Maintenance and transformation of representational formats during working memory prioritization

Daniel Pacheco Estefan[1]*, Marie Christin Fellner[1]*, Lukas Kunz[2], Hui Zhang[1], Peter Reinacher[3,4], Charlotte Roy[5], Armin Brandt[5], Andreas Schulze-Bonhage[5], Linglin Yang[6], Shuang Wang[7], Jing Liu[8], Gui Xue[9], Nikolai Axmacher[1,9]

*Equal contribution

1 Department of Neuropsychology, Institute of Cognitive Neuroscience, Faculty of Psychology, Ruhr University Bochum, 44801 Bochum, Germany

2 Department of Biomedical Engineering, Columbia University, New York, NY, USA

3 Department of Stereotactic and Functional Neurosurgery, Medical Center – University of Freiburg, Faculty of Medicine, University of Freiburg, Freiburg, German

4 Fraunhofer Institute for Laser Technology, Aachen, Germany

5 Epilepsy Center, Medical Center – University of Freiburg, Faculty of Medicine, University of Freiburg, Freiburg, Germany

6 Department of Psychiatry, Second Affiliated Hospital, School of medicine, Zhejiang University, Hangzhou, China

7 Department of Neurology, Epilepsy center, Second Affiliated Hospital, School of medicine, Zhejiang University, Hangzhou, China

8 Department of Applied Social Sciences, The Hong Kong Polytechnic University, Hong Kong SAR

9 State Key Laboratory of Cognitive Neuroscience and Learning and IDG/McGovern Institute for Brain Research, Beijing Normal University, Beijing 100875, PR China

Abstract

Visual working memory depends on both material-specific brain areas in ventral visual stream (VVS) that support the maintenance of stimulus representations and on regions in prefrontal cortex (PFC) that control these representations. These two areas putatively support bottom-up information processing vs. top-down control functions that rely on oscillations in the gamma and beta frequency bands, respectively. Recent studies identified stimulus-specific working memory contents via representational similarity analysis (RSA) and analyzed their representational format using deep neural networks (DNNs) as models of the multi-layered hierarchy of information processing. How executive control prioritizes specific working memory contents and whether this affects their representational formats remains an open question, however. Here, we addressed this issue using a multi-item working memory task involving a retro-cue that prompted participants to either maintain one particular item or the entire item sequence. We exploited the excellent spatiotemporal resolution of intracranial EEG (iEEG) recordings in epilepsy patients and analyzed activity at electrodes in VVS (n=28 patients) and PFC (n=16 patients). During encoding, we identified category-specific information in both VVS and PFC. During maintenance, this information only re-occurred in VVS but not in the PFC – suggesting transformation of PFC representations from encoding to maintenance, putatively driven by the prioritization process. We thus used RSA in combination with two different DNN architectures to investigate the representational format of prioritized working memory contents. When we applied a feedforward DNN, we found that representational formats during maintenance could be well explained in VVS, but not in the PFC. We therefore employed a recurrent DNN architecture (BL-NET) to evaluate the representational formats in PFC during maintenance. We observed that representations in the deepest layer of this network matched the representational format of prioritized stimuli in the PFC. This effect was transient, locked to the presentation of the retro-cue, and specific to the beta (16-29Hz) frequency band. Together, these results suggest that PFC represents prioritized working memory information in an abstract representational format, which may dynamically control working memory representations through top-down mechanisms.

Introduction

Visual working memory (VWM) refers to the ability to store visual information for a short period of time and to flexibly manipulate it according to task demands. A critical aspect of VWM memory is prioritization, i.e., the ability to selectively allocate attention to particular aspects of the stored content depending on behavioral or cognitive requirements. Influential theories proposed that WM prioritization entails the transformation of representational patterns from a purely mnemonic to a task-optimized state (Chatham et al., 2014; Myers et al., 2017). A key prediction of these accounts is that working memory prioritization involves a transformation of representations in executive control areas according to task demands that can be disentangled from mnemonic (i.e., more stable) stimulus representations in sensory brain regions (Liebe et al., 2012; Miller et al., 2018; Myers et al., 2017). Here we set out to empirically test these predictions and analyzed the representational format of VWM stimuli using electrophysiological recordings in human epilepsy patients implanted with electrodes in ventral visual stream (VVS) and/or prefrontal cortex (PFC).

There is now abundant evidence on the neural correlates of VWM control processes in humans (Lepsien et al., 2011; Lepsien & Nobre, 2006; Nee & Jonides, 2008, 2009). Early studies focused on prioritization of to-be-encoded items, using paradigms in which participants were asked to selectively attend to particular items before these items were encoded (e.g., Griffin & Nobre, 2003; Posner, 1980; Schmidt et al., 2002; Vogel & Machizawa, 2004). More recent investigations employed retrospective cueing paradigms, in which visual items are prioritized after information has been stored in WM (Ester et al., 2018; Higo et al., 2011; Nee & Jonides, 2009; Nelissen et al., 2013; Sprague et al., 2016). These studies revealed that prefrontal and parietal regions that underlie the allocation of attention during perception are also engaged during the prioritization of items in VWM (Buschman & Kastner, 2015; Buschman & Miller, 2007; Lepsien et al., 2011; Lepsien & Nobre, 2006; Nee & Jonides, 2008; Nobre et al., 2004). In addition, this research has shown a crucial role of PFC in reformatting working memory content during the process of prioritization (Chatham et al., 2014). In particular, a recent meta-analysis noted that several prefrontal areas respond selectively to retro-cues but not to cues that indicate where to allocate attention before items are encoded in WM (Myers et al., 2017). In addition, the theory of ‘output gating’ suggests a specific role of the PFC for preparing information in a format that can be utilized by downstream regions to facilitate motor output (Sherfey et al., 2020; Wallis et al., 2015).

While previous studies in human have mostly been conducted using non-invasive methods with limited spatiotemporal resolution such as fMRI and M/EEG, recent studies in macaques used invasive recordings in order to investigate the electrophysiological mechanisms underlying the

manipulation of VWM content more directly (Lundqvist et al., 2018; Miller et al., 2018). These studies revealed a critical role of gamma-band activity (50-120Hz) in conveying bottom-up information during VWM encoding and of beta oscillations (20-35Hz) in providing top-down control over VWM contents (Lundqvist et al., 2018; Miller et al., 2018). A recent study on VWM prioritization in monkey PFC revealed that the allocation of internal attention in VWM involves a transformation of coding schemes (Panichello & Buschman, 2021), consistent with the framework proposed in (Myers et al., 2017) and recent computational models (Wan et al., 2022).

In humans, intracranial EEG (iEEG) recordings in epilepsy patients have been used to investigate the neurophysiological patterns underlying content-specific VWM representations. These studies employed multivariate analysis methods such as pattern classification and representational similarity analysis (RSA) in order to identify the representations of specific stimuli (Kriegeskorte & Diedrichsen, 2019). A recent iEEG study revealed the spectral and temporal dynamics of feedforward information processing during visual perception, highlighting a critical role of gamma-band activity (30-75Hz) in conveying bottom-up information from low level visual areas to higher-level regions (Kuzovkin et al., 2018). Similarly, iEEG research has shown that frequency-specific representations in the gamma, beta (13-29Hz) and theta (3-8Hz) frequency bands contain item- and category-specific information and are critical for episodic memory retrieval (Pacheco Estefan et al., 2019, 2021). In addition to identifying the relevant oscillatory frequencies that carry representational content during visual perception and episodic memory, recent human iEEG studies have investigated the formats of VWM representations. This research has employed deep neural networks (DNNs) to elucidate how different aspects of natural images (from low-level sensory features to higher-level contents and conceptual/semantic features) are represented in the brain during mnemonic processing. The idea is that mnemonic representations of natural images, similar to representations formed during perception, might require dedicated circuits specialized in the processing of different aspects or ‘formats’ of these representations. Indeed, some recent studies assessed the different representational formats of VWM encoding and maintenance via DNNs and showed that VWM representations undergo substantial transformations into a representational format that maps onto late layers of a convolutional neural network (Liu et al., 2020, 2021). While these methodological advancements have provided crucial insights into the neural mechanisms of VWM maintenance, no study so far has investigated the representational transformations driving VWM prioritization in humans. Thus, whether the prioritization of VWM representations involves a change in the representational format of the stored content and differential neural patterns for attended (i.e., task-relevant) vs. unattended items is currently unknown.

Here we capitalized on the heuristic potential of DNNs as models of visual representations, the flexibility of RSA and the high spatiotemporal resolution of iEEG to investigate this issue. We analyzed electrophysiological activity from VVS and PFC of epilepsy patients who performed a multi-item visual working memory paradigm involving a retro-cue. In our task, participants encoded a sequence of 3 natural images and were then prompted by a retro-cue to maintain either one specific item or all items (Figure 1A; Methods). Objects pertained to six categories with ten exemplars each (60 images in total). This experimental design allowed us to scrutinize how information about specific contents is represented in the brain during initial encoding and how it is transformed following the retro-cue, both in terms of their representational formats and regarding their underlying frequencies in VVS (441 electrodes) and PFC (147 electrodes; Figure 1B). We predicted that frequency-specific representations would characterize bottom-up storage and top-down information transfer, with a particular role of gamma and beta oscillations, respectively (Engel & Fries, 2010; Kuzovkin et al., 2018; Lundqvist et al., 2018; Miller et al., 2018; Vezoli et al., 2021). In particular, we hypothesized that oscillatory activity in the beta (13-29Hz) range may reflect representational transformations due to top-down control following the retro-cue (Miller et al., 2018).

Results

Participants performed the task with high accuracy (single item condition: proportion correct trials at category level: 0.859 ± 0.11, mean ± standard deviation; correct at item level: 0.80 ± 0.12; multi-item condition: correct at category level: 0.75 ± 0.17; correct at item level: 0.58 ± 0.19). Performance was significantly higher at the category level than at the item level, during both maintenance of single items ($t(31) = 3.82$, $p = 0.0005$) and of all items ($t(31) = 11.65$, $p = 7.3722e-13$). Thus, the participants understood the task correctly and remembered categorical information about the items in most trials.

Maintenance and transformation of category-specific representations

We investigated the electrophysiological patterns supporting the representation of category-specific information in VVS and PFC. We compared activity patterns during encoding (Encoding-Encoding Similarity; EES) and between encoding and maintenance (Encoding-Maintenance Similarity; EMS; Methods) (Figure 2A). As in previous studies (Liu et al., 2020; Pacheco Estefan

et al., 2021), representational patterns were constructed by concatenating multi-electrode epochs of time-frequency resolved power values (52 frequencies between 3 and 150 Hz) in windows of 500 milliseconds (five time points), overlapping by 400ms (Figure 2B). The similarity between feature vectors was calculated using Spearman's Rho. We tracked category-specific representations by comparing activity patterns of (different) items of the same category vs. items of different categories (Figure 2C).

The EES analysis revealed a marked presence of category-specific information during encoding in both VVS and PFC. In the VVS, category-specific information was found in all frequency bands (theta, 3-8Hz; alpha, 9-12Hz; beta, 13-29Hz; low-gamma, 30-75Hz; high-gamma, 75-150Hz; all $p < 0.005$; Figure 2D, top). In the PFC, an analogous analysis revealed category-specific information restricted to the theta, beta and low-gamma frequency bands (all $p < 0.01$; Figure 2D, bottom). We next analyzed the relative timing of effects in greater detail. For this purpose, we increased the temporal resolution by shortening the sliding windows to steps of 10ms and included all available frequencies in the 3-150Hz range as RSA features. This analysis revealed that categorical information reached the PFC at around 400ms after stimulus presentation, with a latency of ~350ms with respect to the VVS (Figure 2E). Moreover, EES was significantly higher in the VVS than the PFC starting ~150ms after stimulus onset ($p = 0.007$), suggesting that category-specific information is transferred from VVS to PFC during VWM encoding.

Next, we analyzed Encoding-Maintenance Similarity (EMS) to investigate whether category-specific representations from the encoding period reoccurred during the maintenance period (and particularly after the retro-cue). During maintenance of single items, we observed significant reoccurrence of category-specific information in all frequency bands (all $p = 0.015$) in the VVS. These effects were transient and occurred most prominently during the first 2s after the retro-cue (Figure 2F, top). By contrast, we did not observe any reoccurrence of category-specific information in the PFC in any frequency band (Figure 2F, bottom).

Together, these results show the formation of category-specific representations in both VVS and PFC during encoding – with differences in frequency contributions and timing – but reoccurrence of these representations during maintenance in VVS only.

Representational formats of category-specific representations

These findings indicate a robust and persistent maintenance of category-specific information in the VVS that was not observed in the PFC. This lack of an effect in the PFC may be due to a transformation of VWM representations driven by the prioritization process. Indeed, recent behavioral (Kerren et al., 2022), neuroimaging (Kwak & Curtis, 2022) and iEEG (Liu et al., 2020, 2021) studies established a crucial role of transformed representational formats for VWM maintenance – and more specifically, of abstract representational formats devoid of specific sensory information. We hypothesized that the PFC may maintain categorical information in an abstract representational format after presentation of the retro-cue. To evaluate this hypothesis, we employed two deep convolutional neural network (DNNs) architectures originally optimized to classify images: the Alexnet network (Krizhevsky et al., 2012) which has been widely employed in cognitive neuroscience to characterize neural representations of natural images during perceptual and mnemonic processes; and the recurrent convolutional network BL-NET, which has been previously applied to model human behavior (reaction times) in a perceptual task (Spoerer et al., 2020). Both the Alexnet and the BL-NET network represent stimuli in various formats ranging from low-level visual features in their superficial layers to more conceptual and abstract features in the deeper layers. While in Alexnet the processing of stimulus features takes place in a single feedforward pass, the BL-NET architecture includes lateral recurrent connections that generate a temporally evolving time-series of stimulus representations in each layer. This architecture recapitulates core features of recurrent dynamics during WM processing in the PFC and constitutes a unique opportunity to investigate dynamic representational transformations during VWM prioritization at different degrees of perceptual abstraction in the human brain (Kietzmann et al., 2019).

We first characterized stimulus representations in different layers of the Alexnet network and evaluated their correspondence with the representations extracted from the iEEG data in VVS and PFC. As expected, the Alexnet network represented stimulus features in a hierarchical fashion, with units in early layers responding to low-level visual features (edges, colors), and deeper layer representing more abstract stimulus features. We constructed representational similarity matrices (RSMs) from network responses by computing the similarities between all unit activations in each layer for all pairs of images using Spearman's Rho (Liu et al., 2020, 2021; Figure 3A, top). For visualization, we projected the data into two-dimensional space using Multidimensional Scaling (Methods, Figure 3A, bottom). Visual inspection of MDS projections suggested that representations of stimuli in Alexnet were clustered according to their categories

already in early layers. We computed the Category Cluster Index (CCI; McKee et al., 2014; Mehrer et al., 2020), defined as the difference in average distances of stimulus pairs from the same category vs. stimulus pairs from different categories (Figure 3B, top; Methods). CCI takes a value of 1 if stimuli from the same category are maximally clustered and approaches 0 if the representational geometry of stimuli shows no categorical organization. Using permutation statistics (i.e., label shuffling), we observed that CCI values were significantly higher than chance in all layers of the network (all $p = 0.008$, Bonferroni-corrected for the 8 layers), and it was highest in layer Fc7 (Figure 3B, top). We correlated the RSMs formed at each layer to evaluate representational changes of Alexnet representations across layers. RSMs formed at successive layers were more similar between layers 2-5 and between layers 6-7, while the input layer 1 and the output layer 8 were the most distinct (Figure 3B, bottom).

We next set out to evaluate the similarity of stimulus representations in Alexnet to neural representations in VVS and PFC. In the VVS, representational geometries during encoding were captured by network representations in all layers and for all frequency bands except from high-gamma activity, where convolutional layers 2 and 3 did not show significant correlations (Figure 3C, top). During maintenance, we observed a significant fit between neural representations in the theta band to network representations in all fully connected layers (i.e., Fc6, Fc7 and Fc8; Figure 3D, top). In addition, we also observed two periods during the maintenance phase in which neural representations based on alpha oscillations matched network representations: (i) following presentation of the retro-cue (match with network representations in layers 5 and 6), and (ii) prior to the probe, i.e. towards the end of the maintenance period (match with network representations in all layers; Figure 3D, top, second panel). In the PFC, we observed a significant match between neural representations in the low-gamma frequency range during encoding with layers 1 and 2 (Figure 3C, bottom). However, PFC representations during the maintenance phase did not match with any representations in the network (Figure 3D, bottom).

Taken together, our RSA results based on the Alexnet network show a very pronounced matching during encoding in the VVS, where all frequency bands showed an effect, and a more limited matching to representations in the PFC. During maintenance, Alexnet representations matched those in the VVS based on theta and alpha oscillations, but did not explain the representational geometry of stimuli in the PFC, consistent with the absence of encoding-maintenance similarity described before.

We therefore employed the recurrent neural network BL-NET to further evaluate representations in VVS and, particularly, PFC. We first assessed the temporal evolution of network

representations in the different layers of BL-NET and correlated RSMs formed at all pairs of successive time-points (Methods). In all layers, representations changed most prominently between intervals 1-2 and least between intervals 7-8 (Figure 4A). In layers 2 to 7, representations remained largely constant at time step 4, while the first layer showed substantial dynamics until the last time interval (Figure 4A). Directly comparing the representations between the initial (1st) and the final (8th) time points separately for each layer revealed larger changes in the first two layers and substantially smaller changes in layers 3-7 (Figure 4B and 4C). CCI values were significantly higher than chance in all layers (all $p = 0.007$), but we observed a fourfold increase of CCI values from the first to the last layer of the network at all time points (Figure 4D).

Together, these results show that the BL-NET network represents the low-level features of our stimuli more dynamically than the abstract high-level features, and as expected, that it clusters categorical information more strongly in deep than superficial layers. We next compared model RSMs to the neural representations, focusing on the RSMs at the last time-point of each layer (Figure 4E). During encoding, results were similar to those in the Alexnet analysis (Figure 3): In the VVS, we observed a significant fit of the neural representations to network representations in all layers and for all frequency bands until the low-gamma range. For the high-gamma range, neural representations only matched network representations in layers 2 and 3 (Figure 4F, left). In the PFC, we found more restricted matching only with the last layer of the network and only for the theta range ($p = 0.005$, Figure 4F, right).

During maintenance, we observed in the VVS significant matches for the theta and alpha bands, again consistent with the results observed in the Alexnet analysis (Figure 4G, top). The effect in theta was observed towards the end of the maintenance period and specifically for the deepest layer of the model (i.e., layer 7). In the alpha range, we observed two time periods with significant fits to network representations: Following the retro-cue and towards the end of the maintenance period, similar to the results from the Alexnet analysis. All layers except from the last layer showed the effect in the late time period (time period ii in Figure 4G, top, first panel), while only layers 4 to 6 showed significant matches following the retro-cue (time period i; Figure 4G top, second panel).

Critically, in the PFC, representations in an early maintenance period (i.e., following the retro-cue) matched those of the last layer of the network specifically in the beta band ($p = 0.005$), but not in any other band (all $p > 0.2650$; Figure 4G, bottom). In order to further characterize the frequency profile of reactivations in VVS and PFC, we performed a frequency-resolved analysis of fits to the BL-NET network (Methods). We constructed RSM time-series for every frequency independently

and grouped them into a time-frequency map of model fits (Figure 4H). During encoding, this analysis revealed significant matching of VVS representations to BL-NET representations for a wide range of frequencies between 3 and 120Hz ($p = 0.001$), while no significant clusters were observed in the PFC (Figure 4I, left). During maintenance, no significant correlations were observed in the VVS (Figure 4I, right). By contrast, we observed a significant fit between neural and network RSMs in the beta frequency range (16-29Hz) starting ~400ms after the presentation of the cue and lasting for ~800ms. This effect was also specific to the last layer of the BL-NET (final layer $p = 0.005$; all other layers: $p = 0.14$), corroborating the abstract representational format of the PFC reactivation.

These results show that PFC representations following the retro-cue matched those of a recurrent network (the BL-NET) but not those of a purely feedforward network (the Alexnet), and that these effects were specific to the beta-frequency range.

Discussion

We analyzed category-specific representations during encoding and working memory prioritization in VVS and PFC and assessed their representational format using feedforward and recurrent DNN models of natural image processing. We report robust category-specific representations in the VVS in all frequencies and confined to specific frequency bands in the PFC during encoding. During maintenance, we observed a reinstatement of encoding representations after the retro-cue in VVS but not in the PFC, and hypothesized that the lack of reinstatement (i.e., EMS) in PFC may be due to a transformation of representational formats from encoding to maintenance. Using a recurrent DNN model to characterize the format of these representations in the PFC we observed abstract representational formats (i.e., matching the deepest network layer) following the retro-cue specifically in a frequency range between 16-29Hz. These results suggest that the prioritization process activates an abstract and transformed representation in the PFC that differs from the format during encoding.

Our results are in line with recent investigations showing a transformation of visual representations during VWM from perceptual to abstract formats during encoding and maintenance in iEEG patients (Liu et al., 2020, 2021). We further extend this research by investigating representational formats in brain regions that are critical for VWM storage and control, including the VVS and the PFC, respectively. Our findings show a matching of neural representations in the VVS during VWM maintenance with representations in all layers of two

DNN architectures (Alexnet and BL-NET), reflecting a presence of both perceptual and high-level abstract information in this region. Importantly, however, the same analysis in the PFC revealed a selective fit to the deepest layer of the recurrent DNN model, indicating presence of transformed representations (Figure 4G, bottom).

Our findings are consistent with the widely accepted role of the PFC in the top-down control of information stored in other brain regions during episodic and working memory (Eichenbaum, 2017; Miller et al., 2018). Indeed, activity in the PFC has been linked to task-dependent executive control over specific contents in several studies (for a review, see Rissman & Wagner, 2012). This could be achieved by modulating the activation state of distributed perceptual and mnemonic representations (Eichenbaum, 2017), for instance through PFC connectivity with the VVS (Oever et al., 2021). Indeed, the beta-frequency reactivation we observed in the PFC is suggestive of a top-down signal prompted by presentation of the cue that might affect information processing in downstream regions. Further analyses are required to investigate this possibility. Nevertheless, our results extend current views of the PFC by demonstrating its critical engagement in the reinstatement of abstract VWM representations during VWM prioritization.

What could be the functional role of this neural process? Our data is consistent with a capacity-limited view of WM which would benefit from compressed stimulus representations in PFC. This notion has been recently supported by behavioral (Kerren et al., 2022) and neuroimaging (Kwak & Curtis, 2022) studies. In particular, (Kerren et al., 2022) demonstrated that semantic aspects of images are selectively prioritized during WM maintenance in a multi-item WM task, while no selective storage of abstract features of images is present in single-item tests. In the fMRI study of (Kwak & Curtis, 2022), on the other hand, representational abstraction was observed in parietal and visual cortices, but not in prefrontal regions. The differences between our results and those of (Kwak & Curtis, 2022) might relate to the particular stimuli employed (natural images versus low-level visual features, which might engage early visual regions more prominently), and to our use of a paradigm involving prioritization, which preferentially engages PFC (Wallis et al., 2015). With regard to Kerren et al. (2022), we observed abstraction in the PFC even during maintenance of single items, while the VVS contained both perceptually detailed and abstract representations.

The abstract reactivation we observed in the PFC was particularly detected in the 16-29Hz frequency range, i.e., strongly overlapping with the traditional beta band (13-29Hz). Beta oscillations have been related to top-down control over information (Lundqvist et al., 2018; Miller et al., 2018), and in particular to maintenance and storage of information in the form of a buffer in the parietal cortex (Gelastopoulos et al., 2019). Although further analyses will be required to

investigate whether the reactivation we observed in PFC is linked to top-down control or mnemonic storage, we favor the former interpretation for several reasons: First, the effect is transient and locked to the presentation of the cue, i.e., to the time period when the prioritization process happens. Second, the abstract representational format is also stored in the VVS, a region more traditionally associated with storage of visual representations, during encoding and maintenance, and a PFC representation in the same format would be redundant. Finally, the reactivation was specific to the beta band, which has traditionally been associated with top-down control in the context of working memory (Miller et al., 2018).

Recent computational models of WM have proposed that prioritization requires a transformation of a subspace of activity in which items are represented in the brain. This transformation involves a rotation or “flip” of the format of prioritized content in neural activity space (Wan et al., 2022). These computational results have recently received some empirical support from studies with monkeys (Panichello & Buschman, 2021) suggesting an efficient neural code that organizes and structures neural representations during the prioritization process. While we have not performed population geometry analysis in our dataset, our work directly adds to this literature showing that WM prioritization depends not only on a transformation of the neural geometry of representations but also on a differential representation of particular visual formats in the context of VWM.

DNNs are increasingly used in cognitive neuroscience to characterize the representational formats and temporal dynamics of perceptual and mnemonic representations. While different feedforward and recurrent architectures have been tested in the domain of vision, resulting in a wide variety of models employed to fit neural data (Cadieu et al., 2014; Kar et al., 2019; Kar & DiCarlo, 2021; Khaligh-Razavi & Kriegeskorte, 2014; Kietzmann et al., 2019; Yamins et al., 2014), this approach has only started to be employed in memory research. Pioneering investigations have applied the deep feedforward convolutional network Alexnet to study representational formats during visual working memory in humans (Liu et al., 2020, 2021). While theoretical and experimental considerations have strongly argued for the use of recurrent architectures in the domain of visual perception (Kietzmann et al., 2019; van Bergen & Kriegeskorte, 2020; Yamins & DiCarlo, 2016), this has so far not been applied to memory research. In our study, we directly compared a feedforward and a recurrent model in their ability to predict human iEEG data. We found largely consistent results between these two types of architectures in the VVS during both encoding and maintenance. During encoding, both networks captured the representational geometry of stimuli across all layers and frequency bands, while this effect was restricted to theta and alpha during maintenance (Figure 4G). In the PFC, on the other hand, the two networks vastly

differed in their ability to fit the neural data: At encoding, we found a match to early layers only in the low-gamma frequency range in Alexnet (Figure 3C, bottom), while only the last layer of the RNN model fitted the neural representations in the theta band. During maintenance, the Alexnet was unable to capture PFC representations, which in the beta range specifically matched BL-NET representations (Figure 4G, bottom). Together, these results demonstrate that the recurrent architecture we employed explains better the representational geometry of stimuli in the PFC during VWM maintenance compared with a feedforward architecture, consistent with the critical role of recurrence observed in recent experimental work (Kar et al., 2019; Kar & DiCarlo, 2021; Kietzmann et al., 2019).

The BL-NET has two important limitations as a model of multi-item VWM. First, it was not trained to memorize stimuli but to solve the task of image recognition, which limits its heuristic value as a model of WM representations. Second, the BL-NET does not include top-down connections but has only lateral connectivity, which does not appropriately reflect PFC-VVS interactions. In future work, novel architectures could be employed that more accurately mimic brain connectivity at least at a high-level of description (i.e., containing top-down as well as within-layer connections), and that implement the process of multi-item VWM.

In summary, we present evidence of a transformation of neural representations after VWM prioritization in the PFC that is critically mediated by information in an abstract representational format in the 16-29Hz range. This result shows that percepts originally formed during encoding are abstracted and reshaped in the PFC to enable flexible task-dependent manipulations during working memory prioritization.

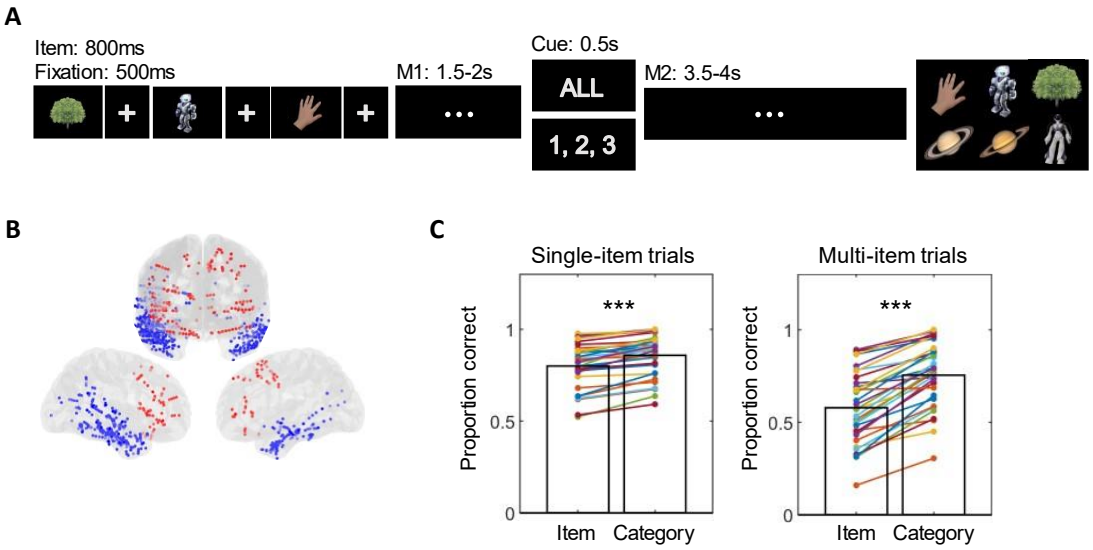

Figure 1: Experimental procedure and stimuli, electrode implantation, and behavioral results

(A) Participants encoded a sequence of 3 images of natural objects and were asked to remember this content during two subsequent maintenance periods that were separated by a retro-cue (M1 – retro-cue – M2). The cue prompted participants to either selectively maintain items at particular list positions (single-item trials, “1, 2, 3” in the figure) or to maintain all items in their order of presentation (multi-item trials, “All” in the figure). During the probe, six items were presented, which included all items shown together with at least one perceptually similar lure from the same category. (B) Electrode implantation included 441 electrodes in the ventral visual stream (VVS, N = 28 participants; blue) and 147 electrodes in the prefrontal cortex (PFC, N = 16 participants; red). (C) Behavioral performance was significantly higher at the category level as compared to the item level both for single and multi-item trials. ***: $p < 0.001$

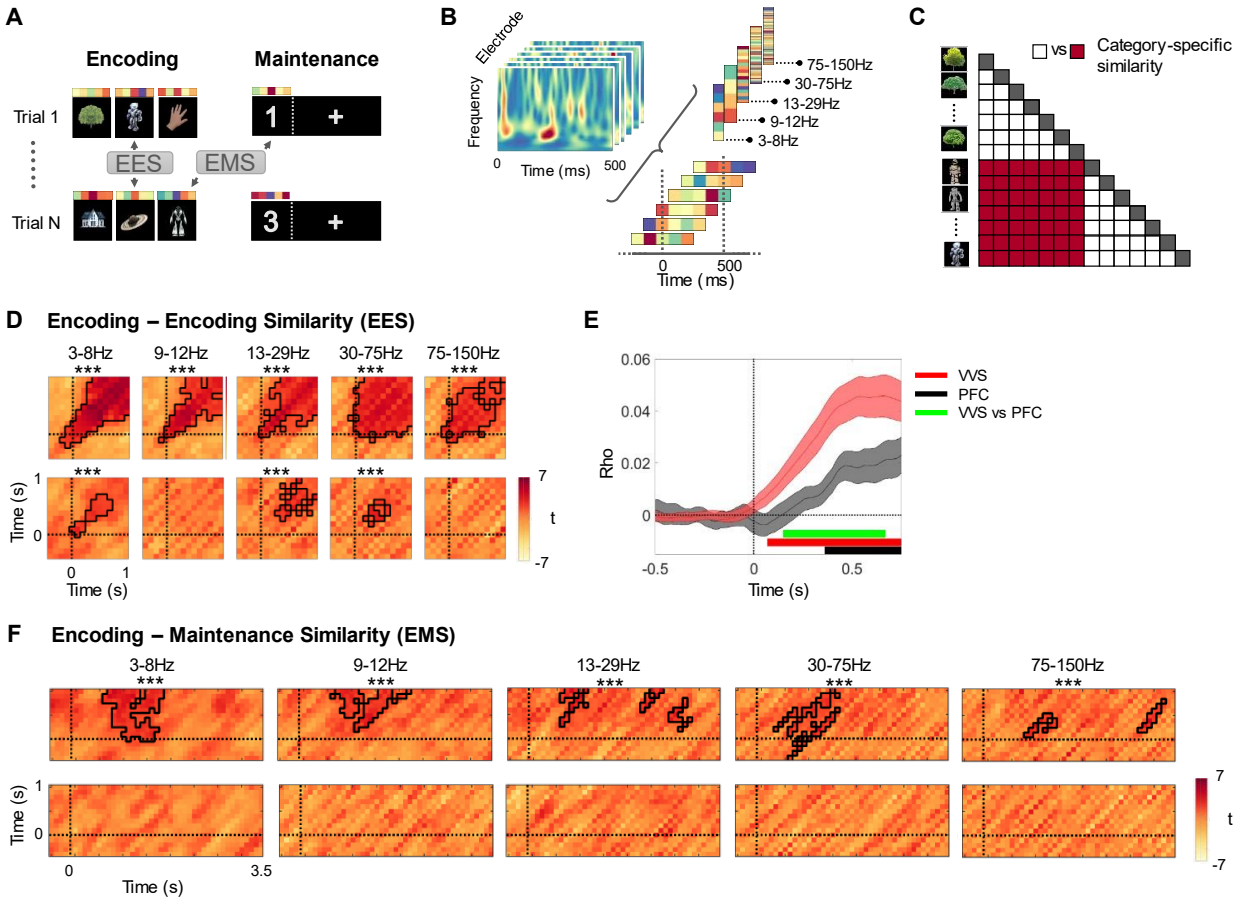

Figure 2: Encoding and maintenance of category-specific representations

(A) Analysis approach. Pattern similarity was calculated across trials either within the individual task stages (i.e., encoding-encoding similarity, EES) or between different task stages (i.e., encoding-maintenance similarity, EMS). (B) Pattern similarity was computed in windows of 500ms, incrementing in 100ms steps, using Spearman's correlation (rho) across electrodes, time points (5 time points in each 500ms window), and frequencies within particular frequency bands (theta, alpha, beta, low gamma, high gamma). (C) Representational Similarity Analysis (RSA): Category-specific representations were assessed by comparing activity patterns of different items from same categories vs. different items from different categories. (D) Category contrasts for five frequency bands during encoding in VVS (top) and PFC (bottom). Clusters with significant differences between same and different categories surviving correction for multiple comparisons are outlined in black. Zero indicates the onset of image presentation on both time axes. (E) Category specificity analysis at higher temporal resolution showing different latencies of effects in VVS (red) and PFC (black). Horizontal bars indicate time-periods where rho values are significantly different from zero in each region, and significantly different between the two regions

(green). (F) Encoding-Maintenance Similarity was computed by correlating patterns of activity at encoding (vertical axis) with those after the presentation of the retro-cue (horizontal axis). Zero indicates image onset and retro-cue onset during encoding and maintenance, respectively.

*** : $p < 0.001$

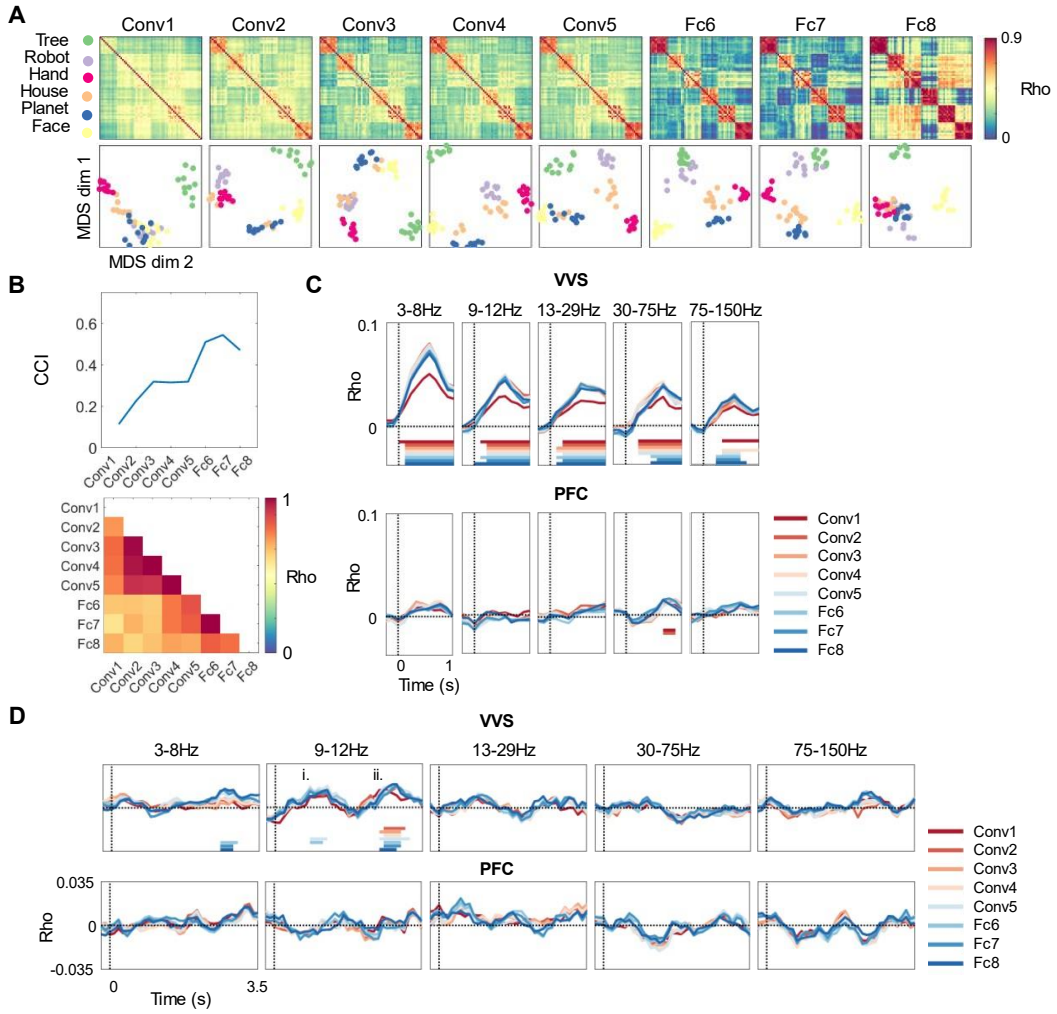

Figure 3: Analysis of representational formats using a feedforward deep neural network

(A) Top: Representations in the feedforward network Alexnet. Representational Similarity Matrices (RSMs) reflecting pairwise correlations of unit activations in each layer of the network (top). Bottom: 2D multidimensional scaling (MDS) projections of RSMs at each layer, color-coded according to categories. (B) Top: Category cluster index (CCI) as a function of network layer. CCI was computed by subtracting within and across category correlations at each individual layer. Bottom: Representational consistency plot showing pairwise correlations (Spearman's rho) of RSMs at each network layer. (C) Correlations between RSMs from the DNN and neural data, for each Alexnet layer and each encoding time window for VVS (top) and PFC (bottom). Each line shows the rho values of a particular layer. Horizontal lines at the bottom of each plot indicate time periods where rho values are significantly higher than zero at the group level (corrected for multiple comparisons across the 5 frequency bands). (D) Same analysis as in C for the

maintenance time-period. Two time-periods of significant group-level correlations, marked as i and ii, were observed in the alpha band (second panel from left to right in the top row).

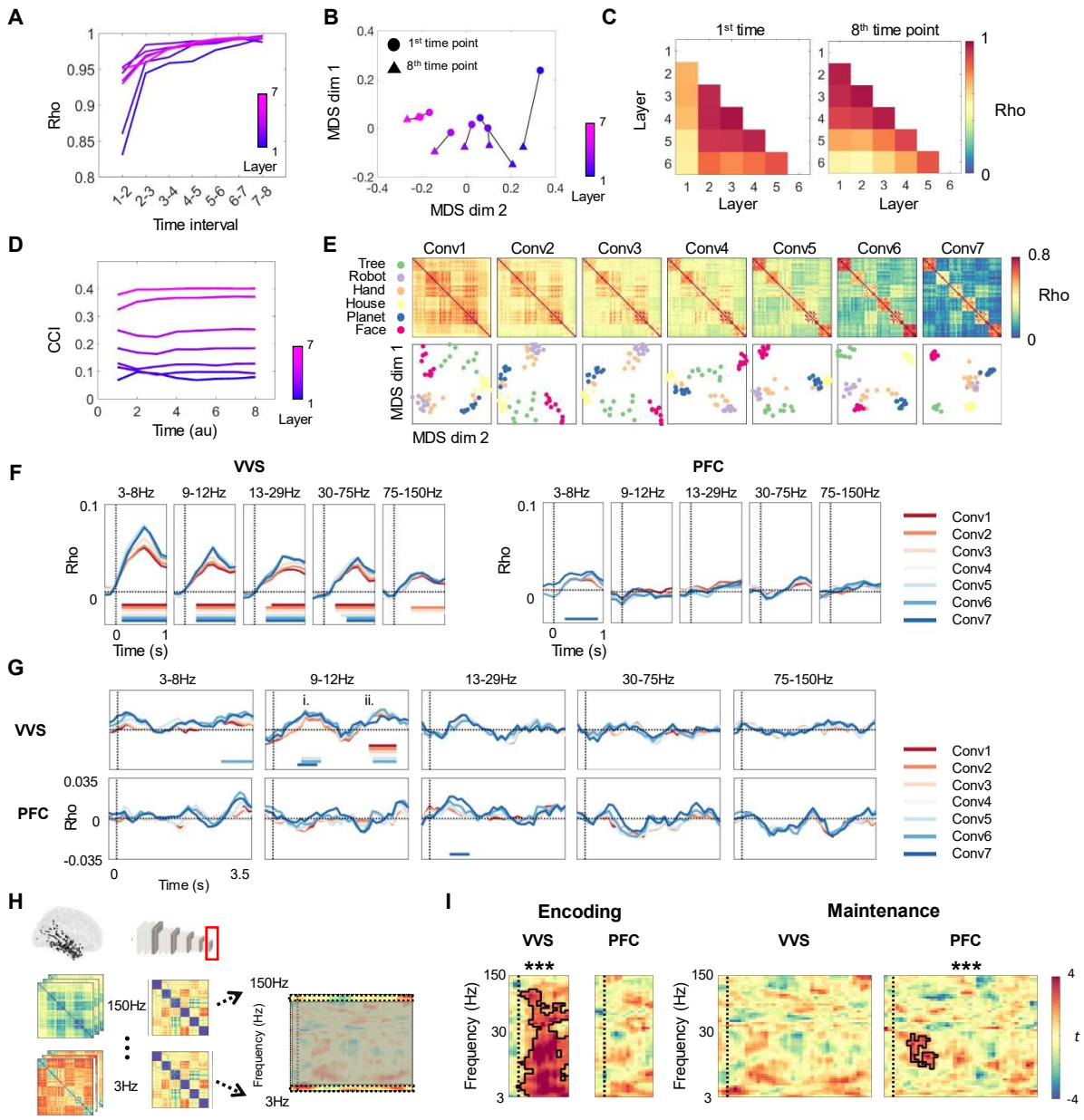

Figure 4: Analysis of representational formats using a recurrent deep neural network

(A) Representational consistency at each time interval of the BL-NET network was computed by correlating representations formed at successive time points. Each curve represents one layer of the network color coded from early (blue) to deep layers (pink). (B) Two-dimensional projections of the first and last time point of each layer in the BL-NET network showing greater representational distances in the first layer than in all other layers. (C) Pairwise correlations of RSMs corresponding to the first (left) and last (right) time points in each layer of BL-NET. (D) Category cluster index (CCI) defined as within vs. across category correlations for each layer and

time point of BL-NET. (E) RSMs and corresponding MDS projections for the last time point of the network in all layers. For the MDS plots, items are color-coded according to category. (F) Correlations between BL-NET RSMs (last time point in each layer) and neural RSMs (5 frequency bands) during encoding. Each line indicates correlations to a particular layer. Horizontal lines at the bottom of each plot indicate regions where rho values are significantly higher than zero at the group level (Bonferroni-corrected for 5 frequency bands). (G) Same analysis as in F for the maintenance period. Two time periods of significant correlations were identified in the alpha frequency range (indicated as i and ii). (H) Frequency-resolved analysis: we analyzed the correspondence of RSMs during the final time-point of the last BL-NET layer with RSM time-series for all individual frequencies in the PFC. (I) Left, encoding: In the VVS, RSMs of BL-NET were significantly correlated with neural RSMs across numerous frequencies in the 3-120Hz range. In the PFC, no significant matching was observed. Right, maintenance: No significant correlations were found in the VVS, but a prominent matching was observed in the PFC with representations in the beta band following presentation of the retro-cue.

***: $p < 0.001$.

Methods

Participants

Thirty-two patients with medically intractable epilepsy participated in the study. Data were collected at the Freiburg Epilepsy Center, Freiburg, Germany; the Epilepsy center, Second Affiliated Hospital, School of Medicine, Zhejiang University, Hangzhou, China; and the Center of Epileptology, Xuanwu Hospital, Capital Medical University, Beijing, China. The study was conducted according to the latest version of the Declaration of Helsinki and approved by the ethics committee of the University Hospital Freiburg and the Institutional Review Board at Xuanwu Hospital.

Experimental design

Participants performed a multi-item working memory paradigm involving a retro-cue. In the task, patients encoded a sequence of 3 images of natural objects from different categories and were asked to remember this content during a subsequent maintenance period. This period consisted of two phases that were separated by a retro-cue. The retro-cue prompted participants to selectively maintain items from particular encoding positions (single-item trials, 50%), or to maintain all items in their order of presentation (multi-item trials, 50%) for a subsequent memory test. In the test, six items were presented, which included all 3 presented items from encoding, two new exemplars from two of the encoding categories, and 1 item from a different category (Figure 1A). In the single-item trials, one of the items in the test was always from the same category as the cued item. Objects pertained to six categories with ten exemplars each (60 images in total). The task was designed to dissociate category-specific and item-specific information, and the test was conceived so that performance could be quantified both at the item and the category level. In order to perform the task correctly at the item level, participants needed to remember not only categorical information about the items but also the specific perceptual information identifying each individual exemplar. Note that performance at the item and the category level have by definition different chance levels. In the single-item trials, chance levels approximate 0.25 for the category level and 0.167 for the item level. In the multi-item trials, chance levels are very low when correct performance is defined as retrieving the entire sequence in the correct order. We computed chance levels through simulations with the real randomized sequences given to our participants. In the single-item trials, chance levels were at 0.25 ± 0.015 for the category level and 0.167 ± 0.001 for the item level. In the multi-item trials, chance levels

were at 0.0254 ± 0.0018 for the category level and at 0.0083 ± 0.00052 for the item level (considering as correct sequences in which all items were correctly retrieved in the right order).

The task was divided into blocks and sessions. Each block consisted of 60 trials. Each session consisted of at least one block, but most participants performed between 1 and 3 blocks in each session (2.59 ± 1.04) and between 1 and 2 sessions (1.19 ± 0.39) in total. The order and frequency of image presentations was pseudorandomized to balance repetitions of images across blocks and sessions. The experiment was programed in Presentation (Neurobehavioral systems, California, USA), and was deployed on Samsung 12'' tablet computers running Microsoft Windows. Patients performed the experiment while sitting in their hospital beds and responded to the memory test utilizing the touch-screen of the tablet.

Intracranial EEG recordings

IEEG data were recorded using amplifiers from Compumedics (Compumedics, Abbotsford, Victoria, Australia), and Brain Products GmbH with sampling rates of 2,000Hz and 2,500Hz, respectively. Patients were surgically implanted with intracranial depth electrodes for seizure monitoring and potential subsequent surgical resection. The exact electrode numbers and implantation locations varied across patients and were determined by clinical needs. Online recording data was referenced to a common scalp reference contact which was simultaneously recorded with the depth electrodes. Data was downsampled to 1,000Hz and bipolarized by subtracting the activity of one contact point with that from the nearest contact of the same electrode, resulting in a total of N-1 virtual channels for an electrode with N channels after bipolarization.

Channel Localization.

Electrodes employed were standard Behnke-Fried electrodes (Ad-Tech Medical Instrument Corporation). Electrodes contained variable number of contacts and inter-contact distances. In the data collected at the Zhejiang University, Hangzhou, and Medical University, Beijing, each depth electrode was 0.8 mm in diameter and had either 8, 12 or 16 contacts (channels) that were 1.5cm apart, with a contact length of 2 mm. Channel locations were identified by coregistering the post-implantation computed tomography (CT) images to the pre-implantation Magnetic Resonance Images (MRIs) acquired for each patient, which were afterwards normalized to MNI space using Statistical Parametric Mapping (SPM; https://www.fil.ion.ucl.ac.uk/spm/). We then determined the location of all electrode channels in MNI space using PyLocator (http://pylocator.thorstenkranz.de/) and 3DSlicer (https://www.slicer.org). To assign a label of

cortical regions to each channel, we segmented each patient's structural MRI using FreeSurfer (surfer.nmr.mgh.harvard.edu), using the pipeline described in (Stolk et al., 2018), and identified the closest cortical or subcortical label for each channel in each patient. We removed channels located in white matter, resulting in 588 clean channels across all patients (18.4 ± 11.9 per patient).

ROI selection

We selected two main regions of interest given their well-known involvement in VWM: the Ventral Visual Stream (VVS) and the Prefrontal Cortex (PFC). The VVS has been widely studied in the context of object recognition during visual perception (Cichy et al., 2014). Previous work employing iEEG and Deep Neural Networks most often applied RSA metrics to activity from distributed electrodes across the whole brain (Liu et al., 2020), and we specifically aimed to extend these studies by investigating region-specific representations in the context of VWM (see also Oever et al., 2021).

The role of the PFC in working memory has been linked to executive control processes that enable the task-dependent manipulation and transformation of information (Miller et al., 2018; Myers et al., 2017). However, relatively little is known about the representational formats of VWM representations in this region during prioritization. Moreover, no previous study investigated region-specific representational similarity during VWM, an approach that has been widely applied in the domain of perception.

Across our group of subjects, electrodes located at the following freesurfer locations were labeled as VVS electrodes: 'inferior temporal', 'middle temporal', 'superior temporal', 'bankssts', 'fusiform', 'cuneus', 'entorhinal'. Electrodes with the following labels were categorized as PFC electrodes: 'medial orbitofrontal', 'pars triangularis', 'superior frontal', 'lateral orbitofrontal', 'pars opercularis', 'rostral anterior cingulate', 'rostral middle frontal', 'superior frontal'. Electrodes from both left and right hemispheres were included in our ROIs.

Preprocessing

We visually inspected raw traces from all channels in each subject independently and removed noisy segments without any knowledge about the experimental events / conditions. All channels located within the epileptic seizure onset zone or severely contaminated by epileptic activity were removed from further analyses. We divided the data into 9-second epochs (from -2 to 7s) around the presentation of each stimulus at encoding or the onset of the retro-cue during the maintenance period. After epoching the data, we completely removed epochs containing artifacts that were

identified and marked in the non-epoched (continuous) data. We visually plotted spectrograms to verify the presence of artifacts in the frequency domain in the resulting epochs. The number of epochs corresponding to item or cue presentation that were removed varied depending on the quality of the signal in each subject (15.10 ± 13.14 in each session, corresponding to around 6.9% of all epochs in each session).

Preprocessing was performed on the entire raw data using EEGLAB (Delorme & Makeig, 2004), and included high-pass filtering at a frequency of 0.1Hz and low-pass filtering at a frequency of 200Hz. We also applied a band-stop (notch) filter with frequencies of 49-51Hz, 99-101Hz, and 149-151Hz.

Time-frequency analysis

Using the FieldTrip toolbox (Oostenveld et al., 2011), we decomposed the signal using complex Morlet wavelets with a variable number of cycles, i.e., linearly increasing in 29 steps between 3 cycles (at 3 Hz) and 6 cycles (at 29 Hz) for the low-frequency range, and in 25 steps from 6 cycles (at 30 Hz) to 12 cycles (at 150 Hz) for the high-frequency range. These time-frequency decomposition parameters were taken following previous research that used iEEG oscillatory power as features for RSA (Liu et al., 2020; Pacheco Estefan et al., 2019; Staresina et al., 2016). The resulting time-series of frequency-specific power values were then z-scored by taking as a reference the mean activity across all trials within an individual session (Fellner et al., 2020). This type of normalization was applied to remove any common feature of the signal unrelated to the encoding of stimulus-specific information. We z-scored across trials in individual sessions in our final analyses, but similar results were obtained when we z-scored the data by considering the activity of all trials irrespective of the session. We employed the resulting time-frequency data to build representational feature vectors in our pattern similarity analyses (see below).

Pattern similarity analysis

We investigated category-specific information by contrasting correlations of different items of the same category with correlations of different items from different categories. In order to obtain a robust estimate of the multivariate patterns representing individual items, we averaged the time-frequency activity across repetitions of items throughout the experiment in each channel independently before computing correlations, resulting in 60 averaged time-frequency maps for each channel. We then computed similarities for the “same category” and the “different category” item pairs and averaged across all combinations of items in the same condition in each subject independently (rho values were Fisher z-transformed before averaging). The “same category”

and “different category” correlations were then statistically compared at the group level using *t*-tests. We specifically excluded item pairs containing stimuli presented in overlapping trials when averaging across repetitions to avoid any possible bias related to the autocorrelation of the signal. In particular, in the different category contrast, we confirmed that none of the trials that were included in the average across repetitions were from the same trials when building the item pairs.

We employed temporally resolved Representational Similarity Analysis to evaluate the dynamics of categorical information in our data following previous work (Cichy et al., 2014; Kietzmann et al., 2019). The main assumption behind the RSA approach is that stimuli from the same categories will have greater neural similarity than stimuli from different categories. We quantified the similarity of neural representations by comparing epochs of brain activity separately in VVS and PFC. As in previous studies (Liu et al., 2020; Pacheco Estefan et al., 2021), we calculated Spearman’s correlations of frequency band-specific oscillatory patterns of activity across multiple contacts, i.e., from all electrodes available for each subject in a given region of interest. We employed five frequency bands: theta (3-8Hz), alpha (9-12Hz), beta (13-29Hz), low-gamma (30-75Hz) and high-gamma (75-120Hz). In addition, we applied RSA in individual frequencies (see below). Before performing similarity comparisons, we defined our representational feature vectors by specifying a 500ms time window in which we included the time courses of frequency-specific power values in time-steps of 100ms (5 time points) for each channel. Note that both the number of channels and the number of individual frequencies included in the feature vectors were variable depending on the number of electrodes available for a particular subject/ROI, and the number of frequencies included in each band. The number of frequencies was of 6, 4, 17, 9 and 16 for theta, alpha, beta, low-gamma and high-gamma, respectively (see section *Time-frequency decomposition* above). In the main analysis, a frequency band-specific representational pattern was thus composed of activity of N electrodes x M frequencies x 5 time-points in each 500ms window. These three-dimensional arrays were concatenated into 1D vectors for similarity comparisons.

We computed the pairwise correlation of these representational patterns across all available time-points using a sliding time window approach proceeding in time steps of 100ms (i.e., with an 80% overlap). We compared correlations at all combinations of time-points resulting in a temporal generalization matrix with two temporal dimensions on the vertical and horizontal axes (Figure 2D and 2F). Note that values in these matrices reflect both lagged (off-diagonal) and non-lagged (on-diagonal) correlations and were thus informative about the stability of neural representations over time.

Pattern similarity maps were computed for each pair of items in a correspondent condition at each time-window and rho values in these maps were Fisher z-transformed for statistical analysis. The temporal generalization maps were averaged across conditions for each subject independently, and the resulting average maps were contrasted via paired *t*-tests across conditions at the group level. Please note that in all pattern similarity plots (and also in the DNN-RSA plots, see below), correlations corresponding to each 500ms window were assigned to the time point at the center of the respective window (e.g., a time bin corresponding to activity from 0 to 500ms was assigned to 250ms).

RSA at high temporal resolution

We increased the temporal resolution of our sliding time window approach to compare the onset of category-specific information in VVS and PFC (Figure 2E). In this analysis, power values were computed with the same method and parameters as in the main analysis, but at an increased temporal resolution (10ms). Feature vectors were constructed in 500ms time windows and the 50 time-points included in each window were averaged separately for electrodes and frequencies, resulting in a two-dimensional representational pattern. We included all individual frequencies in the 3-150Hz range (a total of 52). These two-dimensional frequency x electrodes vectors were concatenated into one-dimensional arrays for similarity analyses. We employed a sliding time-window approach with incremental steps of 10ms resulting in an overlap of 490ms between two consecutive windows, focusing only on matching time points (non-lagged correlations). We performed this analysis separately for the VVS and the PFC and assessed the statistical significance of the resulting time-series in each region. At each time point, we compared the group-level Fisher z-transformed rho values against zero. We also directly compared the values between PFC and VVS at the group level. Given that not all subjects had implanted electrodes in both of our two ROIs, we performed unpaired *t*-tests at each time-point. We corrected for multiple comparisons by applying cluster-based permutation statistics in the temporal dimension in all the pattern similarity analyses (see section *Multiple Comparisons Correction* below).

Feedforward and recurrent DNN models

We compared VWM representations in the iEEG data with those formed in two convolutional deep neural network (DNN) architectures, a feedforward and a recurrent DNN. We used Alexnet (Krizhevsky et al., 2012), a widely applied network in computational neuroscience to model visual perception and WM, as our feedforward model (Kuzovkin et al., 2018; Liu et al., 2020). We also employed a recurrent convolutional DNN, BL-NET, which has been recently applied to model

human reaction times in a perceptual recognition task (Spoerer et al., 2020). Alexnet is a deep convolutional feedforward neural network composed of five convolutional layers and 3 fully connected layers that simulates the hierarchical structure of neurons along the ventral visual stream. Alexnet was trained in the task of object identification, i.e., the assignment of object labels to visual stimuli, using the ImageNet dataset (Deng et al., 2009). When learning to identify images, Alexnet develops layered representations of stimuli that hierarchically encode increasingly abstract visual properties: Early layers reflect low-level features of images such as edges or textures while deeper layers are sensitive to more complex visual information, such as the presence of objects or object parts. Several studies demonstrated the validity of Alexnet as a model of neural representations during biological vision, showing that it can capture relevant features of information processing in the VVS of humans during perceptual and mnemonic processing (Kuzovkin et al., 2018; Liu et al., 2021).

The BL-NET is a deep recurrent convolutional neural network consisting of 7 convolutional layers with feedforward and lateral recurrent connections. Every unit in the BL-NET network receives lateral input from other units within and across feature maps. BL-NET has demonstrated high accuracy in the task of object recognition after being trained with two large-scale image datasets (i.e., ImageNet and Ecoset; Mehrer et al., 2021; Spoerer et al., 2020).

The BL-NET is one out of many models that has recently introduced recurrence into the framework of convolutional neural networks. These studies have shown that recurrent convolutional neural networks can outperform parameter matched feedforward networks in the complex task of object recognition (Spoerer et al., 2020). Theoretical accounts proposing that recurrent DNNs can better explain neural and behavioral data than feedforward networks have recently been supported with experimental results (Kar et al., 2019; Kietzmann et al., 2019; Kubilius et al., 2018; Nayebi et al., 2018; Tang et al., 2018). While previous studies characterized neural representations in humans using recurrent models in the domain of visual perception (Kietzmann et al., 2019), no study so far has used these types of architectures to model VWM, and no study has applied them to iEEG data. Here we investigated whether the BL-NET can fit neural representations formed in the prefrontal cortex of humans performing a VWM task. Recurrence is particularly relevant in the context of PFC computations and has been associated with information maintenance via persistent activity in PFC (Edin et al., 2009).

Stimulus representations in DNNs

In order to analyze how the Alexnet and the BL-NET represent the stimuli in our study, we presented the networks with our images and computed unit activations at each layer. We

calculated Spearman's correlations between the DNN features for every pair of pictures, resulting in a 60 × 60 representational similarity matrix (RSM) in each layer (Kriegeskorte et al., 2008). The Alexnet unit activations were computed using the Matlab Deep Learning Toolbox. Images were scaled to fit the 227x227 input layer of the network. The unit activations in the BL-NET network were extracted using the pipeline described in https://github.com/cjspoerer/rcnn-sat. Images were scaled to 128x128 pixels to fit the input layer of the network as it was originally trained. BL-NET was trained for an arbitrary length of 8 time-steps. We extracted the unit activations at each of these time points and computed RSMs, resulting in a total of 7 (layers) x 8 (time-points) = 56 RSMs.

To visualize the representations of stimuli in our networks, we employed multi-dimensional scaling (MDS). MDS is a dimensionality reduction technique which exploits the geometric properties of RSMs, projecting the high-dimensional network activation patterns into lower-dimensional spaces. To apply the MDS algorithm to our RSMs, we subtracted the correspondent values in the matrix from 1 to obtain a distance metric, and projected the data into two dimensions (Figures 2A Bottom, 3B and E, Bottom).

In addition to computing the representational geometry of stimuli across all layers of our networks, we quantified the similarity between RSMs across layers using Spearman's Rho (a "second-level" similarity metric). We applied MDS to visualize the similarity structure of the BL-NET only for the initial and last time points in each layer (Figure 4B).

To quantify the presence of category information in our networks, we computed a Category Cluster Index (CCI), defined as the difference of average within-category and between-category correlations in the DNN representations of our stimuli. Both within and across category correlations averages were computed after removing the diagonal of the RSM matrices (which only contains values of 1 by definition) and duplicated values due to the symmetry of the RSMs. CCI approaches 1 if representations in all categories are perfectly clustered and 0 if no categorical structure is present in the data (Mehrer et al., 2020). We computed CCI at each layer of the Alexnet (Figure 3B), and for every time point and every layer in the BL-Net network (Figure 4D). To assess whether the observed CCI values were significant, we implemented a permutation procedure. We built a distribution of CCI values expected by chance by shuffling the trial labels of the network RSMs 1,000 times and recomputing CCI values. We considered significant CCI values that exceeded the 95th percentile of this null distribution.

RSA analyses based on DNNs: Modeling neural representations with deep neural networks

We compared the representations formed in the DNN architectures with the iEEG representations using RSA. Neural RSMs were computed in a time-resolved manner for each of our two ROIs. We employed the same parameters as in the pattern similarity analysis: Feature vectors were defined in 500ms time windows, sliding in 100ms steps, and included time as a dimension (5 time points). Separate analyses were conducted for the different frequency bands (theta, alpha, beta, low-gamma, high-gamma), and we averaged across image repetitions before applying the pair-wise correlations, resulting in a 60 by 60 neural RSM (of the same size as the model RSMs, see above). Different from the EES and the EMS analyses (Figure 2) in which we also considered non-matching time points (i.e., temporal generalization approach), we only focused on matching time-point in this analysis. We vectorized the two-dimensional neural and model RSMs by removing the diagonal values (by definition these had a value of 1) and taking only half of the matrix given its symmetry. We correlated the representational feature vectors using Spearman's rho. To assess statistical significance, we compared the resulting Fisher Z-transformed rho-values against zero at the group level, and we corrected for multiple comparisons using cluster-based permutation statistics (see below).

In addition, we performed a time-frequency resolved analysis of fits of neural RSMs and BL-NET RSMs. In this analysis, neural RSM time series (same time windows as described above) were computed with feature vectors comprising information of each individual frequency independently (e.g., for 3Hz, 4Hz, … 150Hz). The resulting RSM time-series were correlated with network RSMs at each individual layer of the BL-NET, focusing specifically on the last time point of each layer for this analysis. Individual frequencies were extracted using the same parameters as in the pattern similarity analysis (see section *Time-frequency analysis* above). The resulting time-series of correlations values were stacked into time-frequency maps of model fits (Figure 4H).

Multiple comparisons corrections

We performed cluster-based permutation statistics to correct for multiple comparisons in the pattern similarity analysis (Figure 2) and in the RSA-DNN analyses (Figures 3 and 4).

In the pattern similarity analysis, we applied cluster-based permutation statistics both for the temporal generalization analysis (Figure 2D and 2F), and for the temporally resolved analysis (Figure 2E). For both analyses, we contrasted same and different category correlations at different time-points using *t*-tests, as in the main analysis, after shuffling the trial labels 1,000 times. We considered significant a time-point if the difference between these surrogate conditions was significant at $p < 0.05$. At every permutation, we computed clusters of significant values defined

as contiguous regions in time where these significant correlations were observed and took the largest cluster at each permutation. Please note that in the temporal generalization analysis, time was defined in two dimensions and clusters were formed by grouping significant values across both of these dimensions, while in the temporally resolved analysis (Figure 2E), correlations were computed at matching time-points and clusters were formed along one temporal dimension. In both analyses, the permutation procedure resulted in a distribution of surrogate *t*-values under the assumption of the null hypothesis. We only considered significant those contiguous time pairs in the non-shuffled data whose summed *t*-values exceeded the summed *t*-value of 95% of the distribution of surrogate clusters (corresponding to a corrected $P < 0.05$; Maris & Oostenveld, 2007).

We also performed cluster-based permutation statistics in the analysis at high temporal resolution in which we directly compared similarity values between VVS and PFC. In this analysis, we computed clusters of significant EES differences between the two regions for every time-point by applying unpaired *t*-tests. We repeated this analysis 1,000 times after shuffling the region labels and kept the summed value of the largest cluster at every permutation. We only considered significant those clusters in the empirical data above the 95th percentile of the shuffled distribution.

In the RSA-DNN analyses, we applied cluster-based permutation statistics both to the DNN analysis in individual frequency bands and to the frequency-resolved analysis. In both analyses, to determine the significance of the correlations between neural and model RSMs, we recalculated the model RSMs at each layer of the network after randomly shuffling the labels of the images. The surrogate model similarity matrices were then correlated with the neural similarity matrix 1,000 times at all times (band-specific analysis) or time-frequency pairs (frequency-resolved analysis). As in the original analysis, we computed the correlations after removing the diagonal of the RSMs and only took half of the matrices given their symmetry. We identified clusters of contiguous time windows in the time domain or in the time-frequency domain where the group-level correlations between neural and network RSMs were significantly different than zero at $p < 0.05$ and selected the maximum cluster size of summed *t*-values for every permutation. This resulted in a distribution of surrogate *t*-values. The statistical significance was then determined by comparing the correlation values for the empirical data with the distribution of correlation values for the surrogate data.

In addition to cluster-based permutations, we also corrected our results for multiple comparisons using the Bonferroni method in the pattern similarity analysis (Figure 2) and in the RSA-DNN analysis conducted in individual bands (Figures 3 and 4). In those analyses, given that we tested five different frequency bands, we only considered *p*-values significant that were below an alpha

of 0.05/5. We also applied Bonferroni correction to our CCI analysis (Figures 3B and 4D). In these analyses, we adjusted the significance threshold according to the correspondent number of layers in each network that were tested (Alexnet= 8; BL-NET = 7).

Data Availability

Anonymized intracranial EEG data and custom-written Matlab code supporting the findings of this study will be made available upon publication.

Acknowledgments

We would like to acknowledge DFG funding via the ORA project “WMREPS Hidden brain states underlying efficient representations in working memory” (project number 396894956). This project was a collaboration with Prof. Mark Stokes (Oxford) and Prof. Elkan Akyürek (Groningen) and is dedicated to the memory of Prof. Stokes.

References

Buschman, T. J., & Kastner, S. (2015). From behavior to neural dynamics: An integrated theory of attention. *Neuron*, *88*(1), 127–144.

Buschman, T. J., & Miller, E. K. (2007). Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices. *Science*, *315*(5820), 1860–1862.

Cadieu, C. F., Hong, H., Yamins, D. L., Pinto, N., Ardila, D., Solomon, E. A., Majaj, N. J., & DiCarlo, J. J. (2014). Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLoS Computational Biology*, *10*(12), e1003963.

Chatham, C. H., Frank, M. J., & Badre, D. (2014). Corticostriatal output gating during selection from working memory. *Neuron*, *81*(4), 930–942.

Cichy, R. M., Pantazis, D., & Oliva, A. (2014). Resolving human object recognition in space and time. *Nature Neuroscience*, *17*(3), 455–462. https://doi.org/10.1038/nn.3635

Delorme, A., & Makeig, S. (2004). EEGLAB: an open-source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*(1), 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). *Imagenet: A large-scale hierarchical image database*. 248–255.

Edin, F., Klingberg, T., Johansson, P., McNab, F., Tegnér, J., & Compte, A. (2009). Mechanism for top-down control of working memory capacity. *Proceedings of the National Academy of Sciences*, *106*(16), 6802–6807. https://doi.org/10.1073/pnas.0901894106

Eichenbaum, H. (2017). Memory: Organization and Control. *Annual Review of Psychology*, *68*, 19–45. https://doi.org/10.1146/annurev-psych-010416-044131

Engel, A. K., & Fries, P. (2010). Beta-band oscillations—Signalling the status quo? *Current Opinion in Neurobiology*, *20*(2), 156–165.

Ester, E. F., Nouri, A., & Rodriguez, L. (2018). Retrospective cues mitigate information loss in human cortex during working memory storage. *Journal of Neuroscience*, *38*(40), 8538–8548.

Fellner, M.-C., Waldhauser, G. T., & Axmacher, N. (2020). Tracking selective rehearsal and active inhibition of memory traces in directed forgetting. *Current Biology*, *30*(13), 2638–2644.

Gelastopoulos, A., Whittington, M. A., & Kopell, N. J. (2019). Parietal low beta rhythm provides a dynamical substrate for a working memory buffer. *Proceedings of the National Academy of Sciences*, *116*(33), 16613–16620. https://doi.org/10.1073/pnas.1902305116

Griffin, I. C., & Nobre, A. C. (2003). Orienting attention to locations in internal representations. *Journal of Cognitive Neuroscience*, *15*(8), 1176–1194.

Higo, T., Mars, R. B., Boorman, E. D., Buch, E. R., & Rushworth, M. F. S. (2011). Distributed and causal influence of frontal operculum in task control. *Proceedings of the National Academy of Sciences*, *108*(10), 4230–4235. https://doi.org/10.1073/pnas.1013361108

Kar, K., & DiCarlo, J. J. (2021). Fast recurrent processing via ventrolateral prefrontal cortex is needed by the primate ventral stream for robust core visual object recognition. *Neuron*, *109*(1), 164–176.

Kar, K., Kubilius, J., Schmidt, K., Issa, E. B., & DiCarlo, J. J. (2019). Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature Neuroscience*, *22*(6), 974–983.

Kerren, C., Linde-Domingo, J., & Spitzer, B. (2022). Prioritization of semantic over visuo-perceptual aspects in multi-item working memory. *BioRxiv*.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Computational Biology*, *10*(11), e1003915.

Kietzmann, T. C., Spoerer, C. J., Sörensen, L. K., Cichy, R. M., Hauk, O., & Kriegeskorte, N. (2019). Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences*, *116*(43), 21854–21863.

Kriegeskorte, N., & Diedrichsen, J. (2019). Peeling the Onion of Brain Representations. *Annual Review of Neuroscience*, *42*, 407–432. https://doi.org/10.1146/annurev-neuro-080317-061906

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 4.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90.

Kubilius, J., Schrimpf, M., Nayebi, A., Bear, D., Yamins, D. L., & DiCarlo, J. J. (2018). Cornet: Modeling the neural mechanisms of core object recognition. *BioRxiv*, 408385.

Kuzovkin, I., Vicente, R., Petton, M., Lachaux, J.-P., Baciu, M., Kahane, P., Rheims, S., Vidal, J. R., & Aru, J. (2018). Activations of deep convolutional neural networks are aligned with gamma band activity of human visual cortex. *Communications Biology*, *1*(1), 1–12.

Kwak, Y., & Curtis, C. E. (2022). Unveiling the abstract format of mnemonic representations. *Neuron*.

Lepsien, J., & Nobre, A. C. (2006). Cognitive control of attention in the human brain: Insights from orienting attention to mental representations. *Brain Research*, *1105*(1), 20–31.

Lepsien, J., Thornton, I., & Nobre, A. C. (2011). Modulation of working-memory maintenance by directed attention. *Neuropsychologia*, *49*(6), 1569–1577.

Liebe, S., Hoerzer, G. M., Logothetis, N. K., & Rainer, G. (2012). Theta coupling between V4 and prefrontal cortex predicts visual short-term memory performance. *Nature Neuroscience*, *15*(3), 456–462. https://doi.org/10.1038/nn.3038

Liu, J., Zhang, H., Yu, T., Ni, D., Ren, L., Yang, Q., Lu, B., Wang, D., Heinen, R., Axmacher, N., & others. (2020). Stable maintenance of multiple representational formats in human visual short-term memory. *Proceedings of the National Academy of Sciences*, *117*(51), 32329–32339.

Liu, J., Zhang, H., Yu, T., Ren, L., Ni, D., Yang, Q., Lu, B., Zhang, L., Axmacher, N., & Xue, G. (2021). Transformative neural representations support long-term episodic memory. *Science Advances*, *7*(41), eabg9715.

Lundqvist, M., Herman, P., Warden, M. R., Brincat, S. L., & Miller, E. K. (2018). Gamma and beta bursts during working memory readout suggest roles in its volitional control. *Nature Communications*, *9*(1), 1–12.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG-and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177–190.

McKee, J. L., Riesenhuber, M., Miller, E. K., & Freedman, D. J. (2014). Task dependence of visual and category representations in prefrontal and inferior temporal cortices. *Journal of Neuroscience*, *34*(48), 16065–16075.

Mehrer, J., Spoerer, C. J., Jones, E. C., Kriegeskorte, N., & Kietzmann, T. C. (2021). An ecologically motivated image dataset for deep learning yields better models of human vision. *Proceedings of the National Academy of Sciences*, *118*(8), e2011417118. https://doi.org/10.1073/pnas.2011417118

Mehrer, J., Spoerer, C. J., Kriegeskorte, N., & Kietzmann, T. C. (2020). Individual differences among deep neural network models. *Nature Communications*, *11*(1), 1–12.

Miller, E. K., Lundqvist, M., & Bastos, A. M. (2018). Working Memory 2.0. *Neuron*, *100*(2), 463–475.

Myers, N. E., Stokes, M. G., & Nobre, A. C. (2017). Prioritizing information during working memory: Beyond sustained internal attention. *Trends in Cognitive Sciences*, *21*(6), 449–461.

Nayebi, A., Bear, D., Kubilius, J., Kar, K., Ganguli, S., Sussillo, D., DiCarlo, J. J., & Yamins, D. L. (2018). Task-driven convolutional recurrent models of the visual system. *Advances in Neural Information Processing Systems*, *31*.

Nee, D. E., & Jonides, J. (2008). Neural correlates of access to short-term memory. *Proceedings of the National Academy of Sciences*, *105*(37), 14228–14233. https://doi.org/10.1073/pnas.0802081105

Nee, D. E., & Jonides, J. (2009). Common and distinct neural correlates of perceptual and memorial selection. *Neuroimage*, *45*(3), 963–975.

Nelissen, N., Stokes, M., Nobre, A. C., & Rushworth, M. F. S. (2013). Frontal and parietal cortical interactions with distributed visual representations during selective attention and action selection. *Journal of Neuroscience*, *33*(42), 16443–16458.

Nobre, A. C., Coull, J. T., Maquet, P., Frith, C. D., Vandenberghe, R., & Mesulam, M. M. (2004). Orienting attention to locations in perceptual versus mental representations. *Journal of Cognitive Neuroscience*, *16*(3), 363–373.

Oever, S. T., Sack, A. T., Oehrn, C. R., & Axmacher, N. (2021). An engram of intentionally forgotten information. *Nature Communications*, *12*(1), 6443. https://doi.org/10.1038/s41467-021-26713-x

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J.-M. (2011). FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience*, *2011*, 156869. https://doi.org/10.1155/2011/156869

Pacheco Estefan, D., Sánchez-Fibla, M., Duff, A., Principe, A., Rocamora, R., Zhang, H., Axmacher, N., & Verschure, P. F. M. J. (2019). Coordinated representational reinstatement in the human hippocampus and lateral temporal cortex during episodic memory retrieval. *Nature Communications*, *10*(1), 1–13.

Pacheco Estefan, D., Zucca, R., Arsiwalla, X., Principe, A., Zhang, H., Rocamora, R., Axmacher, N., & Verschure, P. F. M. J. (2021). Volitional learning promotes theta phase coding in the human hippocampus. *Proceedings of the National Academy of Sciences*, *118*(10).

Panichello, M. F., & Buschman, T. J. (2021). Shared mechanisms underlie the control of working memory and attention. *Nature*, *592*(7855), 601–605.

Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology*, *32*(1), 3–25.

Rissman, J., & Wagner, A. D. (2012). Distributed representations in memory: Insights from functional brain imaging. *Annual Review of Psychology*, *63*, 101–128. https://doi.org/10.1146/annurev-psych-120710-100344

Schmidt, B. K., Vogel, E. K., Woodman, G. F., & Luck, S. J. (2002). Voluntary and automatic attentional control of visual working memory. *Perception & Psychophysics*, *64*(5), 754–763.

Sherfey, J., Ardid, S., Miller, E. K., Hasselmo, M. E., & Kopell, N. J. (2020). Prefrontal oscillations modulate the propagation of neuronal activity required for working memory. *Neurobiology of Learning and Memory*, *173*, 107228. https://doi.org/10.1016/j.nlm.2020.107228

Spoerer, C. J., Kietzmann, T. C., Mehrer, J., Charest, I., & Kriegeskorte, N. (2020). Recurrent neural networks can explain flexible trading of speed and accuracy in biological vision. *PLoS Computational Biology*, *16*(10), e1008215.

Sprague, T. C., Ester, E. F., & Serences, J. T. (2016). Restoring latent visual working memory representations in human cortex. *Neuron*, *91*(3), 694–707.

Staresina, B. P., Michelmann, S., Bonnefond, M., Jensen, O., Axmacher, N., & Fell, J. (2016). Hippocampal pattern completion is linked to gamma power increases and alpha power decreases during recollection. *ELife*, *5*, e17397. https://doi.org/10.7554/eLife.17397

Stolk, A., Griffin, S., van der Meij, R., Dewar, C., Saez, I., Lin, J. J., Piantoni, G., Schoffelen, J.-M., Knight, R. T., & Oostenveld, R. (2018). Integrated analysis of anatomical and electrophysiological human intracranial data. *Nature Protocols*, *13*(7), 1699–1723. https://doi.org/10.1038/s41596-018-0009-6

Tang, H., Schrimpf, M., Lotter, W., Moerman, C., Paredes, A., Ortega Caro, J., Hardesty, W., Cox, D., & Kreiman, G. (2018). Recurrent computations for visual pattern completion. *Proceedings of the National Academy of Sciences*, *115*(35), 8835–8840. https://doi.org/10.1073/pnas.1719397115

van Bergen, R. S., & Kriegeskorte, N. (2020). Going in circles is the way forward: The role of recurrence in visual inference. *Current Opinion in Neurobiology*, *65*, 176–193.

Vezoli, J., Vinck, M., Bosman, C. A., Bastos, A. M., Lewis, C. M., Kennedy, H., & Fries, P. (2021). Brain rhythms define distinct interaction networks with differential dependence on anatomy. *Neuron*, *109*(23), 3862–3878.

Vogel, E. K., & Machizawa, M. G. (2004). Neural activity predicts individual differences in visual working memory capacity. *Nature*, *428*(6984), 748–751. https://doi.org/10.1038/nature02447

Wallis, G., Stokes, M., Cousijn, H., Woolrich, M., & Nobre, A. C. (2015). Frontoparietal and cingulo-opercular networks play dissociable roles in control of working memory. *Journal of Cognitive Neuroscience*, *27*(10), 2019–2034.

Wan, Q., Menendez, J. A., & Postle, B. R. (2022). Priority-based transformations of stimulus representation in visual working memory. *PLoS Computational Biology*, *18*(6), e1009062. https://doi.org/10.1371/journal.pcbi.1009062

Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*(3), 356–365.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624.