Single-neuron spiking variability in hippocampus dynamically tracks sensory content during memory formation in humans

Leonhard Waschke[1,2], Fabian Kamp[1,2], Evi van den Elzen[3], Suresh Krishna[4], Ulman Lindenberger[1,2], Ueli Rutishauser[5-8], & Douglas D. Garrett[1,2]

[1]Max Planck UCL Centre for Computational Psychiatry and Ageing Research, Max Planck Institute for Human Development, 14195 Berlin, Germany. [2]Center for Lifespan Psychology, Max Planck Institute for Human Development, 14195 Berlin, Germany. [3]Tilburg School of Social and Behavioral Sciences, Tilburg University, Tilburg, The Netherlands. [4]Department of Physiology, McGill University, Montreal, Canada. [5]Department of Neurosurgery, Cedars-Sinai Medical Center, Los Angeles, CA, USA. [6]Department of Neurology, Cedars-Sinai Medical Center, Los Angeles, CA, USA. [7]Division of Biology and Bioengineering, California Institute of Technology, Pasadena, CA, USA. [8]Center for Neural Science and Medicine, Department of Biomedical Sciences, Cedars-Sinai Medical Center, Los Angeles, CA, USA.

During memory formation, the hippocampus is presumed to represent the “content” of stimuli, but how it does so is unknown. Using computational modelling and human single-neuron recordings, we show that the more precisely hippocampal spiking variability tracks the composite features that comprise each individual stimulus, the better those stimuli are later remembered. We propose that moment-to-moment spiking variability may provide a new window into how the hippocampus constructs memories from the building blocks of our sensory world.

Prior to memory formation, visual stimulus properties are encoded along the ventral visual pathway (*1*, *2*). Neural selectivity is thought to shift from low-level image properties towards more composite features as information propagates from visual cortex to the medial temporal lobe (MTL) (*3*). In the hippocampus, a high-level end point along this pathway, a variety of visual features are believed to be transformed into conjunct representations during memory encoding (*4*–*6*), but what granularity of information can be conjuncted by the hippocampus is not clear. Beyond the behavioural relevance of basic sensory encoding in cortex (*5*, *6*), it is plausible that sensory features also need to be directly accessible to the hippocampus to enable the formation of conjunctive representations (*7*, *8*). Given the absence of direct visuo-cortical afferents to hippocampal areas (*1*, *3*), it is plausible that hippocampus may preferably track more composite rather than simpler features to form conjunctive representations (*3*). However, the tracking of simpler visual features in hippocampus may nevertheless also be crucial for the formation of detailed memory traces (*5*–*7*). Such a direct comparison of the types of visual “building blocks” required for hippocampal memory encoding has not yet been made. One primary challenge in probing this fundamental question is that tailored experimental approaches that can separate simple and composite sensory features have not been leveraged. We argue that the architecture of multi-layer computational vision models can be used to differentiate between simple and composite visual features of *any* stimulus a participant may encode (see below), in turn permitting direct testing of the sensory feature space the hippocampus leverages during memory formation. But what signature of hippocampal activity should track differential visual features in this context?

In recent years, the moment-to-moment variability of neural activity has emerged as a behaviourally relevant measure that offers substantial insights beyond conventional approaches such as average brain activity (*9*). The processing of different visual features indeed uniquely impacts single-neuron spiking variability in visual cortex (*10*, *11*). Remarkably, individuals who exhibit increased visuo-cortical BOLD fMRI variability in response to more feature-rich stimuli also display superior cognitive performance (*12*). Such dynamic responsivity in visual cortex presumably captures differential perceptual information (*13*), requiring neurons to respond variably depending on what visual input is being processed (*11*). However, it is unknown whether more differentiated visual input also results in variable hippocampal activity during memory formation, and if individual differences therein might account for one’s level of memory performance. We posit that individuals with stronger trial-by-trial coupling between hippocampal variability and the content of visual input should also exhibit superior memory, providing evidence that visual features have successfully been encoded by the hippocampus during memory formation.

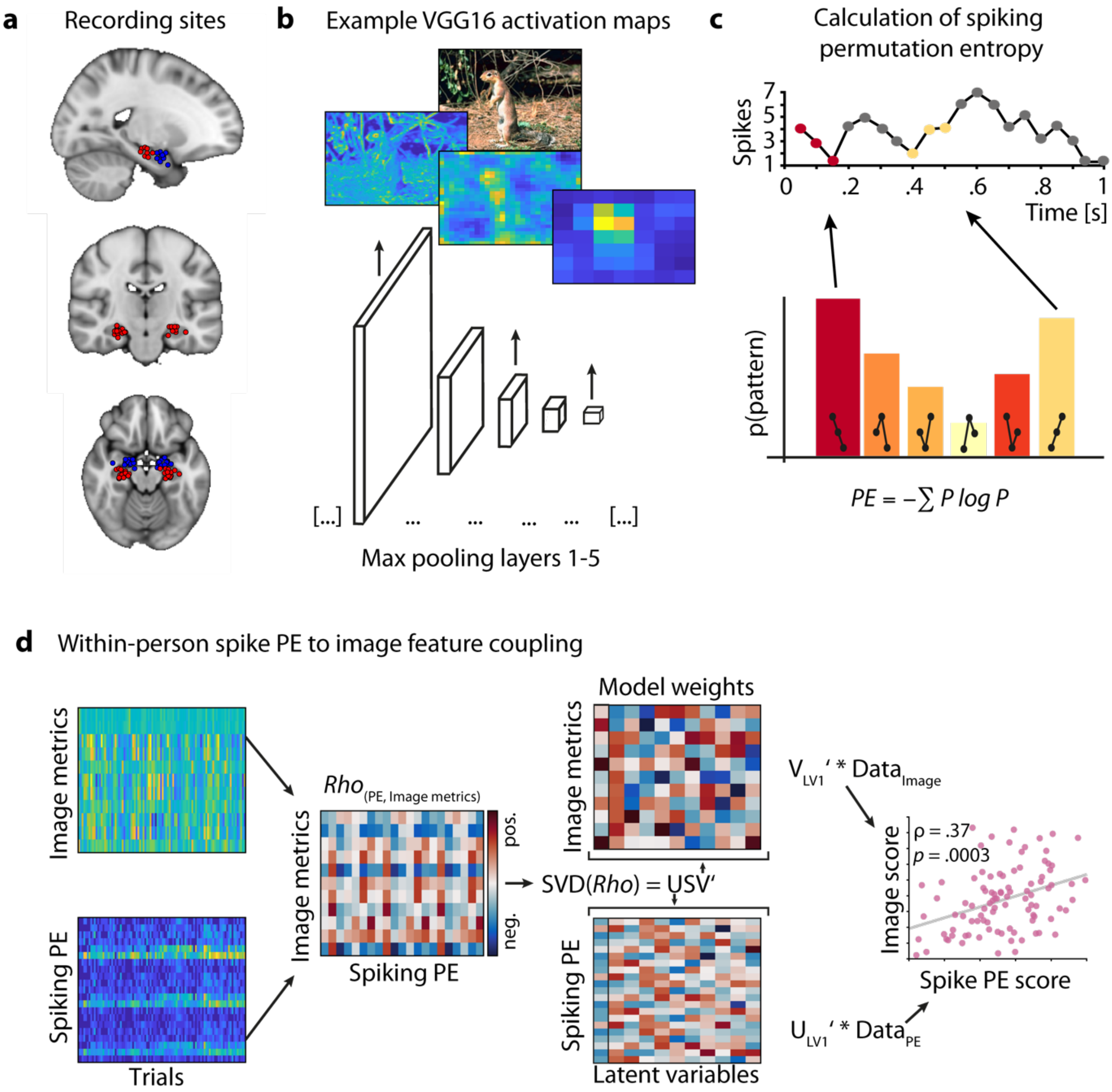

Fig. 1: Estimating latent coupling between image features and hippocampal spike entropy. A: Recording sites of depth electrodes for all participants with available probe coordinates. Hippocampus sites in red, amygdala sites in blue (top: x = –21, middle: y = –19, bottom: z = –17). B: VGG16 (trained on Imagenet) was used to predict activation maps at five layers of varying depth (max pooling layers 1–5) for images previously shown to participants at encoding, resulting in feature-wise activation maps. The mean across layer-wise features is shown for two example images and max pooling layers 1, 3, and 5. We extracted three summary metrics per layer and feature (sum, standard deviation, number of non-zero elements) before subjecting each layer-wise summary matrix (# images * # features) to a principal component analysis (PCA). In all further analyses, we relied on the first component score of each image, layer, and summary metric. C: Spike entropy was calculated per neuron and trial based on the first second of image encoding (for all neurons with PE > .0001 and trials with > 1/3 of neurons spiking). In brief, permutation entropy works by transforming signals into patterns (here: length = 3) and counting these patterns before calculating the Shannon entropy of the pattern distribution. D: Within-person correlations were computed by decomposing the rank-correlation matrix of trial-wise spike PE (per neuron) and image feature metrics using partial least squares (PLS). Singular value decomposition (SVD) of the rank-correlation matrix results in neural and stimulus weights per latent variable (LV). The weights of the first LV (first column outlined in black) were used to reduce the dimensionality of neural and feature matrices into scores for each trial. The rank correlation between both weighted variables represents the latent estimate of across-trial coupling between image features and hippocampus spike PE (right-most panel). Panels in D show data from a single subject in our sample.

Here, we analyse single-neuron hippocampal recordings from 34 human patients (Fig 1A) during a visual encoding and recognition memory task. All neural analysis was done on the single subject level based on simultaneously recorded neurons (12±11 hippocampal neurons per individual and session, total N = 411). We only included high-quality, well isolated units that satisfied all spike sorting quality metrics (*14*). To estimate simple and composite visual features of the images participants saw at encoding, we employed two computational vision models, HMAX and VGG16 (Fig 1B) (*15*, *16*). We measured the variability (entropy) of single-neuron hippocampal activity (Fig 1C) on every trial during encoding (*17*). We then estimate the relationship between different image features and hippocampal spike variability using within-participant latent modelling (*18*) (Fig 1D). Finally, we tested whether stronger coupling between visual features and hippocampus spike modulation yields better memory performance, and examine whether simple or composite visual features are most crucial in this context.

First, we estimated individual trial-level coupling between encoding spike entropy and layer-wise image feature metrics via partial least squares (PLS) (see Fig 1D and Methods for details). Image feature metrics consisted of principal components capturing the spatial sum, standard deviation (SD), and number of non-zero entries per activation map and layer (based on HMAX and VGG16; see Methods for details). Fig 2A depicts within- and across-layer stimulus weights for each subject, highlighting wide individual differences in the relative importance of image features in coupling to hippocampal spike variability. We also revealed significant individual correlations of hippocampal spike entropy for early as well as late-layer features in individual subjects (Fig 2B). For detailed patterns of model-wise feature weights, please see Supplements (Fig S2). Individual late-layer coupling estimates were significantly

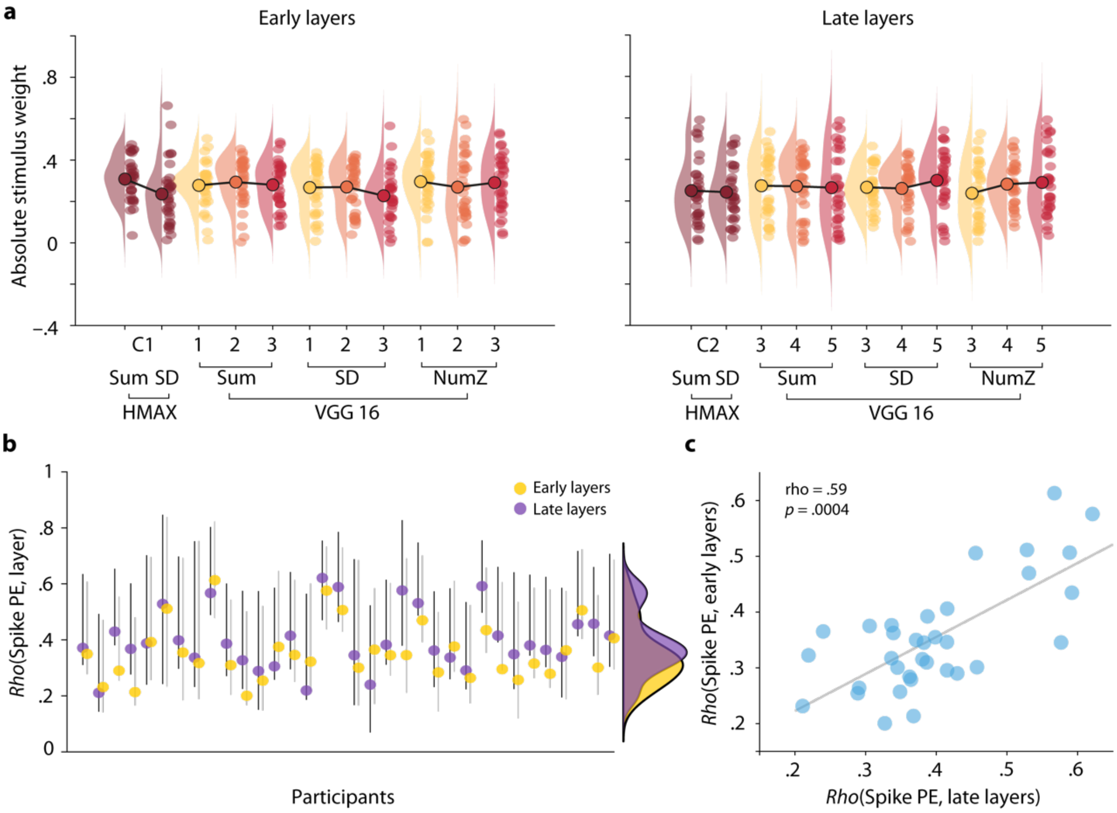

Fig. 2: Coupling of hippocampus spike entropy to image features. A: Feature- and layer wise absolute stimulus weights for early layer (left) and late layer models (right), both based on the same hippocampal spiking data. Raincloud plots (*38*) contain participant wise estimates (single dots), densities, and grand averages (circled big dots, connected by black lines). B: Absolute latent correlation estimates (Spearman) including 95% bootstrapped confidence intervals resulting from individual PLS models estimating hippocampal spike PE coupling to late (purple) and early-layer image features (yellow). Each dot represents one participant. The stronger coupling of hippocampal spike PE to late-layer than early-layer image features (z = 2.9, *p* = .004) illustrates the relative preference of hippocampal neurons for the representation of composite features. C: Positive between-subject correlation between hippocampal spike PE coupling to late and early-layer visual features, illustrating stable individual coupling to image features overall. Dots represent participants.

higher than early-layer coupling estimates (Wilcoxon z_{33} = 2.9, p = .004). Additionally, both were substantially correlated (rho = .59, p = .0004; Fig 2C). Hence, the entropy of trial-wise hippocampus spike entropy during encoding was coupled to simple as well as composite features of images presented during encoding, with stronger coupling to more composite (late-layer) features.

We then tested the relevance of individual spike-to-image feature coupling at encoding to later recognition memory performance (where performance = principal component score capturing various measures of accuracy (mean = .73), dprime (mean = 1.4), and confidence (mean = 2.5 out of 3; see Fig 3 and Methods)). We also contrasted the predictive power of hippocampal spike PE-to-image

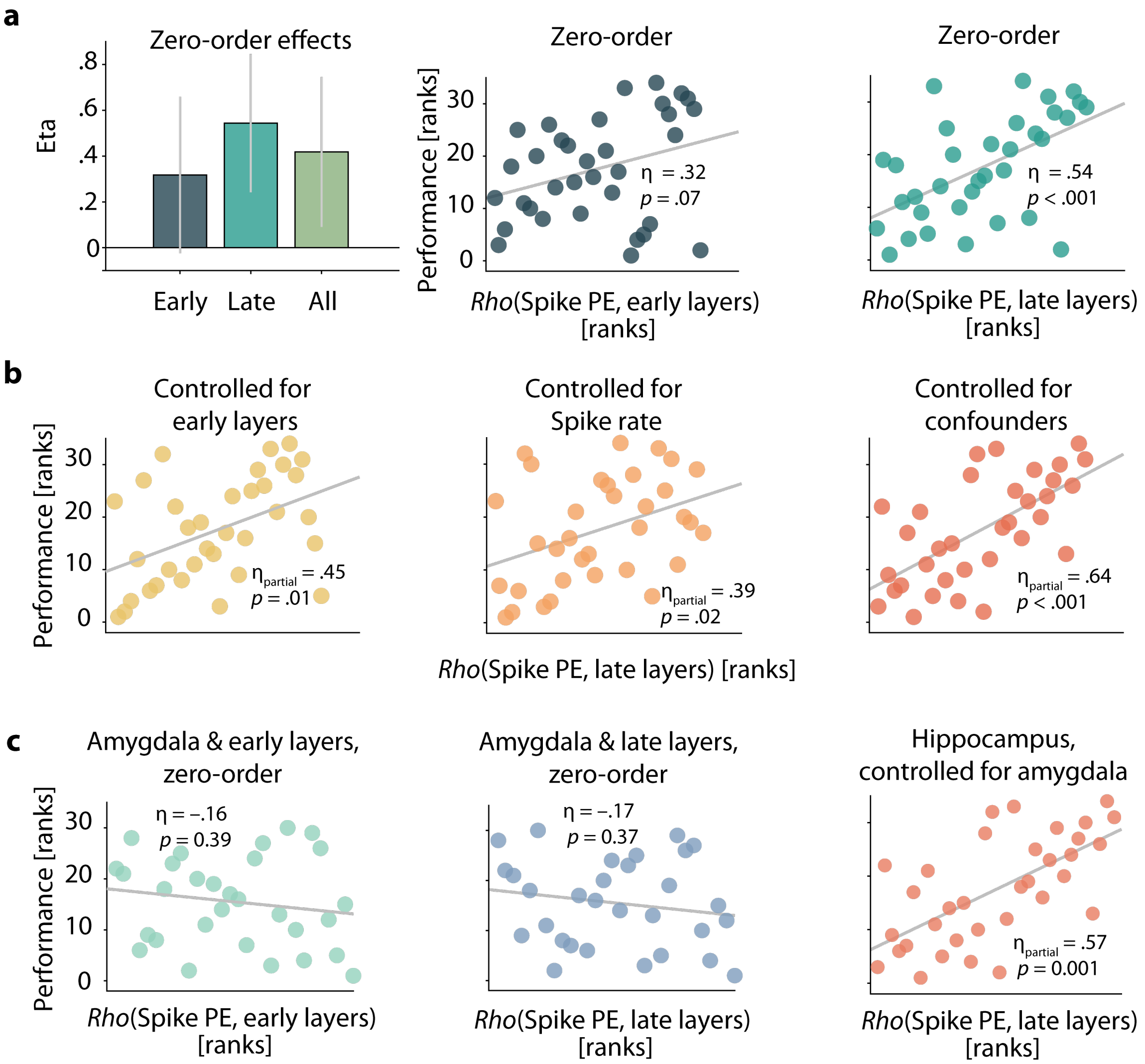

Fig. 3: Coupling of hippocampus spike entropy to late-layer image features specifically predicts memory performance. A: Zero-order eta estimates for early-layer, late-layer, and all-layer coupling vs. performance (left, bar graphs including bootstrapped 95% CI). Recognition performance was captured by a principal component score that combined accuracy, dprime, and confidence. Note that criterion was not related to spike PE coupling (Fig S3), represents bias rather than performance, is very weakly correlated with all performance measures (see Methods), and was hence omitted in this case. Zero-order relationship between hippocampal spike coupling estimates (individual latent correlations) and recognition performance (principal component capturing accuracy, dprime, and confidence) for early-layer models (middle) and late layer models (right). All correlations are non-parametric (Spearman). B: Unique links of hippocampal spike coupling to late-layer features after controlling for coupling to early-layer features (left), coupling of spike rate to late-layer features (middle), and a set of other potential between-subject confounds (number of neurons and trials, task variant, encoding duration, age; right). C: Effects are specific to the hippocampus. Coupling of amygdala spike entropy to either early-layer (left) nor late-layer features (middle) did not predict recognition performance. Finally, controlling for amygdala spike PE coupling to late-layer features did not reduce the link of hippocampus late-layer coupling to performance (right).

feature coupling using early, late, or all computational vision model layers. This way, we sought to unfold their relative importance for the translation of visual features into reliable memory traces during encoding. Later memory performance was positively correlated with the coupling of hippocampal spike variability during encoding to image features for early-layers (Fig 3a, η = .32, t_{31} = 1.9, p = .07) and all layers (η= .42, t_{31} = 2.6, p = .01), but most strongly so for late-layers (η = .54, t_{31} = 3.7, p < .001). See also Fig S3 for behavioural variable-specific distributions and model results. Importantly, the relationship for late layers remained after controlling for the coupling of hippocampal spike variability to early layers (Fig 3b, η_{partial} = .45, t_{31} = 2.8, *p* = .01), and all layers (η_{partial} = .38, t_{31} = 2.3, *p* = .03). In reverse, early-layer coupling (η_{partial} = –.02, t_{31} = .1, *p* = .90) and all-layer coupling (η_{partial} = –.11, t_{31} = .6, *p* = .6) did not predict memory performance once controlled for late-layer coupling. Furthermore, coupling estimates from a latent model that linked trial-wise spike rates (instead of spike PE) to late-layer image features also could not account for the link between spike PE to late-layer coupling and memory performance (η_{partial} = .39, t_{30} = 2.4, *p* = .02), nor could an additional set of potential confounds (number of trials and neurons, task variant, duration of encoding, age; η_{partial} = .64, t_{28} = 4.4, *p* < .001). In sum, individuals who displayed a tighter coupling between hippocampal spiking variability and composite visual features during encoding later exhibited better memory for those encoded images. The dominance of this late-layer specific relationship (relative to early-layer features and other controls) suggests that the hippocampus may indeed encode more composite stimulus features as a central predeterminant of successful memory formation.

Finally, given the potential for memory- and visually-sensitive neurons in the amygdala (*14*, *19*), we repeated all analyses above for neurons recorded in the amygdala within the same group of patients (17±10 amygdala neurons per individual and session, total N = 507). Unlike for hippocampal neurons, the coupling of amygdala spike entropy to visual features was not significantly predictive of memory performance (Fig 3c; all *ps* > .4). Also, controlling for amygdala coupling had minimal impact on the coupling between hippocampal PE and late-layer image features (Fig 3c, η_{partial} = .57, t_{31} = 3.6, *p* = .001). These results speak to the anatomical specificity of a coupling between visual features and neural dynamics that might trace back to the extraction and conjunction of composite information achieved by a diverse set of neurons in the hippocampus specifically.

Collectively, these results represent first evidence that intra-individual coupling between hippocampal spiking variability and image features during encoding is crucial for the successful formation of memories (*19*). Importantly, within individuals, hippocampal spike entropy was coupled more strongly to composite than to simple sensory features (Fig 2b), and this late-layer hippocampal spike coupling dominantly predicted memory performance up to 30 minutes later (Fig 3). These results not only support a ventral representational hierarchy that increases in feature aggregation and composition from visual cortex to hippocampus, but also provide evidence for the intra- and inter-individual behavioural relevance of such hierarchical neural processing (*20*). Our findings of more limited (but still present) behavioural relevance of early-layer features suggest a hierarchy in which the hippocampus downweighs simple features without discarding them entirely (*2*). Based on composite features, the hippocampus might be able to generate conjunct information by combining sensory, object, and relational aspects into a rich and generalizable memory trace (*2*, *5*). The absence of memory-relevant spike variability coupling in the amygdala, despite this structure's known role in memory formation (*19*) and direct afferents from visual cortical areas (*21*), highlights the unique role of the hippocampus as a dynamic conjunction hub of more aggregated visual input during memory formation. Additionally, that trial-level mapping between hippocampal spiking variability and visual content predicted memory formation success over and above standard spike rate (*22*) further buttresses a growing literature revealing the unique behavioural relevance of moment-to-moment fluctuations in brain activity (*9*). Indeed, we and others have argued that control processes may flexibly adapt neural variability (so-called "meta-variability") to meet the resource demands of a given task, thereby enabling optimal behaviour (*9*, *12*, *23*); here, we show that visual feature-driven meta-variability is required for memory success. Although promising, the very idea of meta-variability requires new theories and tools that elucidate the behavioural relevance of within-trial temporal neural variability beyond typically used measures in neuroscience (e.g., the across-trial Fano factor) (*9*).

Using our freely open and available methodological framework (see Methods), future research could test alternative models of "conjunctive" representations in hippocampus in a within-participant, across-trial manner. For example, one could test the presence of object category representations (*24*), or of any map-like representation spanning space (*25*), direction (*26*), or non-spatial relational maps (*27*) within and beyond the hippocampus or MTL (e.g., prefrontal cortex). Importantly, our approach permits the estimation of any joint space between neural activity on the one side and multivariate stimulus properties of any kind on the other, *for each subject*. Doing so allows the optimal expression of individual response profiles that can subsequently be compared across subjects in any desired context, regardless of recording specifics (e.g., exact cells, locations). Crucially, by decomposing this shared space between neural activity and stimulus features, one estimates a low dimensional representation of how neural responses represent stimulus properties of interest, an approach that is immediately complementary to recent large-scale efforts to summarize neural activity alone using dimensionality reduction techniques (*28*, *29*).

Overall, we propose that moment-to-moment spiking variability provides a novel window into how the hippocampus constructs memories from the building blocks of our visual world.

SUPPLEMENTS

METHODS

Sample and electrophysiology

We re-analyzed human hippocampus and amygdala single neuron activity from a previously published dataset (*14*) of 42 patients (total number of sessions = 65) undergoing surgery for intractable epilepsy who performed an encoding and recognition memory task (see below). Electrodes were localized based on post-operative MRI images and locations were only chosen according to clinical criteria. Protocols were approved by the institutional review boards of the Cedars-Sinai Medical Center, Huntington Memorial Hospital and the California Institute of Technology. We analyzed the same single units that were isolated using spike sorting for an earlier release of this data set (*14*), and we focused on spikes fired within the first 1000 ms of stimulus presentation during the encoding phase of the task.

Task

Patients were first presented with images from five out of 10 possible categories (across task variants) during an encoding phase (1-2 sec; 100 trials) and performed an animacy judgement (animal vs. not; unlimited time to respond) on these images. After a 15–30-minute delay, they were presented with a set of images that contained both previously seen and novel images (50% each; 100 trials) and were asked to simultaneously judge images as old or new and provide a confidence rating (from 1-new/confident to 6 – old/confident). While further details on task, recordings, and basic performance can be found elsewhere (*14*, *30*), it is important to note that we limited our analyses of neural activity to the first encoding session of n = 34 patients whose recordings included active hippocampus neurons (average neuron PE > .0001; 12±11 hippocampal neurons per individual and session, total N = 411). Memory performance was quantified using across-trial behavioural data from the corresponding recognition session, for which we focused on recognition accuracy, dprime, confidence, and confidence-weighted accuracy, while additionally including response criterion. We analyzed absolute confidence by collapsing across old and new decisions, resulting in confidence values of 1–3 (low to high confidence) that were averaged within participants and across trials. Of note, all primary results are based on a principal component score of performance which was generated via a PCA on all metrics but response criterion (eigenvalue = 2.8, standardized loadings = .89, .92, .92, .53] for accuracy, confidence-weighted accuracy, dprime, and confidence). Note that we did not include criterion in the PCA estimation because it represents response bias rather than performance and is weakly correlated with all other performance measures (avg_{corr} = .064, ranging from –.02-.13). Separate analyses for each performance metric can also be found in the supplemental material.

Using computational vision models to estimate the content of stimuli participants were asked to encode

With the goal of estimating image features at different levels of aggregation, from simple, orientation-like features to more complex, composite features, we employed two different computational vision models, HMAX (*15*) and VGG16 (*16*). Both models are openly available and have previously been used to estimate image content at different aggregation levels (*12*, *31*).

HMAX

The HMAX model is a biologically-inspired, feedforward model of the ventral visual stream that contains four hierarchical layers, S1, C1, S2, and C2 (*15*). S1 and C1 layers correspond to visuo-cortical areas V1/V2, whereas S2 and C2 correspond to V2/V4 (*32*). Within the first layer (S1) each unit is modeled with a different Gabor filter. These filters vary with respect to their orientation (HMAX defaults: -45°, 0°, 45°, 90°) and their size, the n x n pixel neighborhood over which the filter is applied (sizes: [7:2:37]). The resulting activation map of the S1 layer contains the simple cell responses for every position within the input image. Next, each C1 unit receives the result of a maximization across a pool of simple S1 units with the same preferred orientation but with (a) varying filter sizes and (b) at different positions (spatial pooling). We used 16 filter sizes at the first layer and maximized only across adjacent filter sizes, resulting in 8 "scale bands". In the following, S2 units merge inputs across C1 units within the same neighborhood and scale band, but across all four orientations. Importantly the response of S2 units is calculated as the fit between input and a stored prototype. At the final layer (C2), a global maximum across positions and scales for each prototype is taken, fitting eight C1 neighborhoods [2:2:16] using 400 different prototype features (*32*). For all images seen by participants during encoding, we extracted estimates for C1 and C2 layers for further analysis of within-subject coupling between image features and spiking variability (see below).

VGG16

Additionally, we processed images using VGG16, one of the most commonly used convolutional neural networks of computer vision, characterized by its high number of convolutional layers and its very high accuracy in object classification (*16*). Here, each input image is processed by a stack of 13 convolutional layers, with stride and spatial padding of one pixel and a receptive field of 3×3 pixels. The number of features per convolutional layer gradually increases from early to late-layers ([64, 128, 256, 512]). Convolutional layers are interleaved with five max-pooling layers that carry out spatial pooling (*16*). The stack of convolutional layers is followed by three fully connected layers and one soft-max layer. We used VGG16 as implemented in TensorFlow, pre-trained on the image-net dataset (*33*, *34*). For each image that participants encoded during the experiment, we extracted predicted activation (heat) maps for max-pooling layers 1-5 (corresponding to layers 3, 6, 10, 14, & 18) for further analysis of within-subject coupling between image features and spiking variability (see below).

Extraction of image features from model layer activation maps

We estimated the image features by extracting three layer- and image-wise feature metrics: the spatial sum, spatial standard deviation (SD), of C1 & C2 layers (HMAX) and max-pooling layer 1-5 (VGG16) as well as the number of zero elements ("pixels") for VGG16 max-pooling layer 1-5. This was done to arrive at a comprehensive approximation of image features that incorporates overall saliency (spatial sum), the distribution of salient and non-salient image locations (spatial SD), and the sparsity of saliency maps (number of non-zero entries). Additionally, note that the spatial sum and SD were only computed across non-zero map entries. As the number of features varies across the layers of HMAX and VGG16 models, we extracted the first principal component for each feature metric and layer using layer-wise PCAs across all images. Thus, each image participants saw at encoding was represented by three principal component scores for each model layer of interest, one each capturing the layer-wise spatial sum, standard deviation, and number of non-zero entries. These scores subsequently served as input for individual PLS models to estimate the coupling between image features and spiking variability (see below).

Estimation of spiking variability (permutation entropy)

For each single unit and trial during the encoding phase of the memory task, we extracted the first 1000ms after stimulus onset in non-overlapping bins of 10ms length and extracted the bin-wise spike counts. Based on the resulting spike trains, we then calculated permutation entropy (PE) for each neuron and trial (*35*) to measure the temporal variability of neuronal responses during encoding. Note that permutation entropy is tailored for analyses of this kind as it does not come with distributional assumptions and has been designed with physiological data in mind. We applied PE instead of more commonly used estimates of time series variability (e.g., standard deviation, Fano factor) due to the special distributional properties of single-trial spiking data that often violate normality assumptions (due e.g., to extreme sparsity).

To calculate permutation entropy, a timeseries is first partitioned in overlapping sections of length m (*17*). The data in each section is then transformed into ordinal rankings, so that every section is represented by a unique pattern. For example, the sequence (2,11,14) corresponds to the pattern (0,1,2), whereas the sequence (15,19,1) maps to (1,2,0). Thereafter we can count the relative frequency $\mathrm{p_i}$ of all patterns and compute PE as:

$$\mathrm{PE_m} = -\sum_{\mathrm{i=1}}^{\mathrm{m!}} \mathrm{p_i} \log_2 \mathrm{p_i}$$

where m corresponds to the length of sections and $m!$ describes the number of possible patterns. We computed PE for three different motif lengths ([2,3,4]). Neuron- and trial-wise PE estimates of all three motif lengths were used within individual partial least squares (PLS) models to estimate the individual coupling of image features to spike entropy (see below).

Estimating the within-person coupling of image features and spiking entropy

To quantify the individual multivariate relation between spiking entropy and image features of the presented images, we employed a behavioral partial least squares (PLS) analysis for each subject (*18*, *36*).

Here, PLS first calculated the rank correlation matrix (Rho) between the trial-wise estimates of stimulus features (e.g., $\mathrm{C2_{sum}}$, $\mathrm{C2_{SD}}$, $\mathrm{VGG_{sum}}$, $\mathrm{VGG_{SD}}$, $\mathrm{VGG_{nz}}$ for layers 3-5) and the trial-wise PE estimates of each recorded neuron, within-person (Fig 1A). All neurons included had PE > .0001 and all trials included contained at least a third of neurons spiking at least once (different cut-offs left results qualitatively unchanged). The Rho matrix was subsequently decomposed using singular value decomposition (SVD), generating a matrix of left singular vectors of image feature weights (U), a matrix of right singular vectors of neuron weights (V), and a diagonal matrix of singular values (S).

$$\mathrm{SVD_{Rho}} = \mathrm{USV'}$$

The application of these weights yields orthogonal latent variables (LVs) which embody the maximal relation between feature content of the input and neural spiking entropy. The latent correlation of each LV is calculated by first applying neural weights to neuron-wise PE data and stimulus feature weights to the matrix of stimulus feature metrics, respectively, before correlating the resulting latent scores (Fig 1D). Bootstrapping with replacement was used to estimate confidence intervals of observed latent correlations (1000 bootstraps). Importantly, given the variable and small number of trials and neurons across individuals, non-parametric Spearman correlations were used within PLS and throughout all other analyses.

To test the differential coupling of spike PE to various image features at different levels of image feature aggregation, we obtained individual coupling estimates for “early-layers” of computational vision models ($\mathrm{C1_{sum}}$ and $\mathrm{C1_{SD}}$ from HMAX; $\mathrm{VGG_{sum}}$, $\mathrm{VGG_{SD}}$, and $\mathrm{VGG_{nz}}$ for layers 1–3), “late-layers” ($\mathrm{C2_{sum}}$ and $\mathrm{C2_{SD}}$ from HMAX; $\mathrm{VGG_{sum}}$, $\mathrm{VGG_{SD}}$, and $\mathrm{VGG_{nz}}$ for layers 3–5), and all layers ($\mathrm{C1\&C2_{sum}}$ and $\mathrm{C1\&C2_{SD}}$ from HMAX; $\mathrm{VGG_{sum}}$, $\mathrm{VGG_{SD}}$, and $\mathrm{VGG_{nz}}$ for layers 1–5). This split of layers was performed to keep the number of features within each latent model constant while at the same time separately estimating the coupling of spike variability to early layer, late layer and all image features. To quantify whether spike PE captures unique, behaviourally relevant aspects image feature coupling compared to what may be captured by spike rate, we also ran the same PLS models, but replacing spike PE with trial-wise spike rates. Additionally, to explore the topological specificity of memory-relevant spike-feature coupling, we computed separate PLS models for neurons recorded in hippocampus and amygdala, respectively.

Statistical analyses

We compared the individual strength of spike PE coupling to early and late-layers via a Wilcoxon signed rank test on the absolute latent correlations derived from PLS models based on early and late-layers, respectively.

We then used linear models to regress the performance score (see above) onto individual latent estimates of coupling between spike entropy and image feature metrics (all in rank space). First, we ran zero-order models based on the spike PE coupling estimates of early, late, and all layer PLS models, respectively. Next, we tested the unique explanatory power of late-layer coupling by separately controlling for early and all layer coupling estimates (and vice versa; see Fig 3). To additionally contrast the behavioural relevance of spike PE coupling with more established metrics of single cell activity, we controlled effects of late-layer spike PE coupling for late-layer spike rate coupling. Finally, we controlled effects of late-layer hippocampal spike PE coupling for a set of inter-individual control variables (number of trials, number of neurons used within analysis, task variant, encoding duration, age). For each model we computed estimates of partial eta, marking the unique portion of variance in performance explained by the relation modulation of hippocampal spike PE.

To probe the topological specificity of our findings, we modelled performance as a function of amygdala spike PE coupling for late-layers and additionally controlled the effects of late-layer hippocampus coupling for amygdala effects.

All statistical analyses, PE and HMAX estimation were run in MATLAB 2020a, VGG16 was run in python 3.0.

Data and code availability

Analysed data have been published previously and can be downloaded (https://europepmc.org/article/pmc/pmc5810422). Code to reproduce all main results will be made accessible on Github upon publication.

Acknowledgements

We thank Morgan Barense, Ulrich Mayr, and Markus Werkle-Bergner for fruitful discussions on earlier versions of this work.

Funding sources

LW, FK, and DDG were partially funded by an Emmy Noether Program grant from the German Research Foundation (to DDG) and by the Max Planck UCL Centre for Computational Psychiatry and Ageing Research, Berlin. UR was partially supported by NIH NINDS U01NS117839.

Author contributions

In line with the CRediT framework (*37*), author contributions are listed as follows:

	LW	FK	EvdE	SK	UL	UR	DDG
Conceptualization	■	□	□	□	□	□	■
Methodology	■	□	□	■	□	□	■
Software	■	■	□	■	□	□	■
Validation	■	■	□	□	□	□	■
Formal analysis	■	□	□	□	□	□	■
Investigation	■	□	□	□	□	■	■
Resources	■	□	□	□	■	□	■
Data Curation	■	■	□	□	□	■	□
Writing - Original Draft	■	■	□	□	□	□	■
Writing - Review & Editing	■	■	■	□	■	■	■
Visualization	■	■	□	□	□	□	□
Supervision	□	□	□	□	□	■	■
Project administration	■	□	□	□	□	□	■
Funding acquisition	□	□	□	□	■	■	■

REFERENCES

1. E. Kobatake, K. Tanaka, Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *J Neurophysiol*. 71, 856–867 (1994).
2. B. A. Kent, M. Hvoslef-Eide, L. M. Saksida, T. J. Bussey, The representational–hierarchical view of pattern separation: Not just hippocampus, not just space, not just memory? *Neurobiol Learn Mem*. 129, 99–106 (2016).
3. D. J. Kravitz, K. S. Saleem, C. I. Baker, L. G. Ungerleider, M. Mishkin, The ventral visual pathway: an expanded neural framework for the processing of object quality. *Trends Cogn Sci*. 17, 26–49 (2013).
4. J. R. Manns, H. Eichenbaum, Evolution of declarative memory. *Hippocampus*. 16, 795–808 (2006).
5. T. E. J. Behrens, T. H. Muller, J. C. R. Whittington, S. Mark, A. B. Baram, K. L. Stachenfeld, Z. Kurth-Nelson, What Is a Cognitive Map? Organizing Knowledge for Flexible Behavior. *Neuron*. 100, 490–509 (2018).
6. A. P. Yonelinas, The hippocampus supports high-resolution binding in the service of perception, working memory and long-term memory. *Behav Brain Res*. 254, 34–44 (2013).
7. M. Moscovitch, R. Cabeza, G. Winocur, L. Nadel, Episodic Memory and Beyond: The Hippocampus and Neocortex in Transformation. *Annu Rev Psychol*. 67, 105–134 (2016).
8. A. C. H. Lee, L.-K. Yeung, M. D. Barense, The hippocampus and visual perception. *Front Hum Neurosci*. 6, 91 (2012).
9. L. Waschke, N. A. Kloosterman, J. Obleser, D. D. Garrett, Behavior needs neural variability. *Neuron*. 109, 751–766 (2021).
10. G. Orbán, P. Berkes, J. Fiser, M. Lengyel, Neural Variability and Sampling-Based Probabilistic Representations in the Visual Cortex. *Neuron*. 92, 530–543 (2016).
11. D. Festa, A. Aschner, A. Davila, A. Kohn, R. Coen-Cagli, Neuronal variability reflects probabilistic inference tuned to natural image statistics. *Nat Commun*. 12, 3635 (2021).
12. D. D. Garrett, S. Epp, M. Kleemeyer, U. Lindenberger, T. A. Polk, Higher performers upregulate brain signal variability in response to more feature-rich visual input. *Neuroimage*. 217, 116836 (2020).
13. A. M. Hermundstad, J. J. Briguglio, M. M. Conte, J. D. Victor, V. Balasubramanian, G. Tkačik, Variance predicts salience in central sensory processing. *Elife*. 3, e03722 (2014).
14. M. C. M. Faraut, A. A. Carlson, S. Sullivan, O. Tudusciuc, I. Ross, C. M. Reed, J. M. Chung, A. N. Mamelak, U. Rutishauser, Dataset of human medial temporal lobe single neuron activity during declarative memory encoding and recognition. *Scientific Data*. 5, 180010 (2018).
15. M. Riesenhuber, T. Poggio, Hierarchical models of object recognition in cortex. *Nat Neurosci*. 2, 1019–1025 (1999).
16. K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition. *Arxiv* (2014), doi:10.48550/arxiv.1409.1556.
17. C. Bandt, B. Pompe, Permutation Entropy: A Natural Complexity Measure for Time Series. *Phys Rev Lett*. 88, 174102 (2002).
18. A. R. McIntosh, F. L. Bookstein, J. V. Haxby, C. L. Grady, Spatial Pattern Analysis of Functional Brain Images Using Partial Least Squares. *Neuroimage*. 3, 143–157 (1996).
19. U. Rutishauser, O. Tudusciuc, D. Neumann, A. N. Mamelak, A. C. Heller, I. B. Ross, L. Philpott, W. W. Sutherling, R. Adolphs, Single-Unit Responses Selective for Whole Faces in the Human Amygdala. *Curr Biol*. 21, 1654–1660 (2011).
20. R. A. Cowell, T. J. Bussey, L. M. Saksida, Components of recognition memory: Dissociable cognitive processes or just differences in representational complexity? *Hippocampus*. 20, 1245–1262 (2010).
21. J. L. Price, Comparative Aspects of Amygdala Connectivity. *Ann Ny Acad Sci*. 985, 50–58 (2003).
22. J. T. Wixted, S. D. Goldinger, L. R. Squire, J. R. Kuhn, M. H. Papesh, K. A. Smith, D. M. Treiman, P. N. Steinmetz, Coding of episodic memory in the human hippocampus. *Proc National Acad Sci*. 115, 1093–1098 (2018).
23. W. F. Młynarski, A. M. Hermundstad, Adaptive coding for dynamic sensory inference. *Elife*. 7, e32055 (2018).
24. K. Grill-Spector, The neural basis of object perception. *Curr Opin Neurobiol*. 13, 159–166 (2003).
25. J. O'Keefe, L. Nadel, *The hippocampus as a cognitive map* (Clarendon Press, Oxford, 1978).
26. A. Sarel, A. Finkelstein, L. Las, N. Ulanovsky, Vectorial representation of spatial goals in the hippocampus of bats. *Science*. 355, 176–180 (2017).

27. A. O. Constantinescu, J. X. O'Reilly, T. E. J. Behrens, Organizing conceptual knowledge in humans with a gridlike code. *Science*. 352, 1464–1468 (2016).
28. M. Dabagia, K. P. Kording, E. L. Dyer, Aligning latent representations of neural activity. *Nat Biomed Eng*, 1–7 (2022).
29. C. Stringer, M. Pachitariu, N. Steinmetz, M. Carandini, K. D. Harris, High-dimensional geometry of population responses in visual cortex. *Nature*. 571, 361–365 (2019).
30. U. Rutishauser, S. Ye, M. Koroma, O. Tudusciuc, I. B. Ross, J. M. Chung, A. N. Mamelak, Representation of retrieval confidence by single neurons in the human medial temporal lobe. *Nature Neuroscience*. 18, 1041–50 (2015).
31. S. W. Davis, B. R. Geib, E. A. Wing, W.-C. Wang, M. Hovhannisyan, Z. A. Monge, R. Cabeza, Visual and Semantic Representations Predict Subsequent Memory in Perceptual and Conceptual Memory Tests. *Cereb Cortex*. 31, 974–992 (2021).
32. T. Serre, A. Oliva, T. Poggio, A feedforward architecture accounts for rapid categorization. *Proc National Acad Sci*. 104, 6424–6429 (2007).
33. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: A large-scale hierarchical image database. *2009 Ieee Conf Comput Vis Pattern Recognit*, 248–255 (2009).
34. M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mane, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viegas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, X. Zheng, TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *Arxiv* (2016), doi:10.48550/arxiv.1603.04467.
35. M. W. Flood, B. Grimm, EntropyHub: An open-source toolkit for entropic time series analysis. *Plos One*. 16, e0259448 (2021).
36. A. Krishnan, L. J. Williams, A. R. McIntosh, H. Abdi, Partial Least Squares (PLS) methods for neuroimaging: A tutorial and review. *Neuroimage*. 56, 455–475 (2011).
37. L. Allen, A. O'Connell, V. Kiermer, How can we ensure visibility and diversity in research contributions? How the Contributor Role Taxonomy (CRediT) is helping the shift from authorship to contributorship. *Learn Publ*. 32, 71–74 (2019).
38. M. Allen, D. Poggiali, K. Whitaker, T. R. Marshall, J. van Langen, R. A. Kievit, Raincloud plots: a multi-platform tool for robust data visualization. *Wellcome Open Res*. 4, 63 (2021).

Supplemental figures

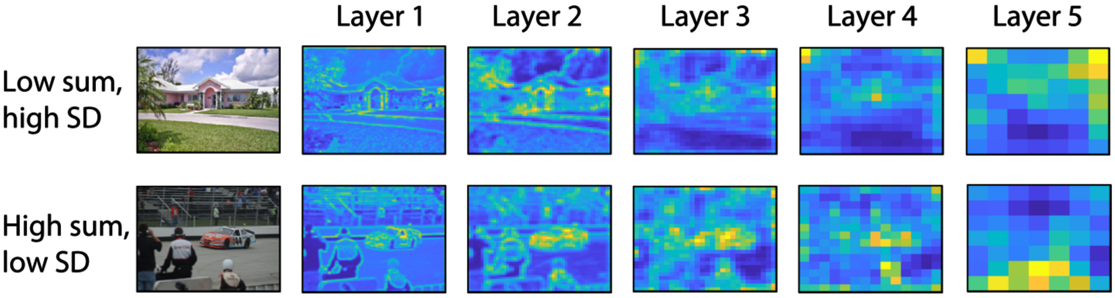

Figure S1. **Layer-wise average activation maps of two example images illustrating differences in feature metrics**. Both rows show average activation maps from all VGG16 max pooling layers (corresponding to layers 3, 6, 10, 14, & 18). The upper row contains maps of an image that comes with a relatively low spatial sum (across pixels per layer and feature) but a high SD, tracing back to the uneven distribution of salient spots across space (grass vs. house). The lower row contains maps of an image with relatively high sum but low SD (salient spots distributed relatively evenly).

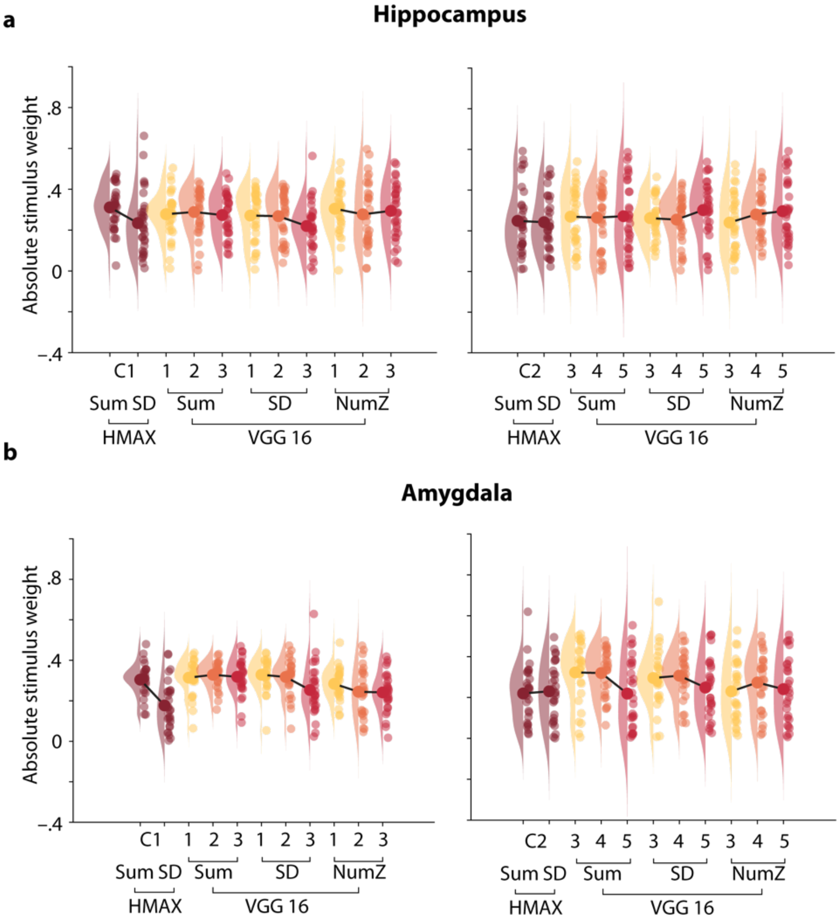

Figure S2. **Participant-, layer- and model wise patterns of absolute stimulus weights.** Absolute weights capture the relative importance of stimulus features for the observed latent correlation. (**a**) absolute weights for early (left) and late (right) layer models, grouped for Computational vision models (HMAX vs VGG16) and feature metrics (Sum, SD, number of non-zero entries NumZ). Dots represent absolute weights from individual PLS models, large non-transparent dots capture grand averages and are connected by black lines. (**b**) As in panel a, but for models based on amygdala neurons. Here, later layers seem to be given lower weights as compared to earlier layers.

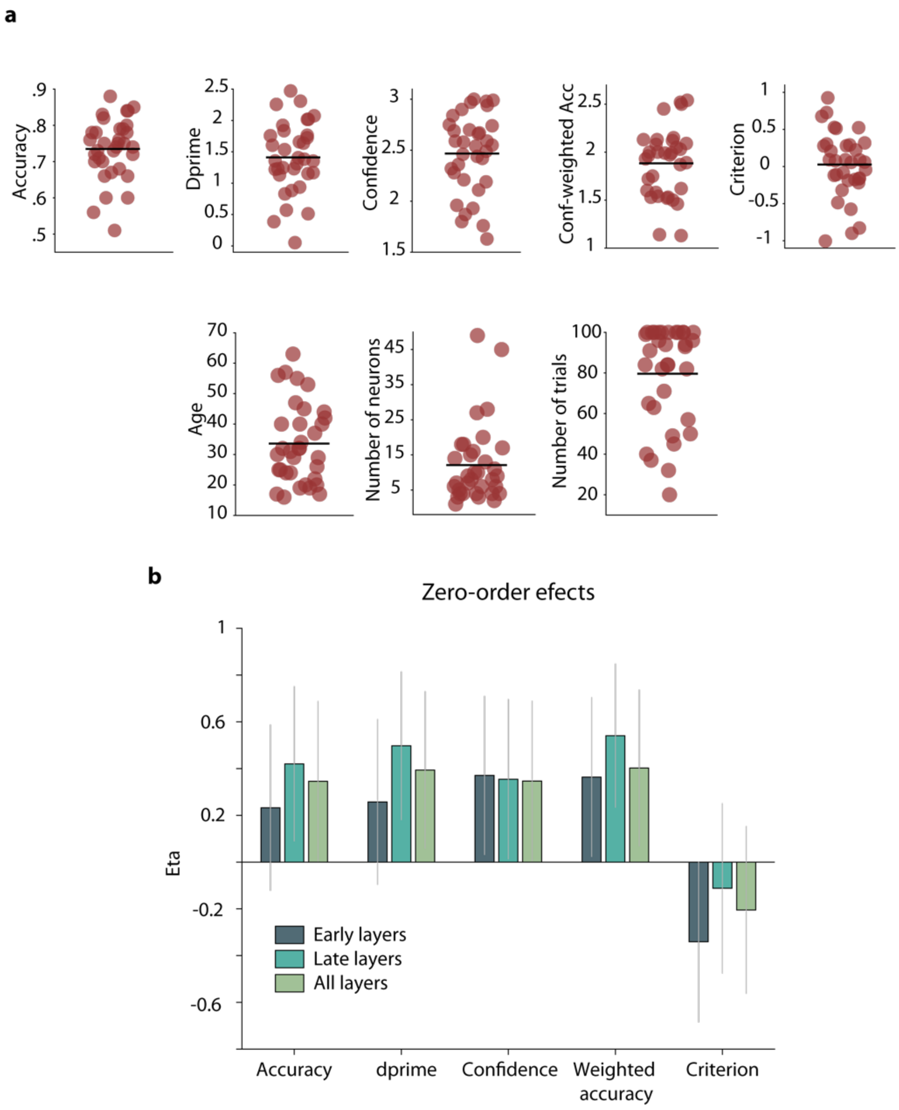

Figure S3. **Descriptive plots of inter-individual variables and their link to spike-image coupling.** (a) Scatter plots of all relevant behavioural variables, displaying across-participant averages (horizontal black line; each dot represents one participant). (b) Zero-order relationships between individual hippocampal spike PE to image feature coupling and behavioural metrics. Vertical lines depict bootstrapped 95% confidence intervals. Memory accuracy and dprime are positively linked to all coupling estimates, but only late and all layer coupling estimates are linked significantly. Confidence is linked positively to coupling estimates from early, late, and all layer models. Similarly, confidence-weighted accuracy is positively correlated with all layer-wise coupling estimates, strongest for coupling to late layers. Response criterion is not significantly linked to any of the three coupling metrics.