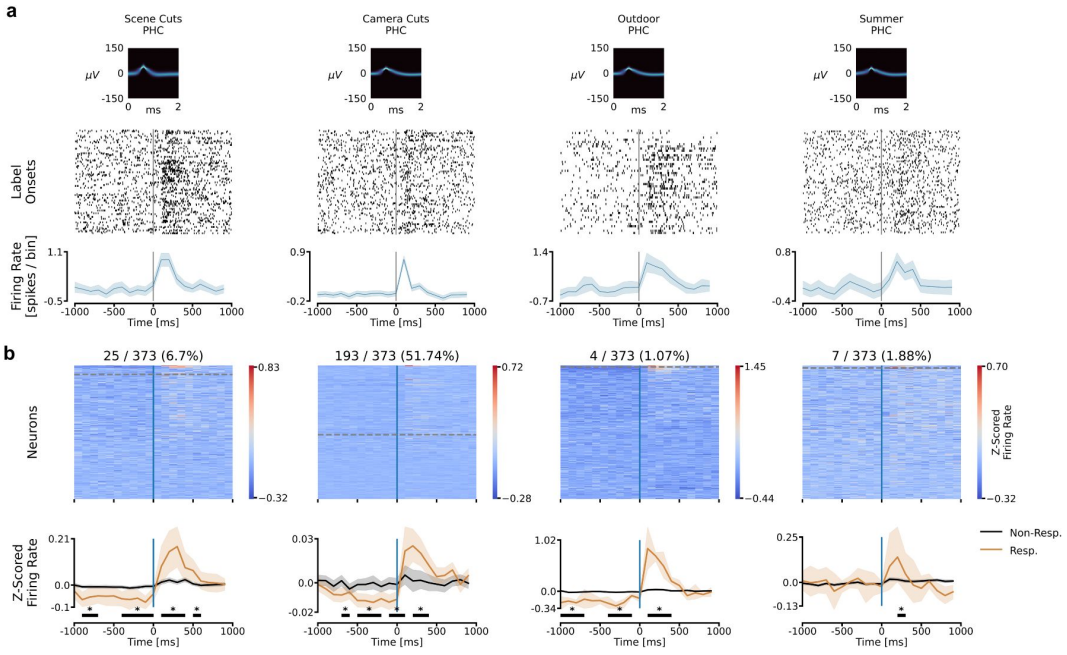

Figure 2.

Responsive single-neurons in the parahippocampal cortex.

a) Example peri-stimulus activity for representative parahippocampal (PHC) neurons, for labels with a significant PHC response. Upper plots: spike density plot showing the waveforms a given responsive neuron (label name given as title). Middle plots: spike time rasters showing the neuron's activity surrounding the onset of the corresponding label throughout the movie. Note: onsets for Scene Cuts and Camera Cuts were randomly subsampled to match the number of Summer appearances. Lower plots: average firing rate across 100 ms bins, for 1000 ms before and 1000 ms after the onset or event. Solid lines show the mean across all neurons, within group, and the transparent area shows the 95% confidence interval. b) Region-wise single-neuron activity surrounding the onset of labeled entity. Number of cells exhibiting a significant response over the total number of PHC cells are given as the title, followed by the corresponding percentage. Upper plots (heatmaps): averages of peri-stimulus spike rates per neuron (spikes per 100 ms bin, z-scored across the pseudotrial) for 1000 ms before and 1000 ms after label onset. Each row of the heatmap represents the average binned activity for one neuron. Neurons are sorted in descending order by the p-value of the response—the dotted grey line shows the threshold for responsive neurons ($p \rightarrow 0.001$). Lower plots (line plots): average z-scored firing rate across bins. Neurons are separated into responsive (orange line) and non-responsive (black line). Solid lines show the mean for each group of neurons (responsive vs. non-responsive) and the transparent area depicts the 95% confidence interval. Significant differences between the responsive and non-responsive firing rates are shown as solid black lines (*, $p \rightarrow 0.05$, cluster permutation test).

location-related label Indoor/Outdoor (0.31 ± 0.06). Note that in contrast to the single unit analysis, there is no need to conduct two separate trainings for the Indoor/Outdoor label since the label combines both features and the network implicitly learns to differentiate between the two. Character-specific labels (Summer, Tom, McKenzie) showed comparable performances, ranging between 0.23 and 0.32. Labels related to visual transitions in the movie exhibited the lowest but nonetheless significant performances of 0.2 ± 0.03 and 0.18 ± 0.01, respectively. Decoding results were consistent across metrics (F1 Score, PR-AUC, and AUROC shown in Supp. Table S2). We evaluated the statistical significance of our results by randomly permuting the test set labels (N = 1000), demonstrating that all decoding performances were significantly above chance level at an alpha level of 0.001 (see Methods). To ensure that this significance was not an artifact of the permutation procedure, we conducted an additional test by circularly shifting the neuronal data relative to the movie features (see Methods for more detail). This approach preserved the temporal structure of each dataset while disrupting the relationship between the neuronal data and the stimulus. Results were consistent between random permutation and circular shifts (Supp. Fig. S11).

To further test the hypothesis that decoding is based on population activity rather than individual neurons, we also trained an LSTM network on individual neurons. From the 46 neurons identified as responsive to Summer in the separate single neuron analysis, we selected a subset of neurons, ensuring an even distribution across both patients and regions. Similar to the threshold model, most models exhibited minimal prediction performances (see Fig. S6). In summary, our decoding network achieved consistent and statistically significant decoding performance on the population level exceeding chance level for all labeled movie features.

Choice of architecture

Our pipeline uses a recurrent neural network (LSTM) to process spiking data as a time series of event counts. We binned spike counts into 80 ms intervals, covering 800 ms before and after label onset, creating sequences with a total length of 20 and a dimensionality of 2286 neurons. We trained a separate model for the prediction of each label, with individually optimized hyperparameters. Given the high degree of imbalance for some labels (i.e., the character McKenzie only appears in 10 % of the movie's frames), we oversampled the minority class during training to mitigate the effects of the uneven distribution. We additionally employed a 5-fold nested cross-validation procedure and carefully selected samples to avoid correlations between samples, as discussed in more detail in the following section. See Methods for additional details regarding model training and architecture.

We compared the LSTM's decoding performance to a simpler logistic regression model, i.e. a linear method that does not consider the neuronal activity as a sequence of spike counts, and therefore ignores the temporal information and non-linear dynamics (**Fig. 3c**). Apart from this, the setup and data split for both pipelines were identical. The logistic regression model showed lower performances for the Scene and Camera Cut features (by 0.1 and 0.07, respectively), whereas no drop for character-related or location-related features was observed. To test our hypothesis that temporal information within spike sequences influences visual transition decoding, we assessed trained models using temporally-modified test data (sequence order of the spike trains was randomly shuned, see Methods for more details) (**Fig. 3d**). A pattern consistent with the LSTM decoding results emerged, with a noticeable decline in performance, especially for the visual transition labels.

Avoiding spurious decoding performance by introducing temporal gaps

Since each frame of the movie shares a high degree of similarity with neighboring frames, we controlled for the temporal correlations in the annotated features induced by the continuous nature of the stimulus. We divided the dataset into training, validation, and test sets, ensuring a

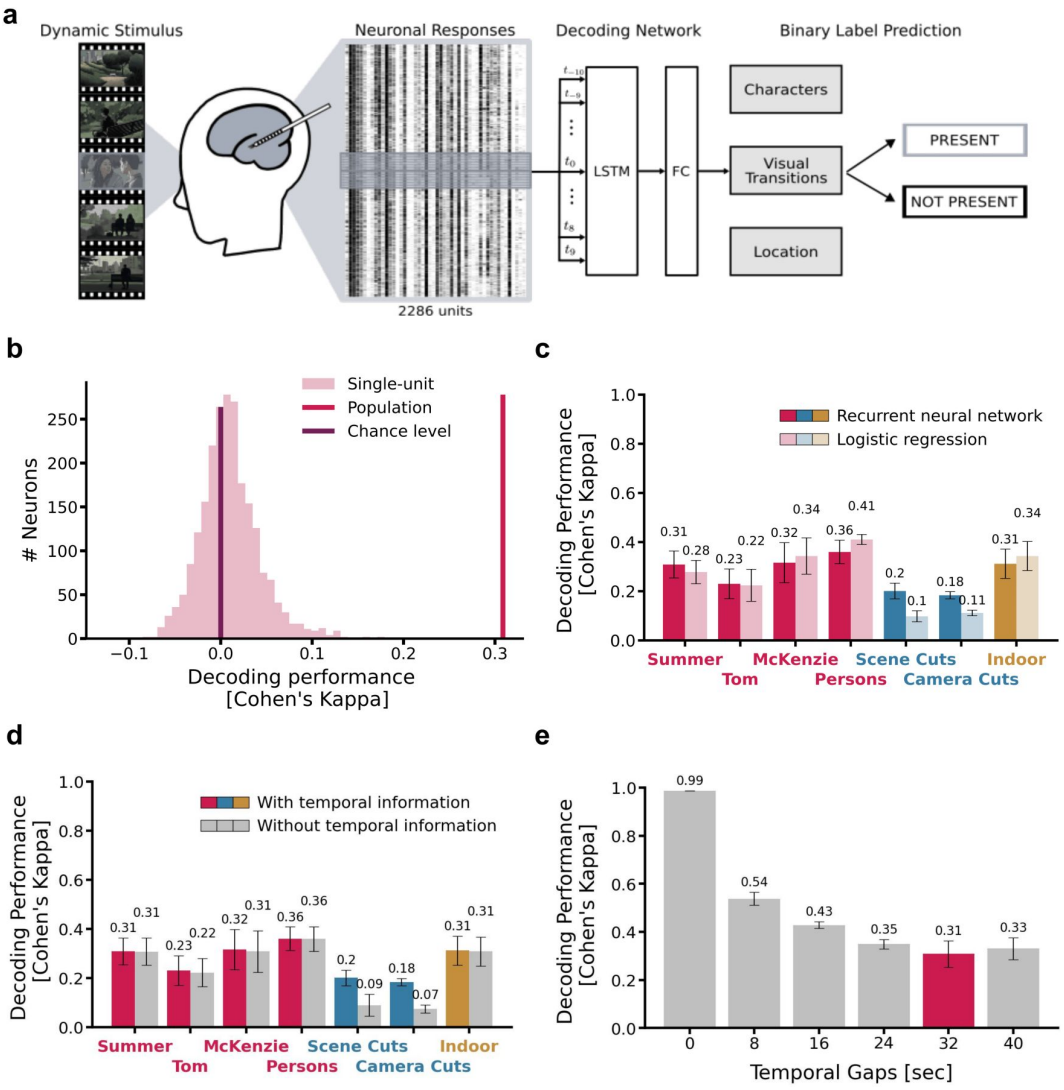

Figure 3.

Categories of labels can be decoded from the neuronal population activity.

a) Overview of the neuronal decoding pipeline. Spiking data (individual neurons shown as columns) was sectioned into 1600 ms sequences, 800 ms before and after a frame onset (purple highlight; bins shown as horizontal lines, not to scale), and given as input to a two-layer Long Short-Term Memory (LSTM) network. The output of the fully connected layer (FC) predicts the presence of a given label in a frame. b) Assessment of individual-neuron decoding performance by classifying data samples into positive or negative predictions for the label Summer based on the firing activity of a neuron. c) Decoding performances on labels of the movie (reported performance using Cohen's Kappa, mean performance across five different data splits, error bars indicate standard error of the mean). Labels fall into one of three categories—characters (pink), visual transitions (blue), or location (yellow)—with a separate model for each label. All performances were significantly better than chance level with an alpha level of 0.001. Decoding performances for the logistic regression model in lighter colors. d) Impact of temporal information in spike trains for recurrent neural networks. Trained models were evaluated using temporally altered test data (sequence order shuned, repeated 100 times). Colored bars depict performance without shuning, while grey bars represent shuned scenarios (reported performance using Cohen's Kappa, results show the mean performance across five different data splits, variance given as standard error of the mean). e) Decoding performance of the main movie character (Summer) for different temporal gap sizes between samples of training, validation, and test sets. Colored temporal gap of 32 s indicates the chosen gap size for all reported performances in this study.

gap of 32 s between samples from different sets to minimize temporal correlations (split visualized in Supp. Fig. S7). We investigated the impact of these gaps by decoding the character Summer using varying gap lengths while keeping the number of samples comparable. We observed that smaller gaps result in substantially higher decoding performance on the test set, raising concerns about potential data leakage between training and test sets. For instance, a random split without any temporal gaps achieves an almost perfect score of 0.99 ± 0.0005. However, as temporal gap sizes increase to 32 s, the performance drops precipitously to 0.31 ± 0.06 (**Fig. 3e**; additional metrics in Supp. Fig. S9). This might explain the higher decoding performance for a comparable task in Zhang et al. [24], which did not report the use of temporal gaps for model evaluation. All subsequently reported results refer to the performance on the held-out test data using 5-fold cross-validation, with data splits incorporating the most conservative temporal gap of 32 s (see Methods). Our analysis underscores the importance of appropriate architecture selection and careful data preparation in complex datasets such as ours, as these choices can exert a significant impact on the results.

Patient-wise decoding performance

The neuronal population analyzed thus far has been pooled from 29 patients, yielding a total of 2286 neurons. Decoding from a pooled population, rather than from individual patients, improves network stability by aggregating activity across a larger neuronal set and enhances the signal-to-noise ratio. However, using this pooled population (or "pseudo-brain") obscures the patient-wise contributions to decoding, which could vary due to the difference in neurons recorded per patient (units per patient range from 30 to 137) or due to differences in the semantic space of recorded neurons. To test for such differences, we assessed decoding performance on a per-patient basis, and analyzed each participant's neuronal population to see if key decoding information is widely distributed or driven by a particular subset.

Decoding performance was obtained for three label categories—Summer, Scene Cuts, and Indoor/Outdoor—representing characters, visual transitions, and locations. To minimize computational load, we retrained a simpler logistic regression model on each patient's neurons, and achieved performance comparable to a more complex recurrent neural network but with lower computational costs. The results are illustrated in **Fig. 4**. Generally, decoding performance was lower at the individual patient level compared to the pooled neuronal population, with no single patient matching the performance of the aggregated data. For the Summer label, we observed substantial variability in decoding performance across patients, with some patients showing near-zero accuracy. However, certain patients (specifically 7,10,15, and 22) achieved decoding performances exceeding 0.2, compared to the overall pooled performance of 0.28. This variability was less pronounced for the Scene Cuts and Indoor/Outdoor labels. For the Scene Cuts label, the already low pooled performance declined further in the per-patient analysis, with patients 2,10,19, and 20 showing slightly better results, while most demonstrated minimal decoding accuracy. The Indoor/Outdoor label elicited a consistently higher accuracy across patients, matching the overall higher decoding performance achieved with the pseudo-brain population. Notably, patients 2,13,20 and 25 achieved performances exceeding 0.2 (Cohen's Kappa), compared to an overall pooled performance of 0.31, indicating robust neuronal responses in patient-specific subpopulations of neurons. Across all labels, the highest-performing patients vary, and no single patient showed consistently superior performance across all three.

Parahippocampal cortex drives decoding of visual transitions and location

Continuously presented stimuli offer a rich array of features. Visual transitions, such as changes in filming angle or scenery, are a commonly studied feature that demarcate the event structure of the dynamic stimulus. In movies, these transitions are relatively well-defined since they consist of identifiable changes in pixel values between frames and are known to elicit time-locked changes

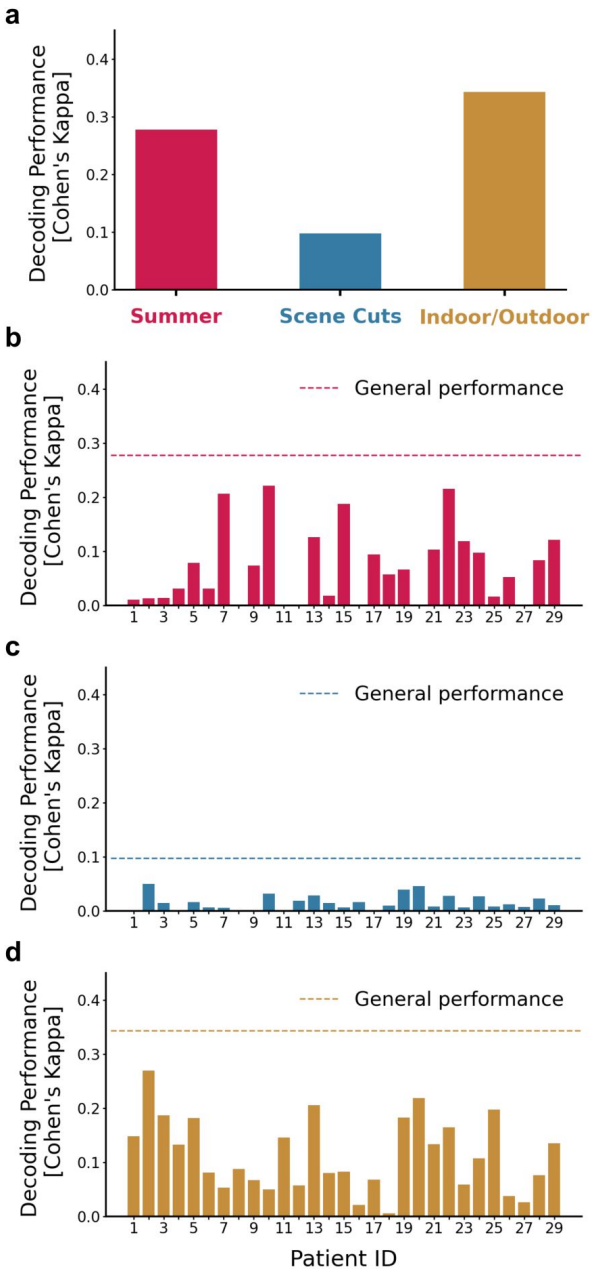

Figure 4.

Patient-specific decoding performance.

Decoding performances for the main character (Summer), visual transitions (Scene Cuts), and location (Indoor/Outdoor) are reported using Cohen's Kappa and compared to the performance obtained from the total population (pooled across all patients, dashed line). a) Decoding performance based on the total population of 2286 units, with neurons pooled across all patients. b-d) Patient-specific decoding performances for Summer, Scene Cuts and Indoor/Outdoor.

in neural activity in fMRI [20], iEEG [22], and single neurons [11]. We investigated two types of frame-wise visual transitions: Scene Cuts (changes in scenery) and Camera Cuts (changes in filming angle). As Scene Cuts consists of visual transitions between locations or points in time and demarcate narrative episodes within the movie, they are related to location but not exclusively. We compared this label to a more straightforward location-related feature, Indoor/Outdoor, which indicates whether a given frame depicts an indoor environment or not. Examples of Scene versus Camera Cuts as well as Indoor/Outdoor scenes are shown in Supp. Fig. S2.

To investigate region-wise differences, we trained separate decoders for neurons in the amygdala (A), hippocampus (H), entorhinal cortex (EC), and parahippocampal cortex (PHC) of the MTL. We excluded the piriform cortex (PIC) due to its relatively lower number of recorded neurons (**Fig. 1b**). We observed a clear dominance of the parahippocampal cortex for both types of visual transitions. The decoding performances reached 0.21 ± 0.02 for Scene Cuts and 0.19 ± 0.03 for Camera Cuts, respectively, when restricting the decoding to the parahippocampal cortex as opposed to 0.20 ± 0.03 and 0.18 ± 0.01 when decoding from the full population. The other regions showed a lower but above-chance decoding performance. Similarly, the parahippocampal cortex yielded the highest performance for predicting indoor versus outdoor content and reached a performance of 0.30 ± 0.06, comparable to the performance on the full population (0.31 ± 0.06). Hippocampus was the second strongest region with a high performance of 0.26 ± 0.04. For entorhinal cortex and amygdala, we observed lower performances of 0.15 ± 0.05 and 0.1 ± 0.03, respectively.

In summary, the parahippocampal cortex consistently achieved the highest decoding performance for labels associated with visual transitions and location, in line with prior research on the MTL [11]. Given that both Scene and Camera Cuts are linked to sharp visual transitions, we anticipated that earlier processing stages in the MTL would show better decoding performances than later processing stages. However, despite the clear dominance of the parahippocampal cortex, our results show that other regions achieve lower, but nonetheless significant performance when detecting event structure and setting information.

Amygdala drives decoding of character presence

The MTL carries information about the identity of specific individuals, in addition to general person-related categories or attribute, primarily through the tuning of individual neurons [31–33]. Unlike visual transitions, character identities are a semantic feature which rely on both visual attributes and higher-level abstract representations. To investigate character-driven representations at the population level, we analyzed neuronal activity during the presence of the movie's three main characters, Summer, Tom, and McKenzie, as well as the more general concept of any character appearance (Persons, see example frames in Supp. Fig. S2). While Summer's appearance throughout the movie frames is balanced (50/50), the remaining labels are highly imbalanced: Tom and Persons appear in the majority of frames (80/20 and 95/5), while McKenzie is predominantly absent (10/90) (**Fig. 1c**). Despite the imbalances, we observed significant decoding performances for all four character labels ranging between 0.23 and 0.36 (**Fig. 3c**). Notably, decoding performance for character identities—despite being abstract and variable—exceeded that of visual transitions (0.20 and 0.18).

Distribution of information across MTL regions

To investigate whether characters were primarily processed in a specific MTL region or in all regions equally, we conducted a similar analysis as before by retraining on region-specific activity (**Fig. 5a**). Our results show that all four tested regions carry information about the character's identities, enabling decoding at above-chance levels ($p < 0.001$, permutation test, see Methods). The amygdala and parahippocampal cortex showed the highest decoding performances for Summer and Tom, respectively, approaching levels similar to decoding from the full population. However, the distribution of information among the other regions was less consistent and varied across

labels (**Fig. 5a**). The hippocampus had the lowest performance for Summer (0.09 ± 0.04) and Tom (0.05 ± 0.05), while the entorhinal cortex performed lowest for McKenzie (0.12 ± 0.02). For Persons, the parahippocampal cortex dominated with decoding performance comparable to the full population (0.36 ± 0.05), while other regions ranged between 0.14 and 0.21.

Differences in character's visual appearances

Since the labels indicate the presence of a given feature within a natural scene rather than a single exemplar shown in isolation, the visual appearance of the labeled entity varied substantially during the movie.

To better control for visual appearance and examine decoding differences across various levels of character presence, we created the additional labels Summer Faces, Tom Faces, and Summer Presence (see Methods for details on annotation creation). As with the character labels, we trained a decoding network on both the full neuronal population as well as the four individual regions (**Fig. 5b**). Face labels for both characters elicited a slight improvement in performance compared to the full neuronal populations, with Summer increasing from 0.31 ± 0.06 to 0.33 ± 0.06 and Tom increasing from 0.23 ± 0.06 to 0.27 ± 0.07. Regionally, the distribution remained consistent with the general character labels, except for Tom Faces, where the entorhinal cortex had the highest performance. The hippocampus performed weakest for both Summer Faces (0.08 ± 0.04) and Tom Faces (0.11 ± 0.03). The Summer Presence label had an overall performance of 0.27 ± 0.04, slightly lower than the general Summer label, with the amygdala clearly dominating the regional distribution (0.26 ± 0.07). For face labels, distributions were more similar across regions than for the general character labels. The amygdala outperformed other regions in decoding the abstract presence of Summer but performed poorly for general person appearances, where the parahippocampal cortex performed best. These findings align with previous research showing that semantically-tuned cells in the human MTL can flexibly activate when their preferred conceptual category is indirectly invoked [34].

Responsive neurons drive decoding of visual transitions but not decoding of characters

Although only a subset of neurons modulated firing in response to the onset of a given movie feature, we nevertheless observed significant decoding performance from the full population. This effect could result from two scenarios: a) information is distributed throughout the neuronal population and decoding does not disproportionately rely on neurons with post-onset increases in firing, or b) the subpopulation of responsive neurons informs the decoder while non-responsive neurons are ignored. We tested each scenario by dividing the full population into two corresponding subsets—non-responsive and responsive neurons—and re-training a neural network on each subset. We analyzed the labels Summer and Camera Cuts, which represent the character-related and visual transition categories, respectively. We then compared the prediction performance of each re-trained model to that of the full population to determine if the decoding of a given label took the entire population into account (scenario a) or relied on responsive neurons (scenario b) (**Fig. 6**). The subsequent analysis evaluates these subsets both at the full population level and within the restricted context of MTL regions.

Decoding with non-responsive versus responsive neurons

First, we tested prediction performance using only the non-responsive neurons to determine if similar decoding could be obtained without neurons which significantly modulated firing after the onset of a feature. For this subpopulation (*Non-responsive (only)*), a separate LSTM was retrained and tested. Despite the exclusion of responsive neurons, these subpopulations yielded performances comparable to those of the complete population (**Fig. 6a**, *Complete*) for the character Summer. Minimal differences in performance were observed across MTL regions, with

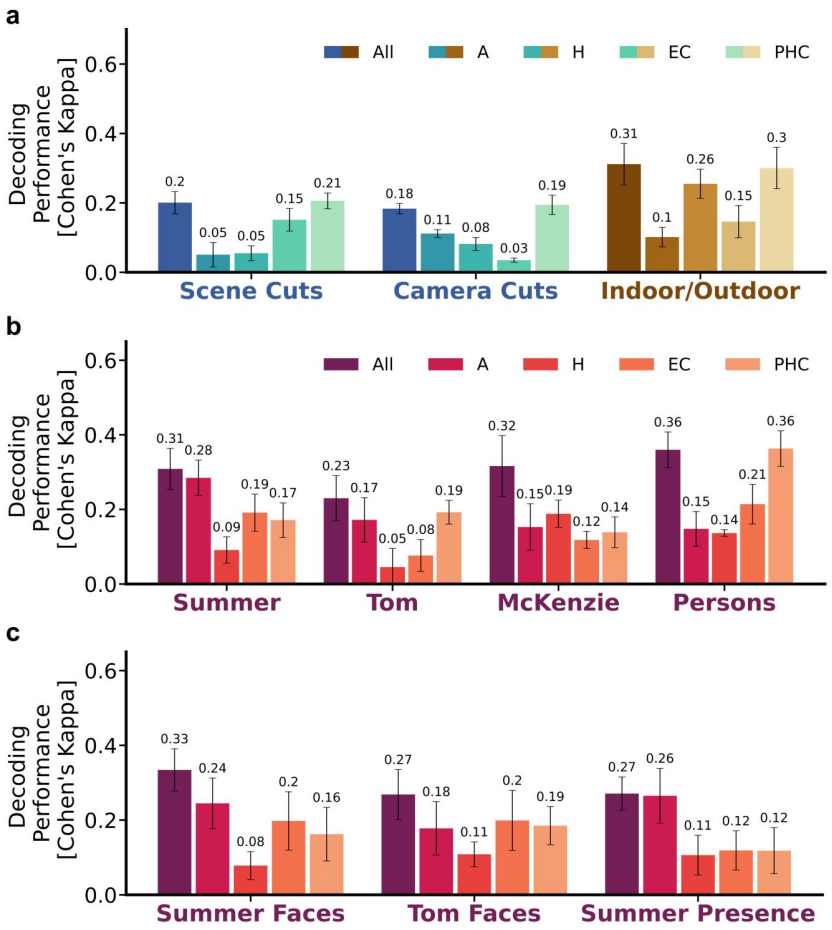

Figure 5.

Semantic information is distributed differently across MTL regions based on category.

Decoding performances for semantic features, by MTL region. All performances were significantly better than chance level with an alpha level of 0.001 (reported performance using Cohen's Kappa, mean performance across five different data splits, error bars indicate standard error of the mean. a) Decoding performances for visual transitions and location features. b) Character visibility could be decoded from the entire population of neurons, and with variable performance when training only on individual MTL regions. c) Decoding performances for face-specific character appearances and Presence features, by region.

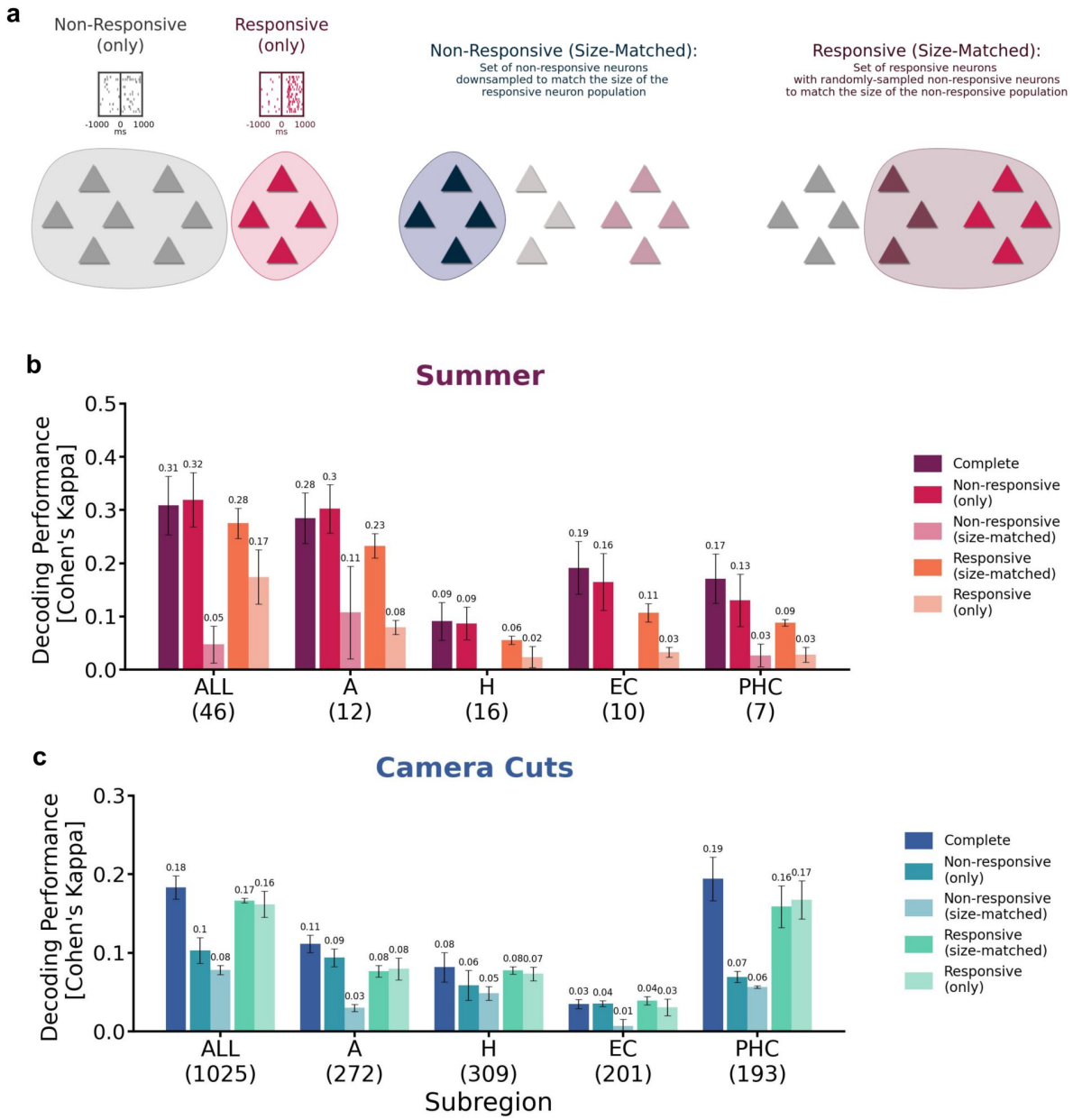

Figure 6.

Responsive neurons drive performance for visual transitions, but not characters.

To assess the contribution of the responsive neurons on decoding (identified in *Stimulus-aligned responsive neurons found primarily in parahippocampal cortex*), we compared the decoding performance for subpopulations which did or did not contain these neurons. a) Illustration of neuronal sets used in the decoding comparisons (triangles represent neurons). An example of the complete population is shown in the left-most section, which depicts *Non-responsive (only)* and *Responsive (only)* cells, with a simulated example of a respective peri-stimulus time histogram (onset raster, grey and magenta). A *Non-responsive (size-matched)* group (middle section) was randomly subsampled from the total population to have a size-matched comparison to the total set of responsive neurons. The *Responsive (size-matched)* set (right-most section) consisted of all responsive neurons padded with randomly selected non-responsive neurons to match the total *Non-responsive (only)* population. b,c) Decoding performances for discussed subpopulations for the character label Summer and Camera Cuts. Number of responsive neurons for the respective subpopulation reported in parentheses.

only the entorhinal and parahippocampal cortex showing qualitative decreases. For comparison, we additionally retrained using only responsive neurons as input, and again tested the decoding performance (*Responsive (only)*). In contrast to the minimal differences observed in the *Non-responsive (only)* model, restricting to only responsive neurons produced a decrease in overall performance across the entire MTL and all regions. This general decrease suggests that the subset of individually responsive neurons is not the primary driver of the decoding performance observed in the full population model. Note that the total number of responsive neurons was less than the total non-responsive, so this effect may be influenced by an overall decrease in neuronal data. This difference in totals is directly addressed below by a size-matching procedure.

A different pattern emerged for the Camera Cuts feature, which exhibited the greatest performance drop when responsive neurons were excluded, for the entire MTL and for all subregions but the entorhinal cortex (**Fig. 6b**). This pattern was most pronounced for the parahippocampal cortex, indicating that responsive neurons carry valuable information for processing Camera Cuts in the movie. This hypothesis is further supported by the decoding performances obtained when restricting the decoding to only responsive neurons. Despite the restricted number of input neurons, the performance dropped marginally, and remained comparable to that of the complete population for the entire MTL, as well as for the amygdala, hippocampus, and entorhinal cortex regions.

Size-matched neuronal populations: comparing responsive and non-responsive subsets

Since the total number of responsive neurons was lower than that of non-responsive neurons, we tested the performance using size-matched versions of both non-responsive and responsive subpopulations. To match the size of the responsive neurons, we randomly selected an equivalent number of non-responsive neurons (*Non-responsive (size-matched)*) and trained and tested a separate neural network. This process was repeated three times with different random selections, and we report the average performance. For both labels, Summer and Camera Cuts, the smaller size-matched non-responsive subpopulation showed an expected decrease in performance compared to the full non-responsive subpopulation. However, the results diverged when comparing the size-matched populations: For Summer, the size-matched non-responsive neurons performed comparably to the responsive neurons within individual MTL regions. Only for the complete population did the responsive neurons show a clear improvement. Conversely, for the Camera Cuts, restricting to only the responsive neurons improved performances, with decoding predictions of all but the amygdala surpassing those of the size-matched non-responsive set of neurons.

A similar pattern emerged for a size-matched version of the responsive neurons (*Responsive (size-matched)*), which we formed within region by padding the set of responsive neurons with randomly selected non-responsive neurons. For Summer, decoding from this subpopulation consistently showed a performance drop compared to the total set of non-responsive neurons across all tested regions. In contrast, for Camera Cuts, the size-matched subpopulation of responsive neurons achieved similar or better performance than the non-responsive neurons across all regions, with the complete population and parahippocampal region showing a clear dominance of the subpopulation containing responsive neurons. We additionally evaluated performances for the labels Tom, Scene Cuts and Indoor/Outdoor, which matched the effects found for Summer and Camera Cuts (Supp. Fig. S10).

In summary, our findings indicate that responsive neurons play distinct roles for different features. Neurons responsive to visual transitions appeared to carry information not equally present in other neurons. On the contrary, for character- and location-related labels, individual

responsive neurons contributed less to decoding, and information appeared to be distributed either across the entire population or a subset of neurons distinct from the previously identified responsive neurons.

Relevant information is carried by a smaller subpopulation of 500 neurons

We observed that subsets of neurons with stimulus-selective responses to characters did not account for the decoding performance of the same character. Previous research in sensory information processing suggests that relevant stimulus information is often encoded by only a subset of neurons within a population [27,35,36]. To explore this further, we adopted a more data-driven approach to define neuronal subsets, ranking their importance using weights extracted from a trained logistic regression model and investigated the minimally sufficient number of neurons required for successful decoding.

Ranking of neurons for the character label Summer

The weights of a logistic regression model can be used to assess each neuron's importance in decoding, allowing for the creation of a ranking across all neurons (see Methods for more details). In contrast to an LSTM, logistic regression models are computationally less expensive to train and achieve comparable decoding results for all movie features, except for Scene Cuts and Camera Cuts, which exhibited reduced but above-chance decoding performance (**Fig. 3c**). For the character label "Summer," we generated a neuron ranking from the trained logistic regression model and defined subsets of neurons by selecting those with the highest rankings.

The above LSTM and logistic regression models were trained on population data using 5-fold cross-validation, with alternating test sets for each of the five splits. However, this approach precludes an independent ranking of neurons across splits, as test data from one split may overlap with training data from another. To address this, we expanded the five original splits to 20, where each subgroup of four splits shared a common test set but varied in the allocation of training and validation data. Any analysis relying on a subset of neurons derived from logistic regression weights was exclusively assessed using the four splits that produced the ranking and shared held-out test data (see more details in Methods, and Supp. Fig. S8).

Training on pre-selections of top-ranked neurons

We trained on progressively smaller subpopulations of top-ranked neurons for the character Summer (**Fig. 7a**) and observed an increase in performance when restricting the input activity to smaller populations (peak at 0.38 (Cohen's Kappa) for 500 neurons, 21.9% of the total population). The decoding performance reported here for the entire population shows a slight variation from the previously reported value of 0.31 due to the modified nested cross-validation procedure. Further reduction of the population lead to a decrease in performance, yet high decoding performance persisted even in small subpopulations of neurons.

We observed that a ranking procedure which did not use a common test set was subject to potential cross-talk between data splits, which substantially impacted and distorted the results. Using a selection of neurons derived from non-independent training and test data led to a stark increase in performance, nearly doubling the original performance of the character Summer to 0.52 (Cohen's Kappa) when restricting to a subpopulation of 150 top-ranked neurons. This again underscores the need to carefully prepare the data for paradigms such as ours, where data samples are highly correlated, as ignoring dependencies between training and test data can greatly skew results (see Supp. Fig. S13).

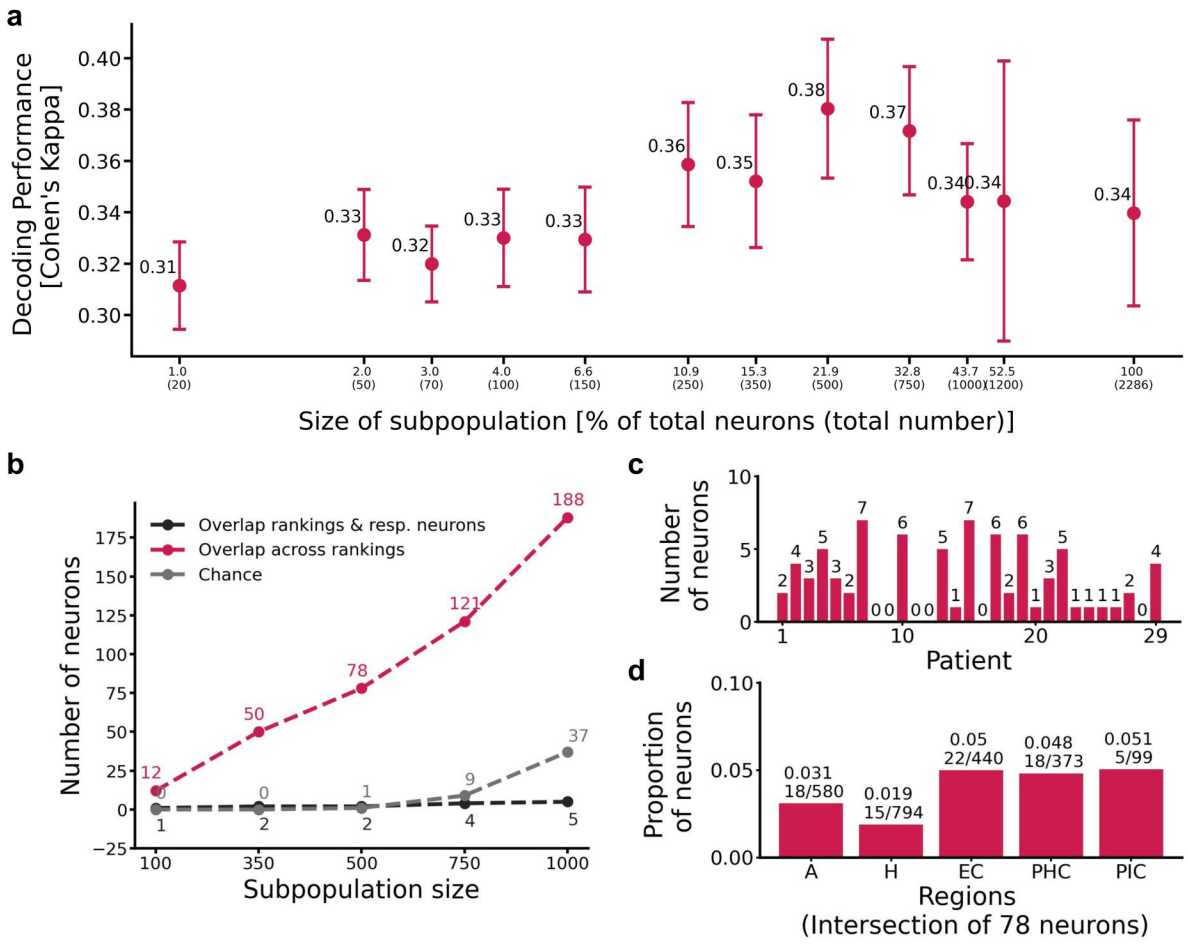

Figure 7.

500 neurons are sufficient to reach peak decoding performance.

a) Decoding performances for the character Summer for subpopulations of top-performing neurons, testing various sizes ranging from 1% to 100% of the full population (absolute numbers of neurons are reported in parenthesis). Mean performance across different splits is reported, and standard error of the mean is visualized by the error bars. b) Number of overlapping neurons across rankings for different sizes of subpopulations of top-performing neurons (pink). As a baseline, we compare the number of overlapping neurons to the number expected by chance (grey), and we observed a notably higher intersection of top-ranked neurons across the splits. Additionally, the overlap between the intersection of top-ranked neurons and the previously defined responsive neurons is shown (black). c, d) Overlapping neurons (in total 78) in subpopulations of 500 top-performing neurons for each ranking were distributed across patients and MTL regions.

Top-ranked neurons and their distribution across patients and MTL regions

As the selection process involved five distinct rankings of neurons, we investigated the consistency of the neuronal composition across rankings. The overlap of neurons within subpopulations of top-ranked neurons assessed across different sizes is shown in **Fig. 7b**. Analyzing the top-performing 500 neurons from each of the five rankings revealed a common set of 78 neurons. We compared this observed overlap to that expected by chance with random subpopulation selections (see Methods), finding a notably higher overlap. This suggests the presence of common neurons crucial for decoding the label Summer. Additionally, we also compared these selected neurons to those previously classified as responsive. The overlap between the two groups of neurons was small, suggesting that the responsive neurons defined by onset-related changes in activity may not be the primary drivers of decoding performance in general, aligning with previous observations for character labels when the decoding pipeline was restricted to responsive neurons. As restricting to 500 neurons yielded the highest decoding performance, we subsequently analyzed the resulting intersection of 78 neurons across rankings. These 78 neurons were distributed across both patients and regions (**Fig. 7c,d**), with no single patient or area of the MTL driving the performance. Visualizations of the spiking activity surrounding the onset of Summer for these neurons do not reveal a clear pattern of stimulus-evoked increases in firing (Supp. Fig. S12).

Our findings reveal that a core subpopulation of approximately 500 neurons drives decoding performance, while additional neurons mainly contribute redundant or noisy information. These neurons are distributed across patients and MTL regions, presenting an important avenue for future research to investigate the mechanisms underlying their organization and the specific functional roles they play in decoding processes.

Discussion

We investigated how the human brain processes semantic and event structure in a naturalistic setting by analyzing the activity of neurons in the medial temporal lobe during the presentation of a full-length commercial movie. Although earlier work has established and characterized the role of MTL neurons in semantic representation, and the processing of dynamic stimuli via fMRI and intracranial electroencephalography, few studies have investigated how human single neurons process dynamic stimuli and none have addressed the relationship between representations on the single-neuron level and the population level.

By analyzing changes in each neuron's activity aligned with the onset of labeled movie features, we identified groups of individual cells that adjust their firing rates in response to specific features. The most pronounced responses occurred following changes in camera angles and scenes, as these features induced activity changes in the largest number of cells. Outdoor scenes and the two main characters also elicited consistent single-unit responses, albeit in far fewer neurons. This lack of explicit single-neuron responses to characters, which might otherwise suggest selective and invariant representations such as those in concept neurons, could be explained by the study participants' lack of familiarity with the movie. Most participants had neither seen the movie before nor encountered the actors in other media prior to its presentation as part of this study. This interpretation aligns with previous research showing that neuronal selectivity to individuals varies with familiarity and personal relevance, as photos depicting personally known individuals are more likely to elicit selective responses in MTL neurons [37].

We anticipated that the decoding performance for each label would reflect the pattern observed at the single-neuron level: Visual transitions would be the most accurately predicted feature, followed by setting, with character presence showing the lowest prediction accuracy. Although not

every feature elicited explicit responses from a significant portion of individual neurons, the network, which takes the collective population activity as input, successfully decoded all tested features with above-chance performance. Interestingly, the decoding performance varied across features and contradicted the pattern found in the individual neurons. Despite eliciting the highest proportion of responses at the single-unit level, visual transitions showed the lowest decoding accuracy out of all tested features. Conversely, characters showed the highest decoding accuracy despite there being few individually responsive neurons. Both approaches link neuronal responses to the movie content but differ markedly in their focus, as the single-unit analysis targets the specific onset of features, emphasizing their initial activation, while the population approach processes data continuously, decoding both the onset and sustained presence of features. This methodological difference may affect the results and contribute to the observed contradiction between them. To bridge these findings, we examined how individual responsive neurons contribute to the decoding network. We hypothesized that the decoding performance for each label would be primarily driven by the subset of neurons that exhibit increased activity in response to that label. This hypothesis held true for the decoding of visual transitions, as the set of individually responsive neurons disproportionately contributed to the decoding performance when using the full population, and performance was strongly affected by their removal from the training population. In contrast, character decoding does not rely on individually responsive cells, as removing neurons that responded strongly to character onset had little impact on performance. When analyzing the network and its prediction behavior, we identified a subset of units which contributed most to the decoding of character information across models, and found that these units had little overlap with the set of neurons which increased firing after character onset. Together, these findings suggest that character-related representations relied on a population code, while visual transitions were encoded by the activity of specific neurons.

When training separately on different regions of the MTL, we observed variations in decoding performance depending on the specific content being decoded. The parahippocampal cortex achieved the highest performance in detecting visual transitions, while the amygdala performed best in predicting character-related information. These results support previous findings which identified that certain regions are more likely to respond to certain categories of images in static screenings. For example, previous studies have shown that the amygdala preferentially responds to images containing faces [1,32], and contains cells that selectively respond to whole faces as opposed to discrete facial features [38]. We extended this analysis to examine different levels of character appearance, distinguishing between face visibility and the general presence of a character in a scene (regardless of face visibility). Although, the amygdala demonstrated strong performance in both cases, it most clearly drove the decoding for the general presence of a character, rather than specifically to face visibility. This effect could be due again to the unfamiliarity of the movie and its actors, as face-specific responses have been shown to form as a function of exposure [39].

Previous work has found that the parahippocampal cortex is especially sensitive to scene information [40], as opposed to objects [40,41], with a higher likelihood of neuronal activation when scenes feature stronger spatial layout cues, such as depth and a recognizable background [42]. Zheng et al. (2022) identified generalized cells, termed “boundary cells”, in the parahippocampal gyrus, hippocampus, and amygdala which modulate their firing after any visual transition event. In our study, we observed significant responses to camera cuts in these same regions, which we interpret as analogous to “soft boundaries” [11]. However, significant responses to scene cuts were only observed in the parahippocampal cortex, whereas Zheng et al. reported responses to their analogous feature (“hard boundaries”) in all measured MTL regions. In addition to their preference for scene-related images, parahippocampal neurons have been shown to respond more frequently to outdoor images compared to indoor ones [42]. In our dataset, a small subset of parahippocampal neurons increased firing in response to the onset of outdoor scenes, whereas none showed increased activity for indoor scene onset. Despite their limited number, these responsive neurons achieved well above-chance decoding performance, albeit falling short

of the performance achieved by the complete neuronal population. Further work is needed to more accurately determine if the activity of single scene-selective parahippocampal neurons in our dataset can be explained by the onset of location-related content.

Visual transitions and character content additionally differed in their use of sequence information. Through a comparison between decoding architectures, we found that temporal information in the spike trains only mattered for the prediction of visual transitions, and that ignoring or even scrambling the sequence information had little effect on character and location features. Although using temporal dynamics improved decoding performance for visual transitions, the temporal dynamics in our data, particularly those inherent to the movie stimulus, also present a significant confounding factor. Both the movie and recorded neuronal activity are subject to a high degree of correlation in time and thus require extra consideration when formulating a pipeline to ensure that the training and test data did not contain adjacent, highly correlated samples. The correlation between training and test samples artificially inflates decoding performance, and does not reflect actual generalization to unseen data. In related research by Zhang et al. (2023), where temporal distance was not considered, high decoding performances were reported on a similar task. Based on our analyses, we anticipate that their reported decoding accuracy is overestimated, and would change if sufficiently large gaps were introduced.

As our dataset consists of single-neuron activity pooled across 29 subjects, each with activity from an average of approximately 80 recorded neurons, the extent to which claims can be made about what an individual brain does is limited. In addition, the participants watched a full-length movie in an clinical setting, where neuronal activity is not solely focused on visual stimuli and likely processes additional information. Although we cannot know whether the neuronal population that we sampled is representative of the human MTL generally, it is one of the largest samples collected to date, both in terms of neurons and patients and offers a unique opportunity to understand the processing and representation of information in the MTL population activity. Despite the inherent limitations, which preclude near-perfect performances in the decoding task, the presented approach demonstrates that movie content can nevertheless be successfully decoded from such a sub-sampled population of neurons. Future work on this dataset could leverage more advanced network architectures to explicitly model between-neuron dynamics within each patient, which could better explain the gain in performance achieved moving from the single-neuron level to the population level. Additionally, in this work we focused specifically on the *visual* content of the movie. A clear next direction would be to integrate the auditory information of the movie, and possibly disentangle the contribution of visual versus audio information streams to the neuronal representation of movie features in the MTL.

Materials and Methods

Participants and recording

The study was approved by the Medical Institutional Review Board of the University of Bonn (accession number 095/10 for single-unit recordings in humans in general and 243/11 for the current paradigm) and adhered to the guidelines of the Declaration of Helsinki. Each patient gave informed written consent for the implantation of micro-wires and for participation in the experiment.

We recorded from 46 patients with pharmacologically intractable epilepsy (ages 19 - 62, median age 37; 26 female patients). Patients were implanted with depth electrodes [31] for locating the seizure onset zone for potential later resection. Micro-wire electrodes (AdTech, Racine, WI) were implanted inside the shaft of each depth electrode. Signal from the micro-wires was amplified using a Neuralynx ATLAS system (Bozeman, MT), filtered between 0.1 Hz and 9000 Hz, and

sampled at a rate of 32 kHz. Spike sorting was performed via Combinato [43] using the default parameters and the removal of recording artifacts such as duplicated spikes and signal interference was performed via the Duplicate Event Removal package [44]. After all data were processed, neuronal signals, experimental variables, and movie annotations were uploaded to a tailored version of Epiphyte [45] for analysis. Due to disruptions in the movie playback caused by clinical interruptions, 13 patient sessions were excluded from further analysis.

Task and stimuli

Patients were shown a German dubbing of the commercial movie *500 Days of Summer* (2009) in its entirety (83 minutes). This film was chosen because the actors portraying the main characters were relatively unfamiliar to a general German audience at the time of the initial recordings. The movie was shown in an uncontrolled clinical setting, where neither gaze nor attention were directly monitored and was presented in letterbox format without subtitles on a laptop using a modified version of the open-source Linux package, *FFmpeg* [46], with a frame rate of 25 frames per second. Due to the length of the movie and the possibility of clinical interruptions, patients and staff were allowed to freely pause the playback. Discontinuity in playback was controlled for within Epiphyte [45]. Pauses and skips in the movie playback were identified through the output of the modified *FFmpeg* program and used as a basis of exclusion for patients. Patients were excluded if they did not watch the entire movie, or watched the movie discontinuously.

Movie annotations

In order to relate the content of the movie to the recorded neuronal activity, we labeled various features on a frame-by-frame basis. These labels are binary and cover the following features:

- ***Main characters:*** *Summer, Tom, McKenzie*
  Frames were labeled as positive if the character could be clearly distinguished by either appearance or context. Characters and persons were considered only on a visual basis (i.e., a frame in which Tom is speaking but not visible is labeled as not containing Tom).
- ***Faces:*** *Summer, Tom*
  Instances of a character's face. Positive samples are frames where the character's face is shown, while negative samples are frames where the character's face is not visible at all. All other frames are excluded.
- ***Presence:*** *Summer*
  Indicates the character's general presence in the scene, even if the character is not visible in the frame. For instance, frames are labeled as positive if the character is part of the scene but is not visible in that particular frame due to factors like the camera angle.
- ***Visual transitions:*** *Camera Cuts, Scene Cuts*
  Marks visual transitions in the movie. Scene Cuts correspond to changes in scenery, while Camera Cuts are primarily based on changes in the visual stimuli.
- ***Location:*** *Inside/Outside*
  Distinguishes between indoor and outdoor locations in the movie. Scenes that do not clearly fit into either category are excluded from the annotation.
- ***Persons***
  General appearance of any person(s)

Main character, Presence, Persons, Location, and Scene Cut labels were obtained manually using the open-source annotation program *Advene* [47]. For face labels of the characters, we developed a deep-learning pipeline for face detection and classification. As backbone we used a pre-trained neural network for face detection and feature extraction [48]. We extended the pipeline by a classification network consisting of fully-connected layers combined with ReLU activation functions. The classification network was fine-tuned on the movie frames to classify detected faces of the main characters, including a "not known" class for faces not belonging to the main characters. In the fine-tuning process, we used the manually created character labels for the

characters. Camera Cuts were labeled automatically using the open-source algorithm *PySceneDetect*[49], run with default parameters and manually reviewed. Camera Cuts mark a cut in the movie by labeling the first frame after cut onset as positive, resulting in cut events associated with a single frame. To adjust for the temporal latency in brain activity, cuts in the movie were associated with frames occurring within 520 ms of the cut onset. This adjustment smooths the cut labels, rendering them more comparable to what we anticipate in neuronal responses.

Calculation of single-neuron response statistics

The "baseline" period was defined as 1000 ms prior to the onset (e.g., the entry of a character into frame) and the "stimulus" period was 1000 ms after the onset or appearance. Pseudo-trials with baseline periods containing frames depicting the label of interest were excluded. Responsive neurons were identified using a modified bin-wise signed-rank test [8]. The spiking activity across pseudo-trials was aligned to the stimulus onset. The baseline period was binned by 100 ms, and the normalized firing rate of the baseline period was compared to nineteen overlapping 100 ms bins defined across the stimulus period using a Wilcoxon signed-rank test (alpha = 0.01, SciPy wilcoxon [50], Simes corrected [51]). Additionally, a neuron was required to have spiked during at least one third of the total pseudo-trials to be tested (otherwise, assigned $p = 1$).

Cluster permutation test

A cluster permutation test [26] was used to test the difference in firing rates between the responsive and non-responsive subsets of neurons. Using the calculated response statistics, neurons were divided into two conditions: responsive ($p \uparrow 0.01$) and non-responsive ($p > 0.01$) for a given label. Activity for a neuron was averaged across bins, yielding a single vector of mean spike counts (spikes / 100 ms) spanning both baseline and stimulus periods for each neuron. This vector was then z-scored relative to its own mean and standard deviation. Mean spike count vectors were combined across conditions, yielding two datasets: A_{resp} and A_{non}, matrices containing the summarized, binned activity for all responsive and non-responsive neurons, respectively. A bin-wise comparison between A_{resp} and A_{non} was performed using a two-sided t-test for independent samples (ttest_ind, SciPy), producing a t-stat and p-value for each bin. Clusters were defined as temporally adjacent bins with $p \uparrow 0.005$ and t-stats were summed within clusters. The procedure was adapted to allow testing of multiple clusters, so no clusters were excluded at this stage. One set of 1000 permutations were performed by randomly assigning neurons to A_{resp} and $_{Anon}$ such that the total number of neurons in each group was conserved and the bin-wise testing procedure was performed on each permuted dataset. Each cluster was compared to the resulting histogram of permuted t-stats. P-values for each cluster were defined as the number of permutations with a higher summed t-stat relative to the total number of permutations, and a p-value less than 0.05 was considered significant.

Decoding from population responses

For decoding, we used population responses (input) to predict the corresponding concept labels (output). We excluded the credits from the dataset and focused solely on the narrative content. The movie was presented at a frame rate of 25 frames per second, with each frame lasting 40 ms. In total, 125, 743 frames were shown. The spiking activity of the recorded neurons was binned using a bin size of 80 ms, corresponding to the activity of two consecutive frames. Each bin was then labeled based on the first frame within that interval.

We sampled activity of the neurons before and after the onset of each frame and found that using a window of 800 ms before and 800 ms after onset yielded the best decoding performance. This resulted in data samples comprising a binary label for the concept and a spike train of length 20 (10 bins before and after the onset of the frame). Given the full population of 2286 recorded neurons, each input data sample had a dimensionality of 2286.

Architecture

To decode concepts from the sequence of neuronal activity, we used a long short-term memory (LSTM) network [30], which is well-suited to process the dynamical structure of the dataset. The output of the LSTM was fed into into several fully-connected layers with ReLU activation functions to obtain binary label predictions. We found that pre-processing the raw spiking data with a linear layer of same size as the input, combined with a batch normalization layer, improved performance. We used the binary cross-entropy loss function, and optimized our network using Adam optimizer [52] in Pytorch [53] with default settings (first and second order moment equal to 0.9 and 0.999, respectively). We obtained the best results with a 2-layer LSTM, with hidden size of 32. We adapted other hyperparameters, e.g., number of linear layers, hidden sizes, batch size, learning rate, weight decay, and dropout rate, for each label.

We trained each network for 700 epochs and used the validation set to estimate the model's ability to generalize to unseen data. As is common practice, we selected the model with the best performance on the validation set and used this for evaluation on the unseen test data. Some labels were highly imbalanced and models trained on these labels were biased towards predicting the majority class. To ensure unbiased predictions and to optimize performance, it is common to force the batches of data samples presented in the training process to be balanced by oversampling the minority class. This oversampling technique was applied to all imbalanced labels, comprising all labels except the label for the main character Summer.

Data split

To ensure that the decoding performance was not affected by correlations between data samples, we carefully split the dataset into training, validation, and test sets. We used 5-fold nested cross-validation, and assigned 70% of the data to training, 15% to validation, and 15% to testing in each split. To avoid correlations between samples, we assigned samples from consecutive segments of the movie to each set (train/val/test) and excluded 32 seconds of data between each set (see Fig. S7). The choice of excluding 32 s was based on the results for the main character Summer (**Fig. 3c**). The total dataset contained 45, 200 samples, resulting in sets of 30, 800/7, 200/7, 200 samples for training/validation/test. We trained a network using the training data, optimized its hyperparameters using the validation set, and evaluated its performance on the corresponding test set. We report the final decoding performance as the average performance on all five test sets.

Evaluation metrics

For each semantic feature, we compared the model's prediction against the binary class labels. While accuracy is a simple evaluation metric, it is not suitable in our case due to the highly imbalanced distribution of most labels, making it challenging to compare accuracy metrics across different labels. We report all decoding performances using the Cohen's Kappa metric [54] which measures the agreement between the ground-truth labels and the predictions of the network, where performance equal to zero is interpreted as chance-level and performance equal to one is interpreted as complete agreement. Cohen's Kappa is defined as

$$\kappa = \frac{p_0 - p_e}{1 - p_e} \tag{1}$$

where p_0 defines the relative observed agreement and p_e the hypothetical probability of chance agreement among prediction and labels. The Cohen's Kappa metric can take negative values, too, implying that the predictions are worse than chance level. The common chance performance equal to zero makes the Cohen's Kappa a useful metric to compare performances across all labels. Additionally, we report the F1 Score, Area under the Precision Recall Curve (PR-AUC) and Area

Under the Receiver Operating Characteristic (AUROC) metric in the Supplementary for all experiments. We briefly explain these metrics, including potential advantages and disadvantages for our analysis:

F1 Score

F1 Score is a metric combining precision and recall by calculating the harmonic mean between these two. Precision and recall are determined based on a classification threshold of 0.5. Designed to perform well on imbalanced datasets, the F1 Score is particularly useful for evaluating our decoding tasks. However, the baseline for chance performance with this metric is not consistent and varies depending on the label distribution.

PR-AUC

The PR-AUC metric, representing the area under the precision-recall curve, extends the F1 Score by evaluating performance across different threshold settings. Similarly to the F1 Score, it can be used for imbalanced datasets. However, as for the F1 Score, the PR-AUC metric is sensitive to changes in the class distributions, resulting in varying chance performance across labels. Being a reliable performance metric in general for our various classification problems, one has to bear in mind that a comparison of performances across concepts can be misleading due to the different chance baselines.

AUROC

The Receiver Operating Characteristics (ROC) curve represents the trade-off between the true positive rate and false positive rate for various threshold settings. AUROC as the area under the ROC curve is a performance measure that is used in settings where one equally cares about positive and negative classes. Performances range in [0, 1] and it is insensitive to changes in the class distribution, which means chance performance is given by a value of 0.5. However, the metric is generally not used for highly imbalanced classification problems and an evaluation of specific labels in our analysis such as McKenzie (class distribution is 90/10) should be taken with caution.

Single-neuron decoding performance

For a fair assessment of decoding from population responses compared to single-neuron activity, we evaluated the decoding performance of individual neurons under a setup comparable to that of the LSTM-based decoding network. Instead of relying on full-population responses, we established a performance based solely on the firing rates of each single neuron. Employing the same data splits, cross-validation approach, and binning procedure used for the LSTM network, we summed the binned firing rates of a neuron surrounding frame onset for the 1600 ms time window utilized by the decoding network. We then individually selected the best threshold for each neuron's activity on the validation set, which was subsequently used to evaluate the neuron's performance on the hold-out test set. We reported the final performance as the average of the five results obtained from the 5-fold cross-validation procedure.

Permutation tests for decoding

To determine if the reported decoding performances were significantly better than chance, we performed two sets of permutation tests—first, we randomly shuned the labels of the held-out test set (Test Set Shune), and second, we shifted the labels while preserving the order of the test labels (Circular Shift). For both tests, the input to the decoding network remained unchanged from the non-permuted version. The only modifications made were to the corresponding feature label in the test set, which were changed in the ways outlined above, and then compared to the original network's prediction scores. The dynamic nature of the visual stimuli implies not only a strong correlation within the neuronal activity but also suggests a temporal correlation for the feature

labels. By modifying only the labels, we ensured that the temporal information embedded in the neuronal data remained unaffected by the permutation test. In the following, we test the significance of both scenarios: one where the temporal correlation within the labels is disrupted (Test Set Shune), and another where the temporal correlation within the labels is maintained (Circular Shift).

Test set shune

The first type of permutation, and the one reported in the main text, consisted of randomly shuning labels in the test set and evaluating the predictions of the models on those. We compared the performance of the model on the heldout test set to a null distribution generated by evaluating the model on the test set with shuned labels (N = 1000). We calculated the probability of our observed performance under the null distribution to obtain the p-value:

$$p = \frac{k+1}{N+1}$$

where k is the number of performances on permuted data outperforming the ground-truth performance of the model. This p-value provided the basis of comparison for describing significant decoding results in our main analyses. By preserving the temporal structure during the model's inference process and only disrupting temporal correlations within the concept labels used for evaluating the model's predictions, we consider this assessment of significance to be the most appropriate for our data setup.

Circular shift

The second type of permutation was performed as a comparison for the above *Test Set Shufle*, as it is a standard method in human single-neuron studies (see [55], [56], [44]). Circular shift permutations maintain both the temporal relationship of neuronal data as well as that of the stimulus information, here the concept labels. Rather than randomly shuning the labels in the test set, we applied a circular shift of labels that maintained the intrinsic temporal structure. The shifting size was randomly chosen N = 1000 times to obtain a null distribution akin to the previous test. This distribution was then utilized to compute the probability of our ground-truth performance and derive the p-value for assessing significance. For studies involving static stimulus presentation, wherein the stimulus is largely uncorrelated with itself, this test provides a useful way to disentangle stimulus-related effects from those endemic to the time-series information. For a comparison between the permutation results, see Supplementary Materials, Fig. S11.

Impact of temporal information on the decoding

To assess the significance of temporal information in neuronal activity sequences, we evaluated the trained models using temporally-altered test data. The input to the decoding network included neuronal activity from 800 ms before to 800 ms after the onset of the frame, divided into 20 bins. For the temporally-altered test data, we randomly shuned the sequence order of the 20 bins (applying a consistent permutation for all neurons and data samples) and evaluated the pre-trained models on the modified test data. This procedure was repeated 100 times, and the performance was averaged. The final performance was calculated by averaging the results across the five data splits, using the standard error of the mean (SEM) as the measure of variance.

Logistic regression models and evaluation of neuron contribution to decoding

We compared the decoding performance of the LSTM to that of a logistic regression model. The dataset used to train the logistic regression model was identical, barring one key change: the spike trains provided to the LSTM, initially of dimension (20, 2286), were reduced to a single bin

representing the total number of spikes around a frame. The data then had a revised shape of (1, 2286), and no longer incorporated temporal information. We trained a logistic regression model using the liblinear solver implementation in Scikit-learn [57]. For training, we employed an L1 penalty and z-scored the neuronal activity per neuron using the mean and standard deviation of the training data. We utilized a nested 5-fold cross-validation, and separately optimized the regularization strength for each data split. The final decoding performance was calculated by averaging the performance across all five test sets.

Logistic regression weights for evaluating individual neuron's contribution

Applying an L1 penalty during training enforced feature sparsity, which facilitated the interpretation of input feature importance through the model's coefficients. A ranking of neurons was generated by evaluating the coefficients of the trained logistic regression models. Caution is needed when combining logistic regression weights from models trained on different splits due to the alternating test sets in each split, which sometimes include data used for training in another split's test data (see Fig. S7). To avoid any interference between training and test data across splits while still accounting for the temporal variation in our data, we further modified the splits used for cross-validation. The original five splits were extended to 20 splits in the following way: each of the original five splits was further divided into four sub-splits which shared a common test set but alternated the division of training and validation data (see visualization in Fig. S8). Any analysis based on neuron selection using logistic regression weights was evaluated exclusively on the corresponding four splits that generated the ranking and shared held-out test data. Final decoding performances for such analyses were derived through a nested cross-validation procedure. This involved initially averaging the decoding performances of models that shared a common test set (i.e. averaging across each set of four sub-splits) and then averaging the resulting five performances. In short, our training procedure for the logistic regression analysis consisted of training $5 \downarrow 4 = 20$ splits, and thus 20 models, where each group of 4 sub-splits shared a common test set.

Neurons were ranked according to their logistic regression coefficients across each set of four sub-splits. Given that there are a total of five such groups, this lead to five distinct rankings. To obtain each ranking, we combined the coefficients of the four trained models using a two-step procedure:

1. Partition the neurons into separate subsets, based on the number of models for which a neuron had a non-zero coefficient (e.g. one group of neurons which had non-zero coefficients on all four sub-splits, then all three sub-splits, etc.).
2. Within each subset of four models, use the average of the absolute coefficients across the five splits to obtain a subset-specific ranking.

By concatenating the partitions of ranked neurons, we obtained a comprehensive ranking of all neurons. The neuron ranked highest displayed non-zero coefficients in all four models (corresponding to four sub-splits) and possessed the greatest average absolute coefficient value among neurons activated in all four models.

Intersection of top-performing neurons and chance-level overlap

In our analysis, we restricted the decoding to subpopulations of neurons that were derived through a ranking of weights of a trained logistic regression model. To ensure that the selection of neurons was independent of the test data, the neuron selection procedure was based on five rankings derived from distinct data splits, each paired with fixed test data. We evaluated the intersection of neurons across subsets of top-ranked neurons from the five rankings, evaluated for varying subpopulation sizes (**Fig. 7b**). For instance, a comparison of the top-performing 500 neurons from each of the five rankings revealed a set of 78 common neurons.

As a reference point for comparison, we report the average count of overlapping neurons anticipated when randomly selecting sets of 500 neurons five times from the entire population (denoted as chance level of overlapping neurons in **Fig. 7**). In the previously mentioned scenario involving subpopulation sizes of 500, the expected number of overlapping neurons is equivalent to one. Mathematically, this is computed as follows: the full population consists of N = 2286 neurons. We refer to the size of the subpopulation as k = 500 and the number of total rankings m = 5. For each neuron n_i in the subpopulation, we define a random variable $X_{i,500}$ as follows:

$$X_{i,500} = \begin{cases} 1, & \text{if neuron } n_i \text{ lies in all five subpopulations} \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

We observe that $\mathbb{P}(X_{i,500} = 1) = \left(\frac{k}{N}\right)^m = \left(\frac{500}{2286}\right)^5$. The expected value of $X_{i,500}$ is given by

$$\mathbb{E}(X_{i,500}) = 1 \cdot \mathbb{P}(X_{i,500} = 1) + 0 \cdot \mathbb{P}(X_{i,500} = 0) = \mathbb{P}(X_{i,500} = 1)$$

We define $X_{500} = \sum_{i=1}^{N} X_{i,500}$ as the number of overlapping neurons across all five rankings. Since the random variables are independently and identically distributed, this implies

$$\mathbb{E}(X_{500}) = \sum_{i=1}^{N} \mathbb{E}(X_{i,500}) = \sum_{i=1}^{N} \mathbb{P}(X_{i,500} = 1) = 2286 \cdot \left(\frac{500}{2286}\right)^5 \approx 1.1443$$

Analogous calculations for k = 100, 350, 750, 1000 yield

$$\mathbb{E}(X_{100}) = 2286 \cdot \left(\frac{100}{2286}\right)^5 \approx 0.0004$$
$$\mathbb{E}(X_{350}) = 2286 \cdot \left(\frac{350}{2286}\right)^5 \approx 0.1923$$
$$\mathbb{E}(X_{750}) = 2286 \cdot \left(\frac{750}{2286}\right)^5 \approx 8.6896$$
$$\mathbb{E}(X_{1000}) = 2286 \cdot \left(\frac{1000}{2286}\right)^5 \approx 36.6180$$

Thus, we derive the chance baselines as 0, 0, 1, 9, and 37 for subpopulation sizes of 100, 350, 500, 750, and 1000, respectively.

Decoding on regions of the MTL

We compared the decoding performance when using the activity of all 2286 recorded neurons to the performance when only using activity from specific regions of the MTL. These regions are the amygdala (580 neurons), hippocampus (794), enthorinal cortex (440), and parahippocampal cortex (373). To use activity from a particular region, we limited ourselves to the activity of neurons in that region and reduced the input dimension to match the number of neurons in the region. The network architecture and data splits remained the same as when using activity from the full population, but the hyperparameters were optimized for the reduced dataset and given label. Training, validation, and test set sizes remained the same as the full dataset condition. In summary, decoding from different regions differed from full population decoding primarily due to reduced input data dimensionality: from a spike train of length 20 and dimension 2286 to a spike train of the same length but with a dimension reduced according to the number of neurons in the specific region.

Author Contributions

Conceptualization

Franziska Gerken, Alana Darcher, Pedro J Gonçalves, Rachel Rapp, Ismail Elezi, Johannes Niediek, Stefanie Liebe, Jakob H Macke, Florian Mormann, Laura Leal-Taixé

Data acquisition

Alana Darcher, Johannes Niediek, Thomas P Reber, Marcel S Kehl, Stefanie Liebe, Florian Mormann

Data curation

Alana Darcher, Marcel S Kehl

Methodology

Franziska Gerken, Alana Darcher, Pedro J Gonçalves, Rachel Rapp, Ismail Elezi, Stefanie Liebe, Jakob H Macke, Florian Mormann, Laura Leal-Taixé

Formal analysis

Franziska Gerken, Alana Darcher

Funding

Jakob H Macke, Florian Mormann, Laura Leal-Taixé

Software

Franziska Gerken, Alana Darcher, Johannes Niediek, Thomas P Reber

Project administration

Jakob H Macke, Florian Mormann, Laura Leal-Taixé

Supervision

Pedro J Gonçalves, Ismail Elezi, Stefanie Liebe, Jakob H Macke, Florian Mormann, Laura Leal-Taixé

Writing - original draft

Franziska Gerken, Alana Darcher

Writing - review and editing

Pedro J Gonçalves, Rachel Rapp, Ismail Elezi, Stefanie Liebe, Jakob H Macke, Florian Mormann, Laura Leal-Taixé

Data Availability Statement

The data will be made fully available in a separate publication following the release of this manuscript. Code will be made available on GitHub.

Acknowledgements

We would like to thank Tamara Müller and Aleksandar Levic for their contributions to the data analysis framework. This work was supported by the German Federal Ministry of Education and Research (DeepHumanVision, FKZ: 031L0197A-C; Tübingen AI Center, FKZ: 01IS18039), as well as the German Research Foundation (DFG) through Germany's Excellence Strategy (Cluster of Excellence Machine Learning for Science, EXC-Number 2064/1, PN 390727645) and SFB1233 (PN 276693517), SFB 1089 (PN 227953431), SPP 2411 (PN: 520287829), and MO 930/4-2.

Additional files

Supplementary materials.

References

[1] Kreiman G., Koch C., Fried I. (2000) **Category-specific visual responses of single neurons in the human medial temporal lobe** *Nature Neuroscience* **3**:946–953 Google Scholar

[2] Kraskov A., Quiroga R.Q., Reddy L., Fried I., Koch C. (2007) **Local field potentials and spikes in the human medial temporal lobe are selective to image category** *Journal of Cognitive Neuroscience* **19**:479–492 Google Scholar

[3] Quiroga R.Q., Reddy L., Kreiman G., Koch C., Fried I. (2005) **Invariant visual representation by single neurons in the human brain** *Nature* **435**:1102–1107 Google Scholar

[4] Reber T.P., Bausch M., Mackay S., Boström J., Elger C.E., Mormann F. (2019) **Representation of abstract semantic knowledge in populations of human single neurons in the medial temporal lobe** *PLOS Biology* **17**:1–17 Google Scholar

[5] Kreiman G., Fried I., Koch C. (2002) **Single-neuron correlates of subjective vision in the human medial temporal lobe** *Proceedings of the National Academy of Sciences* **99**:8378–8383 Google Scholar

[6] Quiroga R.Q., Mukamel R., Isham E.A., Malach R., Fried I. (2008) **Human single-neuron responses at the threshold of conscious recognition** *Proceedings of the National Academy of Sciences* **105**:3599–3604 Google Scholar

[7] Quiroga R.Q., Kraskov A., Mormann F., Fried I., Koch C. (2014) **Single-cell responses to face adaptation in the human medial temporal lobe** *Neuron* **84**:363–369 Google Scholar

[8] Reber T.P., Faber J., Niediek J., Boström J., Elger C.E., Mormann F. (2017) **Single-neuron correlates of conscious perception in the human medial temporal lobe** *Current Biology* **27**:2991–2998 Google Scholar

[9] Mackay S., Reber T.P., Bausch M., Boström J., Elger C.E., Mormann F. (2024) **Concept and location neurons in the human brain provide the 'what'and 'where'in memory formation** *Nature Communications* **15**:7926 Google Scholar

[10] Staresina B.P., Reber T.P., Niediek J., Boström J., Elger C.E., Mormann F. (2019) **Recollection in the human hippocampal-entorhinal cell circuitry** *Nature Communications* **10**:1503 Google Scholar

[11] Zheng J., Schjetnan A.G.P., Yebra M., Gomes B.A., Mosher C.P., Kalia S.K., Valiante T.A., Mamelak A.N., Kreiman G., Rutishauser U. (2022) **Neurons detect cognitive boundaries to structure episodic memories in humans** *Nature Neuroscience* **25**:358–368 Google Scholar

[12] Rey H.G., Ison M.J., Pedreira C., Valentin A., Alarcon G., Selway R., Richardson M.P., Quiroga R.Q. (2015) **Single-cell recordings in the human medial temporal lobe** *Journal of Anatomy* **227**:394–408 Google Scholar

[13] Rutishauser U., Reddy L., Mormann F., Sarnthein J. (2021) **The architecture of human memory: insights from human single-neuron recordings** *Journal of Neuroscience* **41**:883–890 Google Scholar

[14] Nishimoto S., Vu A.T., Naselaris T., Benjamini Y., Yu B., Gallant J.L. (2011) **Reconstructing visual experiences from brain activity evoked by natural movies** *Current Biology* **21**:1641–1646 Google Scholar

[15] Wen H., Shi J., Zhang Y., Lu K.-H., Cao J., Liu Z. (2017) **Neural encoding and decoding with deep learning for dynamic natural vision** *Cerebral Cortex* **28**:4136–4160 Google Scholar

[16] Huth A. G., Nishimoto S., Vu A. T., Gallant J. L. (2012) **A continuous semantic space describes the representation of thousands of object and action categories across the human brain** *Neuron* **76**:1210–1224 Google Scholar

[17] Haxby J. V., Gobbini M. I., Nastase S. A. (2020) **Naturalistic stimuli reveal a dominant role for agentic action in visual representation** *NeuroImage* **216**:116561 Google Scholar

[18] Hasson U., Nir Y., Levy I., Fuhrmann G., Malach R. (2004) **Intersubject synchronization of cortical activity during natural vision** *Science* **303**:1634–1640 Google Scholar

[19] Van der Meer J.N., Breakspear M., Chang L.J., Sonkusare S., Cocchi L. (2020) **Movie viewing elicits rich and reliable brain state dynamics** *Nature Communications* **11**:5004 Google Scholar

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

[21] Baldassano C., Chen J., Zadbood A., Pillow J.W., Hasson U., Norman K.A. (2017) **Discovering event structure in continuous narrative perception and memory** *Neuron* **95**:709–721 Google Scholar

[22] Isik L., Singer J., Madsen J.R., Kanwisher N., Kreiman G. (2018) **What is changing when: Decoding visual information in movies from human intracranial recordings** *NeuroImage* **180**:147–159 Google Scholar

[23] Gelbard-Sagiv H., Mukamel R., Harel M., Malach R., Fried I. (2008) **Internally gener'zated reactivation of single neurons in human hippocampus during free recall** *Science* **322**:96–101 Google Scholar

[24] Zhang Y., Aghajan Z.M., Ison M., Lu Q., Tang H., Kalender G., Monsoor T., Zheng J., Kreiman G., Roychowdhury V., Fried I. (2023) **Decoding of human identity by computer vision and neuronal vision** *Scientific Reports* **13**:1–16 Google Scholar

[25] StabilityAI (2024) **Stable diffusion** https://beta.dreamstudio.ai/generate

[26] Maris E., Oostenveld R. (2007) **Nonparametric statistical testing of eeg-and meg-data** *Journal of Neuroscience Methods* **164**:177–190 Google Scholar

[27] Panzeri S., Macke J.H., Gross J., Kayser C. (2015) **Neural population coding: combining insights from microscopic and mass signals** *Trends in Cognitive Sciences* **19**:162–172 Google Scholar

[28] Kamiński J., Sullivan S., Chung J.M., Ross I.B., Mamelak A.N., Rutishauser U. (2017) **Persistently active neurons in human medial frontal and medial temporal lobe support working memory** *Nature Neuroscience* **20**:590–601 Google Scholar

[29] Jamali M., Grannan B.L., Fedorenko E., Saxe R., Báez-Mendoza R., Williams Z.M. (2021) **Single-neuronal predictions of others' beliefs in humans** *Nature* **591**:610–614 Google Scholar

[30] Hochreiter S., Schmidhuber J. (1997) **Long short-term memory** *Neural Computation* **9**:1735–80 Google Scholar

[31] Fried I., MacDonald K.A., Wilson C.L. (1997) **Single neuron activity in human hippocampus and amygdala during recognition of faces and objects** *Neuron* **18**:753–765 Google Scholar

[32] Mormann F., Niediek J., Tudusciuc O., Quesada C.M., Coenen V.A., Elger C.E., Adolphs R. (2015) **Neurons in the human amygdala encode face identity, but not gaze direction** *Nature Neuroscience* **18**:1568–1570 Google Scholar

[33] Cao R., Li X., Brandmeir N.J., Wang S. (2021) **Encoding of facial features by single neurons in the human amygdala and hippocampus** *Communications Biology* **4**:1394 Google Scholar

[34] Bausch M., Niediek J., Reber T.P., Mackay S., Boström J., Elger C.E., Mormann F. (2021) **Concept neurons in the human medial temporal lobe flexibly represent abstract relations between concepts** *Nature Communications* **12**:6164 Google Scholar

[35] Garcia-Lazaro Jose A., Belliveau L.A.C., Lesica N.A. (2013) **Independent population coding of speech with sub-millisecond precision** *Journal of Neuroscience* **33**:19362–19372 Google Scholar

[36] Ince R.A.A., Panzeri S., Kayser C. (2013) **Neural codes formed by small and temporally precise populations in auditory cortex** *Journal of Neuroscience* **33**:18277–18287 Google Scholar

[37] Viskontas Indre V., Quiroga Rodrigo Quian, Fried Itzhak (2009) **Human medial temporal lobe neurons respond preferentially to personally relevant images** *Proceedings of the National Academy of Sciences* **106**:21329–21334 Google Scholar

[38] Rutishauser U., Tudusciuc O., Neumann D., Mamelak A. N., Heller A. C., Ross I. B., Philpott L., Sutherling W. W., Adolphs R. (2011) **Single-unit responses selective for whole faces in the human amygdala** *Current Biology* **21**:1654–1660 Google Scholar

[39] Cao R., Wang J., Brunner P., Willie J. T., Li X., Rutishauser U., Brandmeir N. J., Wang S. (2024) **Neural mechanisms of face familiarity and learning in the human amygdala and hippocampus** *Cell Reports* **43** Google Scholar

[40] Troiani V., Stigliani A., Smith M.E., Epstein R.A. (2014) **Multiple object properties drive scene-selective regions** *Cerebral Cortex* **24**:883–897 Google Scholar

[41] Dilks D.D., Julian J.B., Paunov A.M., Kanwisher N. (2013) **The occipital place area is causally and selectively involved in scene perception** *Journal of Neuroscience* **33**:1331–1336 Google Scholar

[42] Mormann F., Kornblith S., Cerf M., Ison M.J., Kraskov A., Tran M., Knieling S., Quian Quiroga R., Koch C., Fried I. (2017) **Scene-selective coding by single neurons in the human**

parahippocampal cortex *Proceedings of the National Academy of Sciences* **114**:1153–1158 Google Scholar

[43] Niediek J., Boström J., Elger C.E., Mormann F. (2016) **Reliable analysis of single-unit recordings from the human brain under noisy conditions: tracking neurons over hours** *PLOS One* **11**:e0166598 Google Scholar

[44] Dehnen G., Kehl M.S., Darcher A., Müller T.T., Macke J.H., Borger V., Surges R., Mormann F. (2021) **Duplicate detection of spike events: A relevant problem in human single-unit recordings** *Brain Sciences* **11**:761 Google Scholar

[45] Darcher A., Rapp R., Mueller T. (2024) **Epiphyte** https://github.com/mackelab/epiphyte

[46] (2012) **FFmpeg. https://ffmpeg.org/, 2012.** https://ffmpeg.org/

[47] Aubert O., Prié Y. (2005) **Advene: active reading through hypervideo** *Proceedings of ACM Hypertext* **05**:235–244 Google Scholar

[48] Esler T. (no date) **Facenet pytorch** https://github.com/timesler/facenet-pytorch.git

[49] Castellano B. (no date) **PySceneDetect** https://github.com/Breakthrough/PySceneDetect

[50] Virtanen P., Gommers R., Oliphant T.E., Haberland M., Reddy T., Cournapeau D., Burovski E., Peterson P., Weckesser W., Bright J., van der Walt S.J., Brett M., Wilson J., Millman K.J., Mayorov N., Nelson A.R.J., Jones E., Kern R., Larson E., Carey C.J., Polat İ, Feng Y., Moore E.W., VanderPlas J., Laxalde D., Perktold J., Cimrman R., Henriksen I., Quintero E.A., Harris C.R., Archibald A.M., Ribeiro A.H., Pedregosa F., van Mulbregt P., SciPy 1.0 Contributors (2020) **SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python** *Nature Methods* **17**:261–272 Google Scholar

[51] Simes R.J. (1986) **An improved Bonferroni procedure for multiple tests of significance** *Biometrika* **73**:751–754 Google Scholar

[52] Kingma D.P., Adam J. Ba. (2015) **A method for stochastic optimization, 2014** In: *Published as a conference paper at the 3rd International Conference for Learning Representations* Google Scholar

[53] Paszke A., Gross S., Massa F., Lerer A., Bradbury J., Chanan G., Killeen T., Lin Z., Gimelshein N., Antiga L., Desmaison A., Kopf A., Yang E., DeVito Z., Raison M., Tejani A., Chilamkurthy S., Steiner B., Fang L., Bai J., Chintala S. (2019) **Pytorch: An imperative style, high-performance deep learning library** *Advances in Neural Information Processing Systems* **32**:8024–8035 Google Scholar

[54] Cohen J. (1960) **A coefficient of agreement for nominal scales** *Educational and psychological measurement* **20**:37–46 Google Scholar

[55] Ekstrom A.D., Kahana M.J., Caplan J.B., Fields T.A., Isham E.A., Newman E.L., Fried I. (2003) **Cellular networks underlying human spatial navigation** *Nature* **425**:184–188 Google Scholar

[56] Miller J.F., Fried I., Suthana N., Jacobs J. (2015) **Repeating spatial activations in human entorhinal cortex** *Current Biology* **25**:1080–1085 Google Scholar

[57] Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay E. (2011) **Scikit-learn: Machine learning in Python** *Journal of Machine Learning Research* **12**:2825–2830 Google Scholar

Author information

Franziska Gerken*
Dynamic Vision and Learning Group, Technical University of Munich, Munich, Germany
ORCID iD: 0009-0004-9668-719X

*These authors contributed equally to this work.

Alana Darcher*
Department of Epileptology, University Medical Center of Bonn, Bonn, Germany
ORCID iD: 0000-0001-9048-9732

*These authors contributed equally to this work.

Pedro J Gonçalves
Machine Learning in Science, Excellence Cluster Machine Learning and Tübingen AI Center, University of Tübingen, Tübingen, Germany, VIB-Neuroelectronics Research Flanders (NERF), Leuven, Belgium, imec, Leuven, Belgium
ORCID iD: 0000-0002-6987-4836

Rachel Rapp
Machine Learning in Science, Excellence Cluster Machine Learning and Tübingen AI Center, University of Tübingen, Tübingen, Germany
ORCID iD: 0009-0001-1698-4767

Ismail Elezi§
Dynamic Vision and Learning Group, Technical University of Munich, Munich, Germany
ORCID iD: 0000-0002-7047-5808

§Currently at Huawei.

Johannes Niediek
Department of Epileptology, University Medical Center of Bonn, Bonn, Germany, Machine Learning Group, Technical University of Berlin, Berlin, Germany
ORCID iD: 0000-0003-3323-2986

Marcel S Kehl
Department of Epileptology, University Medical Center of Bonn, Bonn, Germany, Department of Experimental Psychology, University of Oxford, Oxford, United Kingdom
ORCID iD: 0000-0002-6812-8604

Thomas P Reber
Department of Epileptology, University Medical Center of Bonn, Bonn, Germany, Faculty of Psychology, UniDistance Suisse, Brig, Switzerland
ORCID iD: 0000-0002-3969-9782

Stefanie Liebe
Machine Learning in Science, Excellence Cluster Machine Learning and Tübingen AI Center, University of Tübingen, Tübingen, Germany, Department of Epileptology and Neurology, University Hospital Tübingen, Tübingen, Germany
ORCID iD: 0000-0003-2873-2943

Jakob H Macke‡
Machine Learning in Science, Excellence Cluster Machine Learning and Tübingen AI Center, University of Tübingen, Tübingen, Germany, Empirical Inference, Max Planck Institute for Intelligent Systems, Tübingen, Germany
ORCID iD: 0000-0001-5154-8912

For correspondence: Jakob.Macke@uni-tuebingen.de
‡These authors contributed equally to this work.

Florian Mormann‡
Department of Epileptology, University Medical Center of Bonn, Bonn, Germany
ORCID iD: 0000-0003-1305-8028

For correspondence: llealtaixe@nvidia.com
‡These authors contributed equally to this work.

Laura Leal-Taixé†‡
Dynamic Vision and Learning Group, Technical University of Munich, Munich, Germany
ORCID iD: 0000-0001-8709-1133

†Currently at NVIDIA.
‡These authors contributed equally to this work.

Editors

Reviewing Editor
Iris Groen
University of Amsterdam, Amsterdam, Netherlands

Senior Editor
Joshua Gold
University of Pennsylvania, Philadelphia, United States of America

Reviewer #1 (Public review):

Summary:

In this manuscript, Gerken et al examined how neurons in the human medial temporal lobe respond to and potentially code dynamic movie content. They had 29 patients watch a long-

form movie while neurons within their MTL were monitored using depth electrodes. They found that neurons throughout the region were responsive to the content of the movie. In particular, neurons showed significant responses to people, places, and to a lesser extent, movie cuts. Modeling with a neural network suggests that neural activity within the recorded regions was better at predicting the content of the movies as a population, as opposed to individual neural representations. Surprisingly, a subpopulation of unresponsive neurons performed better than the responsive neurons at decoding the movie content, further suggesting that while classically nonresponsive, these neurons nonetheless provided critical information about the content of the visual world. The authors conclude from these results that low-level visual features, such as scene cuts, may be coded at the neuronal level, but that semantic features rely on distributed population-level codes.

Strengths:

Overall, the manuscript presents an interesting and reasonable argument for their findings and conclusions. Additionally, the large number of patients and neurons that were recorded and analyzed makes this data set unique and potentially very powerful. On the whole, the manuscript was very well written, and as it is, presents an interesting and useful set of data about the intricacies of how dynamic naturalistic semantic information may be processed within the medial temporal lobe.

Weaknesses:

There are a number of concerns I have based on some of the experimental and statistical methods employed that I feel would help to improve our understanding of the current data.

In particular, the authors do not address the issue of superposed visual features very well throughout the manuscript. Previous research using naturalistic movies has shown that low-level visual features, particularly motion, are capable of driving much of the visual system (e.g, Bartels et al 2005; Bartels et al 2007; Huth et al 2012; Çukur et al 2013; Russ et al 2015; Nentwich et al 2023). In some of these papers, low-level features were regressed out to look at the influence of semantics, in others, the influence of low-level features was explicitly modeled. The current manuscript, for the most part, appears to ignore these features with the exception of scene cuts. Based on the previous evidence that low-level features continue to drive later cortical regions, it seems like including these as regressors of no interest or, more ideally, as additional variables, would help to determine how well MTL codes for semantic features over top of these lower-order variables.

Following on this, much of the current analyses rely on the training of deep neural networks to decode particular features. The results of these analyses are illuminating, however, throughout the manuscript, I was increasingly wondering how the various variables interact with each other. For example, separate analyses were done for the patients, regions, and visual features. However, the logistic regression analysis that was employed could have all of these variables input together, obtaining beta weights for each one in an overall model. This would potentially provide information about how much each variable contributes to the overall decoding in relation to the others.

A few more minor points that would help to clarify the current results involve the selection of data for particular analyses. For some analyses, the authors chose to appropriately downsample their data sets to compare across variables. However, there are a few places where similar downsampling would be informative, but was not completed. In particular, the analyses for patients and regions may have a more informative comparison if the full population were downsampled to match the size of the population for each patient or region of interest. This could be done with the Monte Carlo sampling that is used in other analyses, thus providing a control for population size while still sampling the full population.

https://doi.org/10.7554/eLife.106758.1.sa3

Reviewer #2 (Public review):

Summary:

This study introduces an exciting dataset of single-unit responses in humans during a naturalistic and dynamic movie stimulus, with recordings from multiple regions within the medial temporal lobe. The authors use both a traditional firing-rate analysis as well as a sophisticated decoding analysis to connect these neural responses to the visual content of the movie, such as which character is currently on screen.

Strengths:

The results reveal some surprising similarities and differences between these two kinds of analyses. For visual transitions (such as camera angle cuts), the neurons identified in the traditional response analysis (looking for changes in firing rate of an individual neuron at a transition) were the most useful for doing population-level decoding of these cuts. Interestingly, this wasn't true for character decoding; excluding these "responsive" neurons largely did not impact population-level decoding, suggesting that the population representation is distributed and not well-captured by individual-neuron analyses.

The methods and results are well-described both in the text and in the figures. This work could be an excellent starting point for further research on this topic to understand the complex representational dynamics of single neurons during naturalistic perception.

Weaknesses:

(1) I am unsure what the central scientific questions of this work are, and how the findings should impact our understanding of neural representations. Among the questions listed in the introduction is "Which brain regions are informative for specific stimulus categories?". This is a broad research area that has been addressed in many neuroimaging studies for decades, and it's not clear that the results tell us new information about region selectivity. "Is the relevant information distributed across the neuronal population?" is also a question with a long history of work in neuroscience about localist vs distributed representations, so I did not understand what specific claim was being made and tested here. Responses in individual neurons were found for all features across many regions (e.g., Table S1), but decodable information was also spread across the population.

(2) The character and indoor/outdoor labels seem fundamentally different from the scene/camera cut labels, and I was confused by the way that the cuts were put into the decoding framework. The decoding analyses took a 1600ms window around a frame of the video (despite labeling these as frame "onsets" like the feature onsets in the responsive-neuron analysis, I believe this is for any frame regardless of whether it is the onset of a feature), with the goal of predicting a binary label for that frame. Although this makes sense for the character and indoor/outdoor labels, which are a property of a specific frame, it is confusing for the cut labels since these are inherently about a change across frames. The way the authors handle this is by labeling frames as cuts if they are in the 520ms following a cut (there is no justification given for this specific value). Since the input to a decoder is 1600ms, this seems like a challenging decoding setup; the model must respond that an input is a "cut" if there is a cut-specific pattern present approximately in the middle of the window, but not if the pattern appears near the sides of the window. A more straightforward approach would be, for example, to try to discriminate between windows just after a cut versus windows during other parts of the video. It is also unclear how neurons "responsive" to cuts were defined, since the authors state that this was determined by looking for times when a feature

was absent for 1000ms to continuously present for 1000ms, which would never happen for cuts (unless this definition was different for cuts?).

(3) The architecture of the decoding model is interesting but needs more explanation. The data is preprocessed with "a linear layer of same size as the input" (is this a layer added to the LSTM that is also trained for classification, or a separate step?), and the number of linear layers after the LSTM is "adapted" for each label type (how many were used for each label?). The LSTM also gets to see data from 800 ms before and after the labeled frame, but usually LSTMs have internal parameters that are the same for all timesteps; can the model know when the "critical" central frame is being input versus the context, i.e., are the inputs temporally tagged in some way? This may not be a big issue for the character or location labels, which appear to be contiguous over long durations and therefore the same label would usually be present for all 1600ms, but this seems like a major issue for the cut labels since the window will include a mix of frames with opposite labels.

(4) Because this is a naturalistic stimulus, some labels are very imbalanced ("Persons" appears in almost every frame), and the labels are correlated. The authors attempt to address the imbalance issue by oversampling the minority class during training, though it's not clear this is the right approach since the test data does not appear to be oversampled; for example, training the Persons decoder to label 50% of training frames as having people seems like it could lead to poor performance on a test set with nearly 100% Persons frames, versus a model trained to be biased toward the most common class. There is no attempt to deal with correlated features, which is especially problematic for features like "Summer Faces" and "Summer Presence", which I would expect to be highly overlapping, making it more difficult to interpret decoding performance for specific features.

(5) Are "responsive" neurons defined as only those showing firing increases at a feature onset, or would decreased activity also count as responsive? If only positive changes are labeled responsive, this would help explain how non-responsive neurons could be useful in a decoding analysis.

(6) Line 516 states that the scene cuts here are analogous to the hard boundaries in Zheng et al. (2022), but the hard boundaries are transitions between completely unrelated movies rather than scenes within the same movie. Previous work has found that within-movie and across-movie transitions may rely on different mechanisms, e.g., see Lee & Chen, 2022 (10.7554/eLife.73693).

https://doi.org/10.7554/eLife.106758.1.sa2

Reviewer #3 (Public review):

This is an excellent, very interesting paper. There is a groundbreaking analysis of the data, going from typical picture presentation paradigms to more realistic conditions. I would like to ask the authors to consider a few points in the comments below.

(1) From Figure 2, I understand that there are 7 neurons responding to the character Summer, but then in line 157, we learn that there are 46. Are the other 39 from other areas (not parahippocampal)? If this is the case, it would be important to see examples of these responses, as one of the main claims is that it is possible to decode as good or better with non-responsive compared to single responsive neurons, which is, in principle, surprising.

(2) Also in Figure 2, there seem to be relatively very few neurons responding to Summer (1.88%) and to outdoor scenes (1.07%). Is this significant? Isn't it also a bit surprising, particularly for outdoor scenes, considering a previous paper of Mormann showing many outdoor scene responses in this area? It would be nice if the authors could comment on this.

(3) I was also surprised to see that there are many fewer responses to scene cuts (6.7%) compared to camera cuts (51%) because every scene cut involves a camera cut. Could this have been a result of the much larger number of camera cuts? (A way to test this would be to subsample the camera cuts.)

(4) Line 201. The analysis of decoding on a per-patient basis is important, but it should be done on a per-session basis - i.e., considering only simultaneously recorded neurons, without any pooling. This is because pooling can overestimate decoding performances (see e.g. Quian Quiroga and Panzeri NRN 2009). If there was only one session per patient, then this should be called 'per-session' rather than 'per-patient' to make it clear that there was no pooling.

(5) In general, the decoding results are quite interesting, and I was wondering if the authors could give a bit more insight by showing confusion matrices, with the predictions of the appearance of each of the characters, etc. Some of the characters may appear together, so this could be another entry of the decoder (say, predicting person A, B, C, A&B, A&C, B&C, A&B&C). I guess this could also show the power of analyzing the population activity.

(6) Lines 406-407. The claim that stimulus-selective responses to characters did not account for the decoding of the same character is very surprising. If I understood it correctly, the response criterion the authors used gives 'responsiveness' but not 'selectivity'. So, were people's responses selective (e.g., firing only to Summer) or non-selective (firing to a few characters)? This could explain why they didn't get good decoding results with responsive neurons. Again, it would be nice to see confusion matrices with the decoding of the characters. Another reason for this is that what are labelled as responsive neurons have relatively weak and variable responses.

(7) Line 455. The claim that 500 neurons drive decoding performance is very subjective. 500 neurons gives a performance of 0.38, and 50 neurons gives 0.33.

(8) Lines 492-494. I disagree with the claim that "character decoding does not rely on individual cells, as removing neurons that responded strongly to character onset had little impact on performance". I have not seen strong responses to characters in the paper. In particular, the response to Summer in Figure 2 looks very variable and relatively weak. If there are stronger responses to characters, please show them to make a convincing argument. It is fine to argue that you can get information from the population, but in my view, there are no good single-cell responses (perhaps because the actors and the movie were unknown to the subjects) to make this claim. Also, an older paper (Quian Quiroga et al J. Neurophysiol. 2007) showed that the decoding of individual stimuli in a picture presentation paradigm was determined by the responsive neurons and that the non-responsive neurons did not add any information. The results here could be different due to the use of movies instead of picture presentations, but most likely due to the fact that, in the picture presentation paradigm, the pictures were of famous people for which there were strong single neuron responses, unlike with the relatively unknown persons in this paper.

https://doi.org/10.7554/eLife.106758.1.sa1

Author response:

> ***Reviewer #1 (Public review):***
>
> *Summary:*
>
> *In this manuscript, Gerken et al examined how neurons in the human medial temporal lobe respond to and potentially code dynamic movie content. They had 29 patients watch a long-form movie while neurons within their MTL were monitored using depth*

> *electrodes. They found that neurons throughout the region were responsive to the content of the movie. In particular, neurons showed significant responses to people, places, and to a lesser extent, movie cuts. Modeling with a neural network suggests that neural activity within the recorded regions was better at predicting the content of the movies as a population, as opposed to individual neural representations. Surprisingly, a subpopulation of unresponsive neurons performed better than the responsive neurons at decoding the movie content, further suggesting that while classically nonresponsive, these neurons nonetheless provided critical information about the content of the visual world. The authors conclude from these results that low-level visual features, such as scene cuts, may be coded at the neuronal level, but that semantic features rely on distributed population-level codes.*
>
> *Strengths:*
>
> *Overall, the manuscript presents an interesting and reasonable argument for their findings and conclusions. Additionally, the large number of patients and neurons that were recorded and analyzed makes this data set unique and potentially very powerful. On the whole, the manuscript was very well written, and as it is, presents an interesting and useful set of data about the intricacies of how dynamic naturalistic semantic information may be processed within the medial temporal lobe.*

We thank the reviewer for their comments on our manuscript and for describing the strengths of our presented work

> *Weaknesses:*
>
> *There are a number of concerns I have based on some of the experimental and statistical methods employed that I feel would help to improve our understanding of the current data.*
>
> *In particular, the authors do not address the issue of superposed visual features very well throughout the manuscript. Previous research using naturalistic movies has shown that low-level visual features, particularly motion, are capable of driving much of the visual system (e.g, Bartels et al 2005; Bartels et al 2007; Huth et al 2012; Çukur et al 2013; Russ et al 2015; Nentwich et al 2023). In some of these papers, low-level features were regressed out to look at the influence of semantics, in others, the influence of low-level features was explicitly modeled. The current manuscript, for the most part, appears to ignore these features with the exception of scene cuts. Based on the previous evidence that low-level features continue to drive later cortical regions, it seems like including these as regressors of no interest or, more ideally, as additional variables, would help to determine how well MTL codes for semantic features over top of these lower-order variables.*

We thank the reviewer for this insightful comment and for the relevant literature regarding visual motion in not only the primary visual system but in cortical areas as well. While we agree that the inclusion of visual motion as a regressor of no interest or as an additional variable would be overall informative in determining if single neurons in the MTL are driven by this level of feature, we would argue that our analyses already provide some insight into its role and that only the parahippocampal cortical neurons would robustly track this feature.

As noted by the reviewer, our model includes two features derived from visual motion: Camera Cuts (directly derived from frame-wise changes in pixel values) and Scene Cuts (a subset of Camera Cuts restricted to changes in scene). As shown in Fig. 5a, decoding performance for these features was strongest in the parahippocampal cortex (~20%), compared to other MTL areas (~10%). While the entorhinal cortex also showed some

performance for Scene Cuts (15%), we interpret this as being driven by the changes in location that define a scene, rather than by motion itself.

These findings suggest that while motion features are tracked in the MTL, the effect may be most robust in the parahippocampal cortex. We believe that quantifying more complex 3D motion in a naturalistic stimulus like a full-length movie is a significant challenge that would likely require a dedicated study. We agree this is an interesting future research direction and will update the manuscript to highlight this for the reader.

> *A few more minor points that would help to clarify the current results involve the selection of data for particular analyses. For some analyses, the authors chose to appropriately downsample their data sets to compare across variables. However, there are a few places where similar downsampling would be informative, but was not completed. In particular, the analyses for patients and regions may have a more informative comparison if the full population were downsampled to match the size of the population for each patient or region of interest. This could be done with the Monte Carlo sampling that is used in other analyses, thus providing a control for population size while still sampling the full population.*

We thank the reviewer for raising this important methodological point. The decision not to downsample the patient- and region-specific analyses was deliberate, and we appreciate the opportunity to clarify our rationale.

Generally, we would like to emphasize that due to technical and ethical limitations of human single-neuron recordings, it is currently not possible to record large populations of neurons simultaneously in individual patients. The limited and variable number of recorded neurons per subject (Fig. S1) generally requires pooling neurons into a pseudo-populations for decoding, which is a well-established standard in human single-neuron studies (see e.g., (Jamali et al., 2021; Kamiński et al., 2017; Minxha et al., 2020; Rutishauser et al., 2015; Zheng et al., 2022)).

For the patient-specific analysis, our primary goal was to show that no single patient's data could match the performance of the complete pseudo-population. Crucially, we found no direct relationship between the number of recorded neurons and decoding performance; patients with the most neurons (patients 4, 13) were not top performers, and those with the fewest (patients 11, 14) were not the worst (see Fig. 4). This indicates that neuron count was not the primary limiting factor and that downsampling would be unlikely to provide additional insight.

Similarly, for the region-specific analysis, regions with larger neural populations did not systematically outperform those with fewer neurons (Fig. 5). Given the inherent sparseness of single-neuron data, we concluded that retaining the full dataset was more informative than excluding neurons simply to equalize population sizes.

We agree that this methodological choice should be transparent and explicitly justified in the text. We will add an explanation to the revised manuscript to justify why this approach was taken and how it differs from the analysis in Fig. 6.

> ***Reviewer #2 (Public review):***
>
> *Summary:*
>
> *This study introduces an exciting dataset of single-unit responses in humans during a naturalistic and dynamic movie stimulus, with recordings from multiple regions within the medial temporal lobe. The authors use both a traditional firing-rate analysis as well*

> *as a sophisticated decoding analysis to connect these neural responses to the visual content of the movie, such as which character is currently on screen.*
>
> *Strengths:*
>
> *The results reveal some surprising similarities and differences between these two kinds of analyses. For visual transitions (such as camera angle cuts), the neurons identified in the traditional response analysis (looking for changes in firing rate of an individual neuron at a transition) were the most useful for doing population-level decoding of these cuts. Interestingly, this wasn't true for character decoding; excluding these "responsive" neurons largely did not impact population-level decoding, suggesting that the population representation is distributed and not well-captured by individual-neuron analyses.*
>
> *The methods and results are well-described both in the text and in the figures. This work could be an excellent starting point for further research on this topic to understand the complex representational dynamics of single neurons during naturalistic perception.*

We thank the reviewer for their feedback and for summarizing the results of our work.

> *(1) I am unsure what the central scientific questions of this work are, and how the findings should impact our understanding of neural representations. Among the questions listed in the introduction is "Which brain regions are informative for specific stimulus categories?". This is a broad research area that has been addressed in many neuroimaging studies for decades, and it's not clear that the results tell us new information about region selectivity. "Is the relevant information distributed across the neuronal population?" is also a question with a long history of work in neuroscience about localist vs distributed representations, so I did not understand what specific claim was being made and tested here. Responses in individual neurons were found for all features across many regions (e.g., Table S1), but decodable information was also spread across the population.*

We thank the reviewer for this important point, which gets to the core of our study's contribution. While concepts like regional specificity are well-established from studies on the blood-flow level, their investigation at the single-neuron level in humans during naturalistic, dynamic stimulation remains a critical open question. The type of coding (sparse vs. distributed) on the other hand cannot be investigated with blood-flow studies as the technology lacks the spatial and temporal resolution.

Our study addresses this gap directly. The exceptional temporal resolution of single-neuron recordings allows us to move beyond traditional paradigms and examine cellular-level dynamics as they unfold in neuronal response on a frame-by-frame basis to a more naturalistic and ecologically valid stimulus. It cannot be assumed that findings from other modalities or simplified stimuli will generalize to this context.

To meet this challenge, we employed a dual analytical strategy: combining a classic single-unit approach with a machine learning-based population analysis. This allowed us to create a bridge between prior work and our more naturalistic data. A key result is that our findings are often consistent with the existing literature, which validates the generalizability of those principles. However, the differences we observe between these two analytical approaches are equally informative, providing new insights into how the brain processes continuous, real-world information.

We will revise the introduction and discussion to more explicitly frame our work in this context, emphasizing the specific scientific question driving this study, while also highlighting the strengths of our experimental design and recording methods.

> *(2) The character and indoor/outdoor labels seem fundamentally different from the scene/camera cut labels, and I was confused by the way that the cuts were put into the decoding framework. The decoding analyses took a 1600ms window around a frame of the video (despite labeling these as frame "onsets" like the feature onsets in the responsive-neuron analysis, I believe this is for any frame regardless of whether it is the onset of a feature), with the goal of predicting a binary label for that frame. Although this makes sense for the character and indoor/outdoor labels, which are a property of a specific frame, it is confusing for the cut labels since these are inherently about a change across frames. The way the authors handle this is by labeling frames as cuts if they are in the 520ms following a cut (there is no justification given for this specific value). Since the input to a decoder is 1600ms, this seems like a challenging decoding setup; the model must respond that an input is a "cut" if there is a cut-specific pattern present approximately in the middle of the window, but not if the pattern appears near the sides of the window. A more straightforward approach would be, for example, to try to discriminate between windows just after a cut versus windows during other parts of the video. It is also unclear how neurons "responsive" to cuts were defined, since the authors state that this was determined by looking for times when a feature was absent for 1000ms to continuously present for 1000ms, which would never happen for cuts (unless this definition was different for cuts?).*

We thank the reviewer for the valuable comment regarding specifically the cut labels. The choice to label frames that lie in a time window of 520ms following a cut as positive was selected based on prior research and is intended to include the response onsets across all regions within the MTL (Mormann et al., 2008). We agree that this explanation is currently missing from the manuscript, and we will add a brief clarification in the revised version.

As correctly noted, the decoding analysis does not rely on feature onset but instead continuously decodes features throughout the entire movie. Thus, all frames are included, regardless of whether they correspond to a feature onset.

Our treatment of cut labels as sustained events is a deliberate methodological choice. Neural responses to events like cuts often unfold over time, and by extending the label, we provide our LSTM network with the necessary temporal window to learn this evolving signature. This approach not only leverages the sequential processing strengths of the LSTM (Hochreiter et al., 1997) but also ensures a consistent analytical framework for both event-based (cuts) and state-based (character or location) features.

> *(3) The architecture of the decoding model is interesting but needs more explanation. The data is preprocessed with "a linear layer of same size as the input" (is this a layer added to the LSTM that is also trained for classification, or a separate step?), and the number of linear layers after the LSTM is "adapted" for each label type (how many were used for each label?). The LSTM also gets to see data from 800 ms before and after the labeled frame, but usually LSTMs have internal parameters that are the same for all timesteps; can the model know when the "critical" central frame is being input versus the context, i.e., are the inputs temporally tagged in some way? This may not be a big issue for the character or location labels, which appear to be contiguous over long durations and therefore the same label would usually be present for all 1600ms, but this seems like a major issue for the cut labels since the window will include a mix of frames with opposite labels.*

We thank the reviewer for their insightful comments regarding the decoding architecture. The model consists of an LSTM followed by 1–3 linear readout layers, where the exact number of layers is treated as a hyperparameter and selected based on validation performance for each label type. The initial linear layer applied to the input is part of the

trainable model and serves as a projection layer to transform the binned neural activity into a suitable feature space before feeding it into the LSTM. The model is trained in an end-to-end fashion on the classification task.

Regarding temporal context, the model receives a 1600 ms window (800 ms before and after the labeled frame), and as correctly pointed out by the reviewer, LSTM parameters are shared across time steps. We do not explicitly tag the temporal position of the central frame within the sequence. While this may have limited impact for labels that persist over time (e.g., characters or locations), we agree this could pose a challenge for cut labels, which are more temporally localized.

This is an important point, and we will clarify this limitation in the revised manuscript and consider incorporating positional encoding in future work to better guide the model's focus within the temporal window. Additionally, we will add a data table, specifying the ranges of hyperparameters in our decoding networks. Hyperparameters were optimized for each feature and split individually, but we agree that some more details on how these parameters were chosen are important and we will provide a data table in our revised manuscript giving more insights into the ranges of hyperparameters.

We thank the reviewer for this important point. We will clarify this limitation in the revised manuscript and note that positional encoding is a valuable direction to better guide the model's focus within the temporal window. To improve methodological transparency, we will also add a supplementary table detailing the hyperparameter ranges used for our optimization process.

> *(4) Because this is a naturalistic stimulus, some labels are very imbalanced ("Persons" appears in almost every frame), and the labels are correlated. The authors attempt to address the imbalance issue by oversampling the minority class during training, though it's not clear this is the right approach since the test data does not appear to be oversampled; for example, training the Persons decoder to label 50% of training frames as having people seems like it could lead to poor performance on a test set with nearly 100% Persons frames, versus a model trained to be biased toward the most common class. [...]*

We thank the reviewer for this critical and thoughtful comment. We agree that the imbalanced and correlated nature of labels in naturalistic stimuli is a key challenge.

To address this, we follow a standard machine learning practice: oversampling is applied exclusively to the training data. This technique helps the model learn from underrepresented classes by creating more balanced training batches, thus preventing it from simply defaulting to the majority class. Crucially, the test set remains unaltered to ensure our evaluation reflects the model's true generalization performance on the natural data distribution.

For the "Persons" feature, which appears in nearly all frames, defining a meaningful negative class is particularly challenging. The decoder must learn to identify subtle variations within a highly skewed distribution. Oversampling during training helps provide a more balanced learning signal, while keeping the test distribution intact ensures proper evaluation of generalization.

The reviewer's comment—that we are "training the Persons decoder to label 50% of training frames as having people"—may suggest that labels were modified. We want to emphasize this is not the case. Our oversampling strategy does not alter the labels; it simply increases the exposure of the rare, underrepresented class during training to ensure the model can learn its pattern despite its low frequency.

We will revise the Methods section to describe this standard procedure more explicitly, clarifying that oversampling is a training-only strategy to mitigate class imbalance.

> *(5) Are "responsive" neurons defined as only those showing firing increases at a feature onset, or would decreased activity also count as responsive? If only positive changes are labeled responsive, this would help explain how non-responsive neurons could be useful in a decoding analysis.*

We define responsive neurons as those showing increased firing rates at feature onset; we did not test for decreases in activity. We thank the reviewer for this valuable comment and will address this point in the revised manuscript by assessing responseness without a restriction on the direction of the firing rate.

> *(6) Line 516 states that the scene cuts here are analogous to the hard boundaries in Zheng et al. (2022), but the hard boundaries are transitions between completely unrelated movies rather than scenes within the same movie. Previous work has found that within-movie and across-movie transitions may rely on different mechanisms, e.g., see Lee & Chen, 2022 (10.7554/eLife.73693).*

We thank the reviewer for pointing out this distinction and for including the relevant work from Lee & Chan (2022) which further contextualizes this distinction. Indeed, the hard boundaries defined in the cited paper differ slightly from ours. The study distinguishes between (1) hard boundaries—transitions between unrelated movies—and (2) soft boundaries—transitions between related events within the same movie. While our camera cuts resemble their soft boundaries, our scene cuts do not fully align with either category. We defined scene cuts to be more similar to the study's hard boundaries, but we recognize this correspondence is not exact. We will clarify the distinctions between our scene cuts and the hard boundaries described in Zheng et al. (2022) in the revised manuscript, and will update our text to include the finding from Lee & Chan (2022).

> ***Reviewer #3 (Public review):***
>
> *This is an excellent, very interesting paper. There is a groundbreaking analysis of the data, going from typical picture presentation paradigms to more realistic conditions. I would like to ask the authors to consider a few points in the comments below.*
>
> *(1) From Figure 2, I understand that there are 7 neurons responding to the character Summer, but then in line 157, we learn that there are 46. Are the other 39 from other areas (not parahippocampal)? If this is the case, it would be important to see examples of these responses, as one of the main claims is that it is possible to decode as good or better with non-responsive compared to single responsive neurons, which is, in principle, surprising.*

We thank the reviewer for pointing out this ambiguity in the text. Yes, the other 39 units are responsive neurons from other areas. We will clarify to which neuronal sets the number of responsive neurons corresponds. We will also include response plots depicting the unit activity for the mentioned units.

> *(2) Also in Figure 2, there seem to be relatively very few neurons responding to Summer (1.88%) and to outdoor scenes (1.07%). Is this significant? Isn't it also a bit surprising, particularly for outdoor scenes, considering a previous paper of Mormann showing many outdoor scene responses in this area? It would be nice if the authors could comment on this.*

We thank the reviewer for this insightful point. While a low response to the general 'outdoor scene' label seems surprising at first, our findings align with the established role of the parahippocampal cortex (PHC) in processing scenes and spatial layouts. In previous work using static images, each image introduces a new spatial context. In our movie stimulus, new spatial contexts specifically emerge at scene cuts. Accordingly, our data show a strong PHC response precisely at these moments. We will revise the discussion to emphasize this interpretation, highlighting the consistency with prior work.

Regarding the first comment, we did not originally test if the proportion of the units is significant using e.g. a binomial test. We will include the results of a binomial test for each region and feature pair in the revised manuscript.

> *(3) I was also surprised to see that there are many fewer responses to scene cuts (6.7%) compared to camera cuts (51%) because every scene cut involves a camera cut. Could this have been a result of the much larger number of camera cuts? (A way to test this would be to subsample the camera cuts.)*

The decrease in responsive units for scene cuts relative to camera cuts could indeed be due to the overall decrease in “trials” from one label to the other. To test this, we will follow the reviewer’s suggestion and perform tests using sets of randomly subsampled camera cuts and will include the results in the revised manuscript.

> *(4) Line 201. The analysis of decoding on a per-patient basis is important, but it should be done on a per-session basis - i.e., considering only simultaneously recorded neurons, without any pooling. This is because pooling can overestimate decoding performances (see e.g. Quian Quiroga and Panzeri NRN 2009). If there was only one session per patient, then this should be called 'per-session' rather than 'per-patient' to make it clear that there was no pooling.*

The per-patient decoding was indeed also a per-session decoding, as each patient contributed only a single session to the dataset. We will make note of this explicitly in the text to resolve the ambiguity.

> *(6) Lines 406-407. The claim that stimulus-selective responses to characters did not account for the decoding of the same character is very surprising. If I understood it correctly, the response criterion the authors used gives 'responsiveness' but not 'selectivity'. So, were people's responses selective (e.g., firing only to Summer) or non-selective (firing to a few characters)? This could explain why they didn't get good decoding results with responsive neurons. Again, it would be nice to see confusion matrices with the decoding of the characters. Another reason for this is that what are labelled as responsive neurons have relatively weak and variable responses.*

We thank the reviewer for pointing out the importance of selectivity in addition to responsiveness. Indeed, our response criterion does not take stimulus selectivity into account and exclusively measures increases in firing activity after feature onsets for a given feature irrespective of other features.

We will adjust the text to reflect this shortcoming of the response-detection approach used here. To clarify the relationship between neural populations, we will add visualizations of the overlap of responsive neurons across labels for each subregion. These figures will be included in the revised manuscript.

In our approach, we trained separate networks for each feature to effectively mitigate the issue of correlated feature labels within the dataset (see earlier discussion). While this

strategy effectively deals with the correlated features, it precluded the generation of standard confusion matrices, as classification was performed independently for each feature.

To directly assess the feature selectivity of responsive neurons, we will fit generalized linear models to predict their firing rates from the features. This approach will enable us to quantify their selectivity and compare it to that of the broader neuronal population.

> *(7) Line 455. The claim that 500 neurons drive decoding performance is very subjective. 500 neurons gives a performance of 0.38, and 50 neurons gives 0.33.*

We agree with the reviewer that the phrasing is unclear. We will adjust our summary of this analysis as given in Line 455 to reflect that the logistic regression-derived neuronal rankings produce a subset which achieve comparable performance.

> *(8) Lines 492-494. I disagree with the claim that "character decoding does not rely on individual cells, as removing neurons that responded strongly to character onset had little impact on performance". I have not seen strong responses to characters in the paper. In particular, the response to Summer in Figure 2 looks very variable and relatively weak. If there are stronger responses to characters, please show them to make a convincing argument. It is fine to argue that you can get information from the population, but in my view, there are no good single-cell responses (perhaps because the actors and the movie were unknown to the subjects) to make this claim. Also, an older paper (Quian Quiroga et al J. Neurophysiol. 2007) showed that the decoding of individual stimuli in a picture presentation paradigm was determined by the responsive neurons and that the non-responsive neurons did not add any information. The results here could be different due to the use of movies instead of picture presentations, but most likely due to the fact that, in the picture presentation paradigm, the pictures were of famous people for which there were strong single neuron responses, unlike with the relatively unknown persons in this paper.*

This is an important point and we thank the reviewer for highlighting a previous paradigm in which responsive neurons did drive decoding performance. Indeed, the fact that the movie, its characters and the corresponding actors were novel to patients could explain the disparity in decoding performance by way of weaker and more variable responses. We will include additional examples in the supplement of responses to features. Additionally, we will modify the text to emphasize the point that reliable decoding is possible even in the absence of a robust set of neuronal responses. It could indeed be the case that a decoder would place more weight on responsive units if they were present (as shown in the mentioned paper and in our decoding from visual transitions in the parahippocampal cortex).

https://doi.org/10.7554/eLife.106758.1.sa0