Stable maintenance of multiple representational formats in human visual short-term memory

Jing Liu[a,b,1], Hui Zhang[c,1], Tao Yu[d], Duanyu Ni[d], Liankun Ren[e], Qinhao Yang[a,b], Baoqing Lu[a,b], Di Wang[e], Rebekka Heinen[c], Nikolai Axmacher[a,b,c,2], and Gui Xue[a,b,2,3]

[a]State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University, 100875 Beijing, China; [b]IDG/McGovern Institute for Brain Research, Beijing Normal University, 100875 Beijing, China; [c]Department of Neuropsychology, Institute of Cognitive Neuroscience, Faculty of Psychology, Ruhr University Bochum, 44801 Bochum, Germany; [d]Beijing Institute of Functional Neurosurgery, Xuanwu Hospital, Capital Medical University, 100053 Beijing, China; and [e]Comprehensive Epilepsy Center of Beijing, Department of Neurology, Xuanwu Hospital, Capital Medical University, 100053 Beijing, China

Edited by Wolf Singer, Max Planck Institute for Brain Research, Frankfurt, Germany, and approved November 5, 2020 (received for review April 9, 2020)

Visual short-term memory (VSTM) enables humans to form a stable and coherent representation of the external world. However, the nature and temporal dynamics of the neural representations in VSTM that support this stability are barely understood. Here we combined human intracranial electroencephalography (iEEG) recordings with analyses using deep neural networks and semantic models to probe the representational format and temporal dynamics of information in VSTM. We found clear evidence that VSTM maintenance occurred in two distinct representational formats which originated from different encoding periods. The first format derived from an early encoding period (250 to 770 ms) corresponded to higher-order visual representations. The second format originated from a late encoding period (1,000 to 1,980 ms) and contained abstract semantic representations. These representational formats were overall stable during maintenance, with no consistent transformation across time. Nevertheless, maintenance of both representational formats showed substantial arrhythmic fluctuations, i.e., waxing and waning in irregular intervals. The increases of the maintained representational formats were specific to the phases of hippocampal low-frequency activity. Our results demonstrate that human VSTM simultaneously maintains representations at different levels of processing, from higher-order visual information to abstract semantic representations, which are stably maintained via coupling to hippocampal low-frequency activity.

visual short-term memory | intracranial EEG | representation | deep neural network | hippocampus

Visual short-term memory (VSTM) refers to the active maintenance of visual information for a short period of time (1, 2). Classical models assume that VSTM representations are built during initial perception and maintained via persistent firing of neurons in prefrontal cortex (3, 4). By contrast, recent human neuroimaging and primate neurophysiological studies suggest that VSTM maintenance relies neither on a stable code (5, 6) nor on persistent neuronal activity (7, 8). Instead, VSTM may involve processes of "dynamic coding" which lead to substantial transformations of neural representations. The transformation processes may reflect the encoding of information along the ventral visual stream (9), the transformation of perceived stimuli into internal representations (5, 6), or the mapping of VSTM representations onto appropriate motor plans (10). After these transformations, task-relevant neural representations of VSTM may be retained in a more stable form (11). Meanwhile, despite highly dynamic coding at the single-neuron level, neural activities at the population level contain subspaces in which stimulus representations are stable across VSTM encoding and maintenance (12, 13).

Regarding the persistency of neuronal activations during maintenance, it has been shown that item-specific VSTM representations can be retained in an "activity-silent" state during the maintenance period (11) which does not require persistent activity increases. Activity-silent representations can still be identified by multivariate decoding algorithms (14–16) and can be recovered to an active state by transcranial magnetic stimulation (TMS) impulses (17). Meanwhile, cross-frequency coupling models of VSTM suggest that individual items are represented by neural assemblies which are synchronized via high-frequency (i.e., gamma) oscillations that are locked to specific phases of hippocampal low-frequency oscillations (18–20). This coupling may result in phase coding, such that neural representations of specific items are coupled to distinct phases of low-frequency oscillations, according to either the identity of an item (21, 22) or its position on a list (23). Notably, a very recent study integrated the concepts of activity-silent VSTM representations and phase coding by showing that the amplification of activity-silent working memory representations depends on the phase of ongoing electroencephalography (EEG) oscillations at which the impulse is applied (24).

In light of these dynamic processes, it remains unclear whether and how humans can maintain stable representations of complex visual images in VSTM. First, although simple visual features can be decoded (14, 15) or reconstructed via encoding models (25–27), it is unknown how faithfully (e.g., item-specific) and how stably (i.e., temporally generalizable) natural images can be maintained in VSTM. Second, classical models of VSTM assume

Significance

Visual short-term memory (VSTM) is the ability to actively maintain visual information for a short period of time. Classical models posit that VSTM is achieved via persistent firing of neurons in prefrontal cortex. Leveraging the unique spatiotemporal resolution of intracranial EEG recordings and analytical power of deep neural network models in uncovering the neural code of visual processing, our results suggest that visual information is first dynamically extracted in multiple representational formats, including higher-order visual format and abstract semantic format. Both formats are stably maintained across an extended period via coupling to phases of hippocampal low-frequency activity. These results suggest human VSTM is highly dynamic and involves rich and multifaceted representations, which contribute to a mechanistic understanding of VSTM.

Author contributions: J.L. and G.X. designed research; J.L., T.Y., D.N., L.R., Q.Y., B.L., and D.W. performed research; J.L., H.Z., R.H., N.A., and G.X. analyzed data; and J.L., H.Z., R.H., N.A., and G.X. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]J.L. and H.Z. contributed equally to this work.

[2]N.A. and G.X. contributed equally to this work.

[3]To whom correspondence may be addressed. Email: gxue@bnu.edu.cn.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2006752117/-/DCSupplemental.

First published December 7, 2020.

that maintenance relies on either visuospatial or verbal information (1). However, natural visual objects are encoded at multiple processing stages which contain different "representational formats" (28). This processing cascade starts with the extraction of lower-level visual properties (e.g., color and contrast) to more complex texture information, to superordinate categories (e.g., animate/inanimate), up to abstract conceptual and semantic information (9, 29). Whether human VSTM maintenance relies on one or several of these representational formats is unknown. Finally, given the complexity of these representations, a certain form of temporal dynamics (e.g., rhythmic fluctuations and/or phase coding) might be essential to support their maintenance.

Combining the high temporal resolution of human intracranial EEG recordings with a deep neural network (DNN) and a semantic model that enable the characterization of representational formats (29–33), the current study aimed to examine the nature and dynamics of representations during short-term maintenance of natural visual images. An extended encoding period (3 s) was used to capture the entirety of the dynamic processes during visual object encoding and to clearly separate them from the subsequent maintenance period. We also used a relatively long maintenance period (7 s) to examine the stability of representations. Our results revealed substantial transformations during the encoding period, but stable item-specific maintenance in two distinct representational formats. At a finer temporal resolution, maintenance of both representational formats exhibited arrhythmic fluctuations, which were coupled to phases of hippocampal low-frequency activity.

Results

Behavioral Results. Intracranial EEG was recorded during a delayed matching to sample (DMS) task (Fig. 1*A*) in 19 epilepsy patients (mean age ± SD, 27.9 ± 7.1; 5 female) with depth electrodes implanted for clinical purposes (total, 529 channels; mean ± SD, 27.8 ± 12.4 channels per patient; Fig. 1*B*). Participants first encoded a word–picture pair for 3 s and then maintained the picture for a 7-s delay period, during which the picture disappeared but the cue word remained on screen. After the maintenance period, a probe was displayed and participants were asked to indicate whether it matched the previously presented target picture. To encourage participants to maintain visual details of the picture, the nonmatching probes were highly similar to the respective target pictures. Participants performed well in the DMS task (accuracy rate 0.89 ± 0.04, mean ± SD; d-prime 2.59 ± 0.49, mean ± SD; *SI Appendix*, *SI Text*); thus the following analyses focused only on correct trials.

Item-Specific Representations during VSTM Maintenance. We first tested whether there were item-specific neural representations during the VSTM maintenance period. To this end, we conducted a global representational similarity analysis (RSA), correlating intracranial EEG (iEEG) power across channels and frequencies

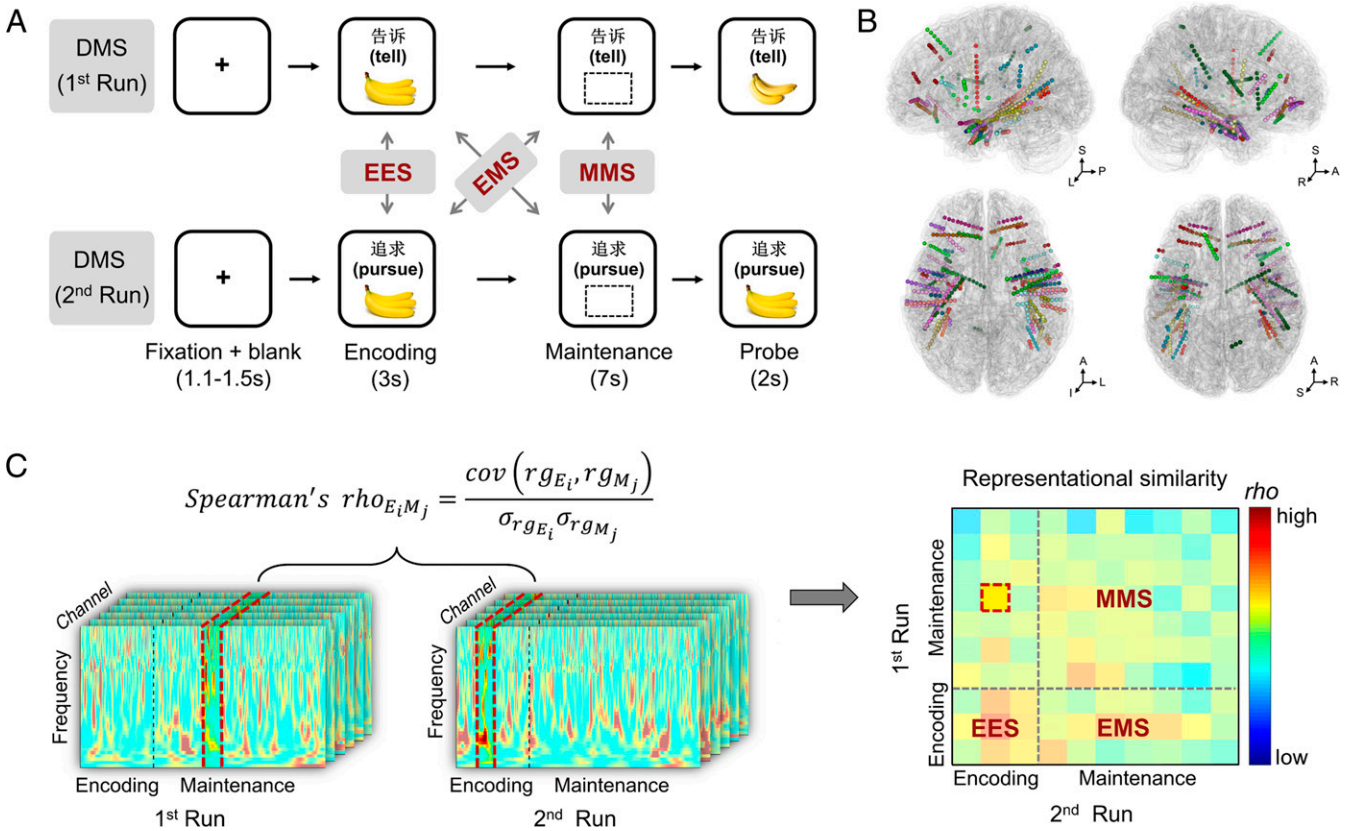

Fig. 1. Experimental paradigm, intracranial EEG electrodes, and analysis approach. (*A*) A DMS task was used in the experiment. The same picture was paired with different cue words in two consecutive experimental runs. The representational similarity was then calculated between trials from different runs, both within the individual task stages (i.e., encoding–encoding similarity [EES], maintenance–maintenance similarity [MMS]) and across different task stages (i.e., encoding–maintenance similarity [EMS]). (*B*) Normalized electrode localization map. Each colored sphere indicates one channel, with different colors representing different participants. (*C*) Global RSA was performed by correlating iEEG power across 43 frequencies (between 2 and 100 Hz) and across all clean channels in consecutive time windows of 200 ms, sliding in steps of 10 ms. (*Left*) We correlated activity patterns during one trial from the first run (left, between dashed lines) with activity patterns during another trial in a second run (right, between dashed lines). (*Right*) Resulting similarity map across various time windows. The different sections (separated by dashed lines) show within- and across-stage similarity, corresponding to EES, EMS, and MMS in *A*.

(from 2 to 29 Hz in steps of 1 Hz and from 30 to 100 Hz in steps of 5 Hz) (34) within consecutive overlapping time windows of 200 ms, incrementing in steps of 10 ms (Fig. 1*C*). Item-specific representations were identified by comparing neural pattern similarity of trial pairs with the same picture (within-item [WI] similarity) versus trial pairs with different pictures (between-item [BI] similarity) (Fig. 1*A* and *Materials and Methods*). We averaged WI and BI maintenance–maintenance similarities (MMS) across the entire delay period (Fig. 2*A*) and compared them using a paired *t* test. This analysis revealed greater WI than BI MMS ($t(18) = 2.57$, $P = 0.02$, Fig. 2*B*), providing clear evidence for item-specific representations during the VSTM maintenance period.

Representations from two different encoding periods are maintained. We next examined which encoding periods contained representational formats that were maintained during VSTM. To this end, we calculated the encoding–maintenance similarity (EMS) between each encoding time window and each maintenance time window (Fig. 2 *C*, *Left*). Item-specific EMS was then identified via the contrast of WI EMS and BI EMS. For each encoding time window, item-specific EMS was averaged across the entire maintenance period and then tested against zero. Two temporal clusters during encoding showed significant item-specific EMS that survived cluster-based correction for multiple comparisons across all encoding time windows (*Materials and Methods*): an early cluster (250 to 770 ms, $t(18) = 3.51$, $P = 0.003$; $P_{corrected} = 0.012$) and a late cluster (1,000 to 1,980 ms, $t(18) = 4.00$, $P < 0.001$; $P_{corrected} < 0.001$) (Fig. 2 *C*, *Right*). A two-way ANOVA with "cluster" (early vs. late cluster) and "item specificity" (WI vs. BI EMS) as repeated measures revealed a significant interaction effect ($F(1,18) = 6.44$, $P = 0.021$; Fig. 2*D*), indicating greater item specificity for the late than the early cluster. The WI EMS was also greater in the late than in the early cluster ($t(18) = 2.93$, $P = 0.009$). These results suggest that information from both encoding periods contributes to representations during maintenance, with a more prominent role of the late encoding period.

Distinct representational formats in early and late encoding clusters. The above analyses indicate that representations from two different encoding periods are maintained in VSTM. However, they do not indicate whether these two clusters contain similar or different representational formats. To address this question, we recomputed item-specific EMS in the early cluster while regressing out the representations in the late cluster and vice versa. We still found significant item-specific EMS in both the early ($t(18) = 2.51$, $P = 0.022$) and the late encoding clusters ($t(18) = 3.66$, $P = 0.002$) (*SI Appendix*, Fig. S1), indicating that VSTM contains information in two distinct and separable representational formats.

To further characterize the representational formats of the two encoding clusters, we compared the corresponding neural representations with representations in a DNN and a semantic model (Fig. 3*A* and *Materials and Methods*). In this DNN, low-level visual features (e.g., color, contrast) are processed in early layers and more complex object features are processed in deeper layers [see *SI Appendix*, Fig. S2 for detailed depictions of the "AlexNet" (35) and *SI Appendix*, Fig. S3 for the representational structure in all DNN layers] (*SI Appendix*, *SI Text*). We constructed neural similarity matrices in individual encoding time windows based on correlations between all possible pairs of items. The DNN similarity matrices were obtained by correlating activations of artificial neurons between the corresponding pairs of items in each DNN layer. The neural similarity matrices were then correlated with these DNN similarity matrices (Fig. 3*B*). We averaged the Fisher Z-transformed correlation values within the early and the late encoding cluster, separately for each of the eight DNN layers.

Comparing the correlation values for individual layers against zero showed that neural representations in the early cluster were significantly correlated with DNN representations in deep layers (layer 5, $P_{FDR} = 0.049$; layer 6, $P_{FDR} = 0.049$; layer 7, $P_{FDR} = 0.022$; layer 8, $P_{FDR} = 0.022$, corrected for multiple comparisons by false discovery rate [FDR]), marginally significantly correlated with DNN representations from layer 2 to layer 4 (layer 2, $P_{FDR} = 0.050$; layer 3, $P_{FDR} = 0.058$; layer 4, $P_{FDR} = 0.050$), and not significantly correlated with representations in the first layer (layers 1, all $P_{FDR} = 0.16$). By contrast, representations in the late cluster were not significantly correlated with representations in any DNN layers (all $P_{FDR} > 0.46$) (Fig. 3*C*). We confirmed this

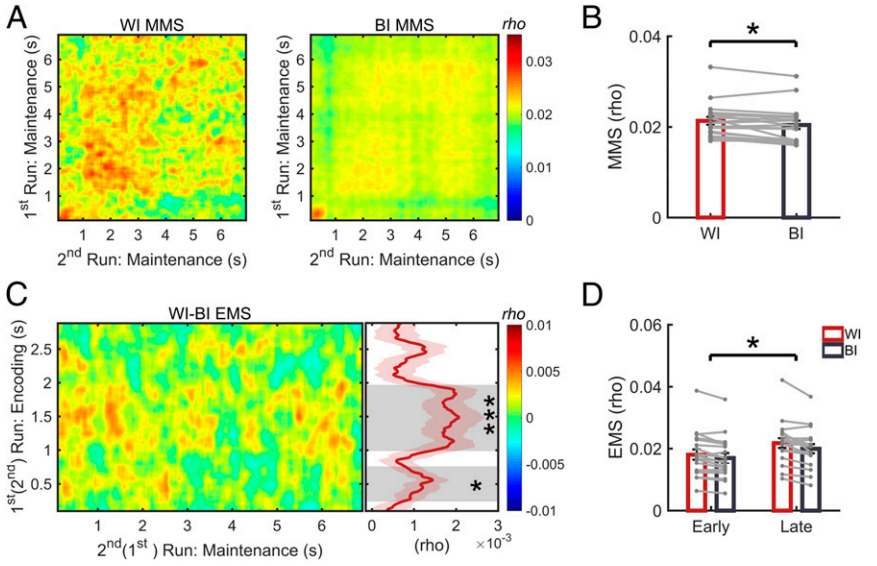

Fig. 2. Maintenance of item-specific representations from two distinct encoding periods. (*A*) WI and BI MMS. WI MMS was calculated between pairs of items with same pictures but different cue words, while BI MMS was calculated between pairs with different pictures and different cue words. (*B*) Averaged WI and BI MMS across the entire maintenance period (7 s). Each gray dot indicates one participant. (*C*) Two encoding clusters showed item-specific EMS. Item-specific EMS (WI-BI EMS) was calculated for consecutive encoding time windows (vertical axis). (*Left*) EMS between individual time windows of encoding and maintenance period. (*Right*) EMS averaged across the entire maintenance period, separately for each encoding time window. The two shaded areas mark the two encoding clusters with significant item-specific EMS. (*D*) WI and BI EMS in the early and the late cluster. Greater item-specific EMS was found in the late cluster than in the early cluster. Each gray dot indicates one participant. Error bars reflect one SEM. $*P < 0.05$; $***P < 0.001$.

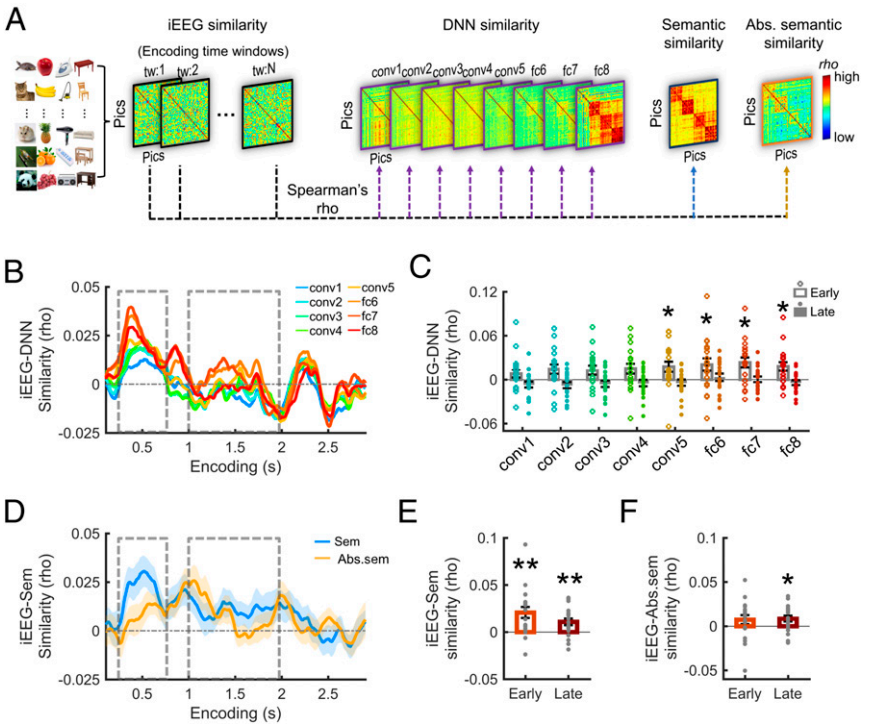

Fig. 3. Distinct representational formats in the early and late encoding clusters. (*A*) Linking neural representations to visual, semantic, and abstract semantic representations, respectively. Neural similarity matrices were created by correlating activities between all pairs of different pictures in individual encoding time windows. Visual similarity matrices were obtained by correlating activations of the artificial neurons in each DNN layer. The semantic similarity matrix was generated by calculating the cosine similarities of the word vectors between picture labels. An abstract semantic similarity matrix resulted from iteratively regressing out the visual similarity in each DNN layer from the semantic similarity matrix. Exemplar pictures of items from the four categories (animals, fruits, electrical devices, and furniture) used in the study are shown at *Left*. (*B*) Correlation (Spearman's rho) between DNN and neural similarity maps for each DNN layer and each encoding time window. Two dashed boxes indicate the time window of the early and the late encoding cluster showing significant item-specific EMS. (*C*) Averaged correlation between neural similarity in the early and late encoding clusters with the similarity matrices in each DNN layer. Each color dot indicates the correlation value from one participant. (*D*) Correlation values resulting from comparing the neural similarity with the semantic (blue line) and the abstract semantic similarity (orange line) in each encoding time window. Two dashed boxes mark the time windows corresponding to the early and the late encoding cluster. Color shaded areas around the lines reflect one SEM. (*E*) Semantic representations correlated with neural representations in both the early and late encoding clusters. (*F*) Abstract semantic representations correlated with neural representations in the late encoding cluster. Each color dot indicates one participant. Error bars reflect one SEM. $*P < 0.05$; $**P < 0.01$.

result by performing a permutation analysis in which we recreated the neural similarity matrices after randomly shuffling the picture labels (*SI Appendix*, Fig. S4). A two-way ANOVA with "layer" (eight layers) and "encoding cluster" (early vs. late cluster) as repeated measures revealed significant main effects of encoding cluster ($F(1,18) = 13.23$, $P = 0.002$) and layer ($F(7,126) = 2.41$, $P = 0.02$), but no significant interaction effect ($F(7,126) = 0.54$, $P = 0.80$), indicating stronger higher-order visual representations in the early cluster.

To examine whether the representations in the two encoding clusters contained semantic information, we extracted the word labels for all pictures in the experiments (*SI Appendix*, *SI Text*) and then used a Chinese semantic model, i.e., Directional Skip-Gram (36), to convert each label into a vector with 200 semantic features. We obtained a semantic similarity matrix between the semantic features of all possible picture pairs (*SI Appendix*, Fig. S3). The neural representational similarity matrices in individual encoding time windows were then correlated with the semantic similarity matrix (Fig. 3*D*). We found that the semantic similarity matrix was significantly correlated with the neural similarity matrices in both the early ($t(18) = 3.52$, $P = 0.002$) and the late ($t(18) = 3.04$, $P = 0.007$) encoding clusters (Fig. 3*E*). This result was also confirmed by the permutation analysis described above (early cluster, $P < 0.001$; late cluster, $P = 0.003$; *SI Appendix*, Fig. S4).

To disentangle the contributions of perceptual versus abstract semantic information, we iteratively regressed out the similarity matrices of all eight DNN layers from the semantic similarity matrix. We correlated the resulting "abstract semantic similarity matrix" with the neural representational similarity matrices in the early and the late cluster, respectively (Fig. 3*D*). The analysis revealed significant correlations for the late encoding cluster ($t(18) = 2.53$, $P = 0.02$), but not for the early encoding cluster ($t(18) = 1.44$, $P = 0.17$) (Fig. 3*F*). The results indicate that only the late encoding cluster contains abstract semantic representations, while semantic representations in the early cluster are shaped by higher-order perceptual information.

Stable maintenance of both representational formats during VSTM. We next examined the temporal dynamics during the maintenance of item-specific representations in VSTM. Specifically, two aspects of temporal dynamics were examined: first, whether the representational formats were stable or whether they underwent a systematic transformation across the maintenance period and, second, whether the strength of reactivated representational formats was stable across time.

We performed two analyses to examine the stability of representational formats. In the first analysis, we calculated the similarity between the neural activities at different encoding and maintenance time windows across two independent runs, which shared the same items. A lower off-diagonal than on-diagonal similarity would indicate transformation of representational formats from one time point

to another, reflecting dynamic coding (13). This analysis revealed two clusters that showed dynamic coding during the encoding period ($P_{corrected} < 0.001$, corrected for multiple comparisons by cluster-based permutation test; Fig. 4*A*), but not during the maintenance period ($P_{corrected} > 0.24$). To better characterize the representational dynamicity, we computed the dynamicism index (*di*) across time (*Materials and Methods*). The dynamicism index reached relatively large values during the first 2 s of encoding and then dropped to near zero across the entire maintenance period (Fig. 4 *A*, *Bottom*). These results suggest that representational formats were dynamically transformed during the encoding period, but remained stable across the maintenance period.

In a second analysis, we further tested whether the representational formats were systematically transformed toward the format in the preprobe period (12). We hypothesize that if neural representations gradually change from one time point (t_1) to another time point (t_2), the pattern of neural activities between these time points would become increasingly less similar to the pattern at t_1 and more similar to the pattern at t_2. We defined three temporal clusters as reference points, the early encoding cluster (250 to 770 ms), the late encoding cluster (1,000 to 1,980 ms), and the preprobe time period, i.e., the last second of the maintenance period (9,000 to 10,000 ms). Neural activities in these three clusters of one trial were then correlated with the activities in each encoding and maintenance time point of a trial with the same picture from a different run. This analysis showed that the similarity with the early and late encoding clusters peaked within the time windows of the respective clusters. However, the similarity with the preprobe time period showed no distinctive peak, not even during the last second of the maintenance period (Fig. 4*B*). This result suggests that the representational formats in the early and late encoding clusters did not systematically change to prepare for the upcoming response in our task.

To systematically quantify the dynamics of representations in these three clusters, we divided the encoding and maintenance period into nonoverlapping 1-s time bins (e.g., first time bin, 0 to 1 s; second time bin, 1 to 2 s, etc.). We then applied a linear fit to the similarity values within each time bin, separately for similarities with each of the three clusters. We found that neural representations changed dynamically during the first 2 s of the encoding period (Fig. 4 *B*, *Bottom Right* and *SI Appendix*, *SI Text*). They also changed during the first second of the maintenance

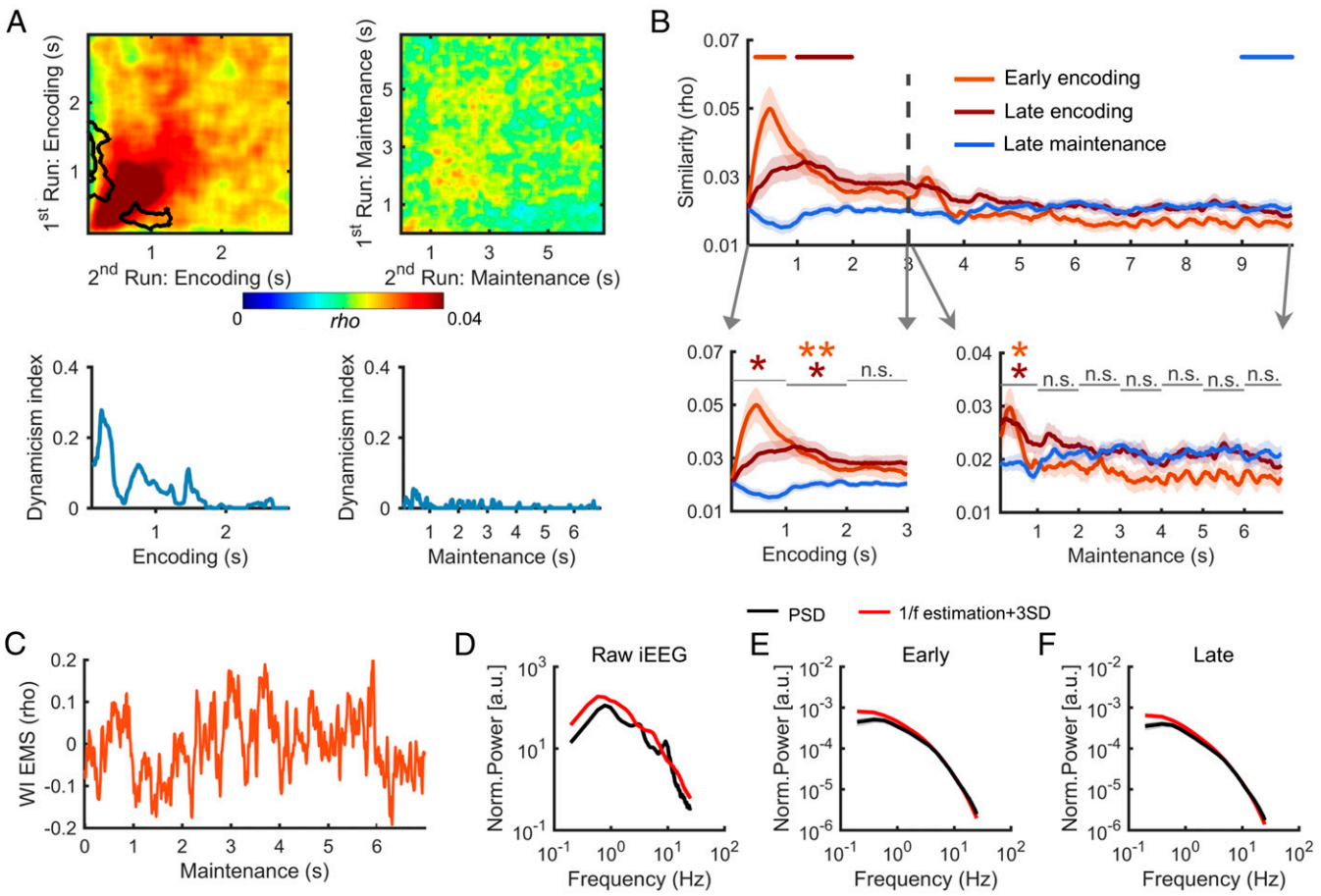

Fig. 4. Stable representational formats during the maintenance period. (*A*) Representational formats were dynamically transformed during the encoding period but remained stable during the maintenance period. Significant off-diagonal similarity reduction was found during the encoding period (*Upper Left*) but not the maintenance period (*Upper Right*). Black clusters show significant dynamicity after cluster-based permutation testing. *Bottom* shows time course of the "dynamicism index" that characterizes dynamic coding across the encoding (*Bottom Left*) and maintenance (*Bottom Right*) periods. Indexes approaching zero indicate stable representations. (*B*) Temporal generalization analysis showing the overall stability of representational formats during the short-term memory delay period. (*Top*) Fluctuations of similarities between each time window with the early (orange) and late (brown) encoding cluster and the last 1 s of the maintenance period (blue). (*Bottom*) Statistical results (FDR corrected) about the linear change of the fluctuations in the *Top* within each 1-s time bin. The colors of the stars correspond to the respective color of lines. (*C*) Exemplary single-trial WI EMS (early encoding cluster), which shows pronounced temporal fluctuations. (*D*) Rhythmic activity in raw iEEG data during the pretask resting period from one participant. IRASA identified a significant oscillatory component (black line) which peaked at 9.38 Hz above the estimated critical threshold of 1/f spectrum (red line: mean + 3 SD). (*E* and *F*) Lack of rhythmicity in EMS time courses for either the early (*E*) or the late cluster (*F*). Shaded areas depict one SEM. *$P < 0.05$; **$P < 0.01$; not significant (n.s.), $P > 0.05$.

period, probably due to the sudden offset of the stimulus (37). Importantly, however, they remained stable during the remaining encoding period and the last 6 s of the maintenance period. As a control, we computed item-specific EMS in the last 6 s of the maintenance period and found significant item-specific EMS for both the early ($t(18) = 2.66$, $P = 0.016$) and the late encoding clusters ($t(18) = 3.61$, $P = 0.002$), ruling out that the stability of EMS was due to an overall lack of item-specific information.

Together, these results show that representational formats were substantially transformed during stimulus encoding. However, except for the first second of the maintenance period, they neither decayed further nor were systematically transformed to match activity prior to probe presentation, providing converging evidence for an overall stability of neural representational formats during VSTM maintenance.

Arrhythmic fluctuations in representational strengths during VSTM. Despite this overall stability, we observed substantial fluctuations of WI EMS (Fig. 4*C*). Further tests revealed nonpersistent item-specific representation across the maintenance period (*SI Appendix*, Fig. S5). We thus tested whether the time courses of WI EMS showed rhythmic fluctuations. To improve the temporal resolution of this analysis, we recomputed EMS using shorter maintenance time windows of 10 ms. We used a well-established method (irregular-resampling auto-spectral analysis [IRASA]) (38) to assess possible rhythmicity of WI EMS time series for both the early and the late encoding clusters. Supporting the validity of the IRASA method in our iEEG dataset, we detected prominent alpha oscillations during the pretask resting period (see Fig. 4*D* for exemplar data from a participant). However, IRASA did not reveal any oscillatory components in the time courses of WI EMS with either the early (Fig. 4*E*) or the late cluster (Fig. 4*F*). Thus, we did not find any evidence for rhythmic fluctuations of stimulus-specific activity during VSTM maintenance.

Hippocampal phase coding of VSTM representations. We next investigated whether the fluctuating reactivation of representational formats was related to the phase of hippocampal low-frequency activity during the maintenance period (18, 39). Only 14 participants with at least one clean hippocampal channel were included in this analysis. Using the Multiple Oscillations Detection Algorithm (MODAL) (40), we identified hippocampal oscillations across a broad frequency range (1 to 10 Hz) across trials and participants (*SI Appendix*, Fig. S6). Following previous work (41, 42), we extracted the phases of hippocampal activity from 1 to 10 Hz and analyzed cross-frequency coupling (*Materials and Methods*). We first replicated previously found cross-frequency coupling effects, showing that the amplitude of high-frequency activities (30 to 100 Hz) was clustered to the phases of low-frequency activity in the hippocampus (after removing one outlier, $t(12) = 4.06$, $P = 0.002$; *SI Appendix*, Fig. S7).

Following a previous study (21), we then tested whether item-specific representations occurred predominantly during specific phases of concurrent hippocampal low-frequency activity. We used the Moore–Rayleigh test (43) to quantify the "representation-to-phase-clustering" value r*, indicating the dependence of WI EMS values on the phase of hippocampal low-frequency activity (Fig. 5 *A* and *B* and *Materials and Methods*). This was done separately for the early and late clusters (see exemplar data of single trials in *SI Appendix*, Fig. S8). We obtained a surrogate r* as the baseline for each trial by circularly shifting the WI EMS values with respect to the concurrent phases of hippocampal low-frequency activity. This analysis showed that the empirical r* values were significantly greater than surrogate r* values for both the early ($t(13) = 2.26$, $P = 0.042$) and the late encoding cluster ($t(13) = 2.95$, $P = 0.011$) (Fig. 5*C*), indicating significant coupling between WI EMS and the phases of hippocampal low-frequency activity.

In a second analysis, we tested for the existence of phase coding—i.e., whether representations of specific items were locked to similar phases across repeated presentations (Fig. 5*D*). To this end, we computed the difference of preferred phases between repetitions of the same items (WI) and between different items (BI) (Fig. 5*E*). We found smaller phase differences for WI pairs than for BI pairs, for both the early ($t(13) = -5.55$, $P < 0.001$) and the late cluster ($t(13) = -4.71$, $P < 0.001$; Fig. 5*F*). These results demonstrate that representations of the same item tend to occur at similar phases of hippocampal low-frequency activity.

Several control analyses revealed that EMS phase coding was not only driven by electrodes that were consistently activated during repetitions of the same item or by activations that were locked to a consistent phase of hippocampal low-frequency activity (*SI Appendix*, *SI Text* and Fig. S9). Finally, all of the above analyses were based on the early and late clusters that were separated by a dip. To ensure our results did not critically depend on how encoding clusters were defined, we redefined these clusters as 500-ms time windows centered around pronounced peaks of the EMS time series. Analyses based on the newly defined clusters replicated the main results reported above (*SI Appendix*, *SI Text* and Fig. S10), indicating that our results are robust to different methods to define the clusters.

Discussion

Can human VSTM achieve stable representations of complex natural images? If so, which representational formats and temporal dynamics characterize VSTM maintenance in humans? The results described here support a different view on these fundamental questions. First, our results provide clear evidence for item-specific representations of natural visual images during an extended maintenance period. Second, we show that these representations originate from two distinct time windows during encoding that contain higher-order visual representations and abstract semantic representations, respectively. Third, the reactivation of these different representational formats was coupled to stimulus-specific phases of hippocampal low-frequency activity during the maintenance period (*SI Appendix*, Fig. S11).

Item-Specific and Stable Representations of Complex Images in VSTM. Previous studies have shown that simple visual stimuli such as colors or the orientation of gratings can be successfully decoded using multivariate classification analyses (14, 15, 44). We extended these findings by showing that representational similarity analysis could be successfully employed to identify item-specific representations of natural visual objects during VSTM maintenance. Using this approach, we observed stimulus-specific representations that were widely distributed across the maintenance period without systematic decline or transformation.

Several factors might have contributed to the relative stability of the representations. First, we used a simple DMS task where participants needed only to maintain one object at a time in VSTM and no interference was introduced. By contrast, some of the studies that reported dynamic coding employed paired associate learning tasks, in which representations were transformed from a perceptual to a response-related code during the maintenance phase (6). Second, iEEG recordings reflect a neural population coding (45). In the presence of heterogeneous activities of single neurons, more sustained stable representations may still be achieved through population coding (12, 13). Third, our results suggest that the employment of an extended encoding period (3 s) and a rather long maintenance period (7 s) might have enabled us to clearly separate different processing stages and observe stable representational formats across the VSTM maintenance period. Previous studies have shown that, during visual processing, brain activities gradually and progressively changed from representing low-level visual information to representing superordinate category information within the first few hundred milliseconds (9, 46, 47). In the current study, we found that neural representations underwent substantial transformations

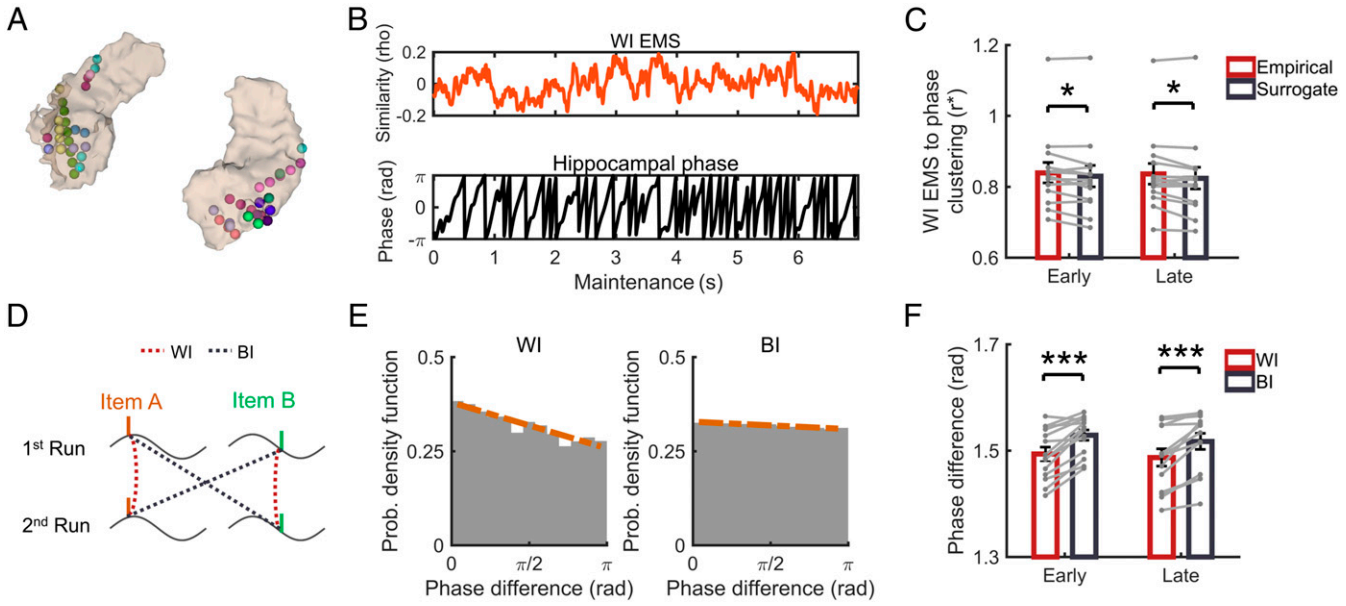

Fig. 5. Hippocampal phase coding during VSTM maintenance. (*A*) Depiction of hippocampal channels from 14 participants. Spheres with the same color are from the same participant. (*B*) Exemplar data from a single trial showing the time course of WI EMS values from the early encoding cluster (*Top*) and the concurrent phase of hippocampus low-frequency activity (1 to 10 Hz) (*Bottom*). (*C*) WI EMS clustering to the phase of hippocampal low-frequency activity for both the early and the late encoding cluster. (*D*) Schematic depiction of phase coding. The same item from different trials should be locked to similar phases of hippocampal low-frequency activity, whereas different items should be locked to different phases. As a result, the phase differences between WI pairs should be smaller than those of BI pairs. (*E*) Probability density function of phase differences for WI pairs and BI pairs for one participant. (*F*) Group-averaged phase differences for WI and BI pairs in the early and the late cluster. Each gray dot indicates one participant. Error bars reflect one SEM. $*P < 0.05$; $***P < 0.001$.

during the first 2 s of the encoding period and were overall stable afterward, including the extended maintenance period. This suggests that the encoded representations went through an extended transformation process before reaching a stable state for maintenance. As a large body of previous studies used only a brief stimulus presentation period (e.g., less than 1 s) (6, 13), the dynamics of early maintenance in these studies might reflect this transformation of representational formats.

Higher-Order and Abstract Representational Formats Were Maintained in VSTM. By linking the neural representations to visual representations in a DNN, we found that neural representations in an early encoding cluster (250 to 770 ms after stimulus onset) were correlated with deeper but not early layers of a visual DNN. These representations correspond to higher-order visual processing steps that eventually support the extraction of object categories. Our results contribute to a growing body of literature indicating that representations within various layers of DNNs match representations in the ventral visual processing stream (30, 32, 48). Neural activities in the late encoding cluster were not associated with DNN representations in any layers but could be explained by semantic similarities. This suggests that neural representations in the late cluster contain fewer visual components and instead reflect abstract long-term knowledge. Consistently, after regressing out the perceptual contributions to the semantic similarity, only the representations in the late cluster were correlated with the abstract semantic similarity patterns. Notably, the lack of association between the later cluster representation and the visual DNN might be due to the choice of a specific DNN model, as the AlexNet is a pure feed-forward network. Future studies could examine the nature of representation in more complex and arguably more biologically realistic DNNs, such as recurrent models (49) or deep reinforcement learning networks (50).

Why were the abstract semantic representational formats more robustly maintained than the visual representations, with the encoding–maintenance similarity in the late encoding cluster showing stronger item-specificity? One possibility is that maintaining these representations reduces memory load. Moreover, they might reflect information that integrates bottom–up external inputs and top–down long-term knowledge (51), and this long-term knowledge has been shown to contribute to short-term memory maintenance of complex images (52). When only simple visual stimuli, such as colors, orientations, or contrasts, are maintained, the early visual representations from the primary visual cortex may need to be faithfully maintained (15, 26). Future work should collect more samples from different brain regions and use "human super-EEG" analysis (53) to provide a better characterization of the representational nature of VSTM at a higher spatial resolution.

Studies on long-term memory encoding and retrieval further support the idea that representations from late processing stages can be stably remembered. For example, scalp-EEG studies suggest that the stability of representations across repetitions in a late time window (500 ms after stimulus onset) supports subsequent memory (54, 55). Moreover, representations from a late encoding phase (~1,000 to 2,000 ms after stimulus onset) were found to be reinstated during successful memory retrieval (56). Another recent study demonstrated that stimulus-specific activity from a late encoding stage was reactivated during offline periods and sleep, thereby supporting long-term memory consolidation (57). Finally, functional magnetic resonance imaging (fMRI) studies revealed that pronounced item-specific memory reactivations during retrieval occurred in the parietal lobe but not in the ventral visual cortex (58). Together, these results suggest that late, abstract, and transformed representational formats support stable short-term maintenance, as well as long-term storage and reactivation.

VSTM Representational Formats Reactivated at Specific Phases of Hippocampal Low-Frequency Activity. In contrast to the persistent activity model of VSTM, the occurrence of sparse and irregular

neuronal bursts has been posited as a mechanism to optimize information transferring (59). Between these bursts, working memory representations are assumed to be maintained "silently" by spiking-induced changes in synaptic weights (60, 11). The spikes couple with slow oscillations to refresh synaptic weight changes (60), and the refresh rate may account for the limited capacity of short-term memory (61). Compared with the persistent activity model of VSTM, this dynamic coding scheme is more robust to interference and disruption, consumes less energy, and enables the simultaneous maintenance of multiple items (62). Consistently, we found that reactivations of item representations were not persistent, but showed pronounced irregular fluctuations. These fluctuations were locked to specific phases of hippocampal low-frequency activity. These results support the idea that VSTM does not entail persistent activation of item-specific representations. However, the electrodes we used could not record single-unit activities, and existing studies have revealed persistent activities in single neurons during VSTM (63, 64).

Some theoretical models suggest that items in working memory are coded by neural assemblies which are synchronized in the gamma frequency band and locked to the phase of low-frequency oscillations, in particular in the theta band (19, 39). Human intracranial EEG studies have found support for this model by demonstrating that high-frequency activity is indeed nested within phases of low-frequency activity in hippocampus (18). Recent work also found that letter selective activity is coupled to theta phases during working memory maintenance (23). The current study extends this work by directly showing hippocampal phase coding of stimulus-specific representational formats during short-term memory maintenance. More importantly, the same item tended to lock to a similar phase of hippocampal low-frequency activity during VSTM, using a similar coding scheme to that in a recent iEEG study during virtual navigation (21). Interestingly, unlike the rhythmic reactivation that is repeated on every cycle of low-frequency oscillations (65), we did not find evidence for rhythmic reactivations, consistent with a monkey multiunit recording study which showed that feature-specific information reoccurred in discrete and irregular bursts (7). Corroborating this finding, a recent study found that despite the lack of low-frequency oscillations in the hippocampus of bats, the spiking was locked to specific phases of hippocampal activity (66).

Exactly how such irregular reactivation of item-specific representational formats at specific phases of hippocampal low-frequency activity could support VSTM maintenance remains to be further examined. One possibility is that the arrhythmic reactivation reflects the intermittent interaction between short-term memory and long-term knowledge, which is known to rely on the hippocampus (67). In addition, when only a single item is maintained, it might not be necessary to rigorously refresh its representation in each oscillatory cycle, especially when a higher-order abstract representational format is maintained. Future studies should examine how the nature of stimulus and working memory load could affect the oscillation of VSTM representations and its coupling with the phase of hippocampal low-frequency activity.

To summarize, our study suggests that stable item-specific VSTM maintenance of natural images is achieved via reactivations of multiple higher-order and abstract representational formats that are phase locked to hippocampal activity. These results contribute to the development of a more rigorous, mechanistic understanding of VSTM.

Materials and Methods

Participants. Nineteen patients (27.9 ± 7.1 y, 5 female) with medically intractable epilepsy participated in the study. All patients were implanted with depth electrodes for diagnostic purposes using a stereotactic procedure with the Leksell frame. Each depth electrode (0.8 mm in diameter) had either 12 or 16 contacts (channels) that were 1.5 mm apart, with a contact length of 2 mm. Recordings were performed at 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 Institutional Review Board at Xuanwu Hospital. All participants gave written informed consent.

Experimental Design. Fifty-six pictures from four categories (i.e., familiar fruits, animals, electrical devices, and furniture) and 112 two-character Chinese verbs were used in this study. There are 14 pictures in each category. Each picture was paired with two different cue words across two runs. Associations were randomized across participants. For each picture, a very similar picture was also selected, which was used as a lure in the probe phase of the experiment (e.g., Fig. 1*A*).

A DMS task was used in the study. Each trial consisted of three phases, i.e., an encoding phase, a maintenance phase, and a probe phase. During the encoding phase, a word–picture pair was presented at the center of the screen for 3 s. Participants were instructed to pay attention to the details of the picture and memorize the associations. During the maintenance phase, the picture disappeared while the cue word remained on the screen, and participants were asked to maintain the picture as vividly as possible for 7 s. During the probe phase, a picture appeared on the screen and participants were asked to indicate whether it was the same as the target picture, by pressing one of two buttons within 2 s. The next trial started after 0.3 s of fixation, followed by 0.8 to 1.2 s of a blank screen. For half of the trials in a run, the probe picture was the same as the target picture (match trials); for the other half, a very similar lure picture was presented (nonmatch trials) to encourage participants to pay attention to the visual details of the target picture. Following each run of the VSTM task, there was a long-term memory task, during which subjects were shown the cue words and asked to recall the category of the associated picture. The current study focuses only on the VSTM task. The cue on the screen was presented to encourage participants to remember the association between the cue and the picture and, meanwhile, reduce the processing load and prevent the distraction from processing the picture during the maintenance period.

We randomly selected 14 target pictures for each session. Each session consisted of two runs. In the first run, each picture was paired with one of a randomly selected set of 14 words, resulting in 14 unique word–picture pairs. In the second run, we randomly selected 14 new words and paired them with the same pictures as in the first run, resulting again in 14 word–picture pairs. Each pair was repeated three times within a run and the lag between repetitions was 7 to 12 trials. This allowed us to calculate the representational similarity of the same pictures across the two runs and, meanwhile, to avoid the possible confound of cue words on WI (similarity between pairs with the same picture but different words) versus BI (similarity between pairs with different pictures and words) pattern similarity comparisons (58). Participants finished 2 to 4 sessions (3.36 ± 0.83 sessions) of the VSTM task.

Intracranial EEG Recordings and Analyses. iEEG data were recorded using amplifiers from Brain Products GmbH, NeuroScan (Compumedics Limited) or the Nicolet electroencephalogram system (Alliance Biomedica Pvt Ltd.), with sampling rates of 2,500, 2,000, and 2,048 Hz, respectively. Online recording data were referenced to a common contact placed subcutaneously, which was simultaneously recorded with the depth electrodes. During offline preprocessing, channels that were within the epileptic loci or severely contaminated by epileptic activity were removed from further analyses. To preserve the relative activation patterns across channels when examining distributed item-specific representations and conserve neural activities across a broad frequency band (e.g., 1 to 100 Hz), we rereferenced all raw iEEG data to the average activity across all clean channels. Notably, when using the white-matter rereferencing scheme (i.e., the average activity of channels in white matter), we obtained a highly similar pattern of item-specific representations (*SI Appendix*, Fig. S12).

Data analysis was performed by EEGlab (https://sccn.ucsd.edu/eeglab/) and the Fieldtrip toolbox (68) implemented in Matlab (MathWorks Inc.), as well as using in-house Matlab code. To remove 50 Hz line noise, data were band-stop filtered at 50 ± 2 Hz and its harmonic frequencies. The filtering was done using fourth-order Butterworth filters. Trials contaminated by artifacts were identified based on visual inspection and excluded from further analyses. Each trial was epoched from 4 s before stimulus onset to 4 s after the end of the maintenance phase (i.e., −4 s to 14 s with respect to stimulus onset). This long trial duration was used to eliminate edge effects in time-frequency transformation, and we focused only on the results between −500 ms and 10 s. Time-frequency transformation was performed within each trial using the complex Morlet wavelets. Wavelet kernels with six cycles were used to extract the power at each frequency from 1 to 100 Hz in 1-Hz steps. All power spectral data were down-sampled to 100 Hz after time-frequency transformation.

This resulted in a 100 (frequencies) × 1,050 (time points) time-frequency matrix for each channel and each trial. Afterward, we normalized the power separately for each frequency and each channel by subtracting the mean power of all trials and then dividing by the SD of the power across trials.

Channel Localization. Channel locations were identified by coregistering a postimplantation computed tomography (CT) image to a preimplantation MRI, which was afterward normalized to MNI space in Statistical Parametric Mapping (SPM)12. The anatomical locations of channels were then identified and plotted using 3D slicer (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) and identified the closest cortical or subcortical label for each channel in each patient. Channels in subcortical regions, including hippocampus, were visually verified in each patient's native anatomical space. Across all patients, 529 clean channels were included in the analyses (mean number of clean channels per patient, 27.8; SD, 12.4).

Representational Similarity Analysis. Representational similarity was estimated between trials from two consecutive runs within a session. The WI similarity was calculated between trials using the same picture but paired with different cue words, so that the pattern similarity was not confounded by identical cue words. The BI similarity was calculated between trials using different pictures and different cues. The WI and BI pairs were matched in terms of lags and calculated only between correct trials.

We used a sliding time window approach to calculate the representational similarity. A 200-ms sliding time window was used, with incremental steps of 10 ms (i.e., 190 ms overlap between two consecutive windows). Power spectral values were first averaged across time points within each sliding window of 200 ms, separately for each channel and frequency. Forty-three frequencies in the range between 2 and 100 Hz were then extracted in steps of 1 Hz between 2 and 29 Hz and in steps of 5 Hz between 30 and 100 Hz, as in a previous study (34).

To identify item-specific representations across channels and frequencies (i.e., global RSA), the power spectral values of all channels and frequencies within each time window were vectorized. The representational similarity between two trials was then obtained by calculating the Spearman's correlation between the features of the two trials. All correlation values were Fisher Z-transformed before further statistical analyses.

Notably, global RSA can be calculated either separately for encoding and maintenance periods or between encoding and maintenance periods. To examine item-specific representations during the maintenance period, we calculated MMS. Likewise, item-specific representations during the encoding period were identified by calculating EES. To examine which processing steps during the encoding period contained representational formats that contribute to stimulus-specific maintenance, we calculated EMS. In this analysis, we correlated neural activities in each encoding window with activities in each maintenance window. For all these analyses, item-specific representations were obtained by contrasting the WI versus BI similarity.

Nonparametric Cluster-Based Permutation Test. A nonparametric statistical test based on the cluster-level permutation was implemented in Matlab to correct for multiple comparisons (69). Specifically, for individual time windows, statistical tests were performed between conditions (e.g., WI vs. BI), and time windows with statistical values larger than a threshold ($P = 0.05$) were selected and combined into contiguous clusters on the basis of adjacency. Cluster-level statistics were computed by taking the sum of the t values within a cluster. The distribution of cluster-level statistics under the null hypothesis was constructed by randomly permuting condition labels (e.g., WI vs. BI) for 1,000 times, and the maximum cluster-level statistic in each permutation was extracted. If no time point showed a significant t value for a given surrogate, a value of 0 was assigned for that surrogate. The nonparametric statistical significance was obtained by calculating the proportion of surrogates within the permutation distribution that exceeded the observed cluster-level statistics.

Linking Neural Representations to Deep Neural Network Representations and Semantic Representations. We characterized the representational formats within different encoding time periods via a DNN, AlexNet (35) and a well-established semantic model, Directional Skip-Gram (36). The AlexNet implements a network for object identification, i.e., the assignment of object labels to visual stimuli (*SI Appendix, SI Text* and Fig. S2). It was pretrained using the ImageNet dataset (70). The AlexNet consists of eight layers, five convolutional layers and three fully connected layers, which simulate the hierarchical structure of neurons along the ventral visual stream. To verify that the pretrained DNN can be applied to the pictures in our experiment, we classified all target pictures using the DNN and only pictures whose labels were correctly identified were included. We found that 47 of 56 pictures were successfully classified (i.e., their labels were among the top five labels provided by the DNN). For each of these 47 images, we then extracted the simulated activations from each layer of DNN, which served as features for RSA. We calculated Spearman's correlations between the DNN features of every pair of pictures, resulting in a 47 × 47 similarity matrix in each layer.

The semantic similarity matrix was calculated based on the labels of the pictures, which were generated by five independent raters (*SI Appendix, SI Text*). For each picture label, we extracted the semantic features (i.e., word vectors) from a well-trained Chinese word embedding model, Directional Skip-Gram (36). In this model, each word vector consists of 200 values, with each value indicating the meaning of a picture label in one semantic dimension. The semantic similarities of the 56 picture labels were accessed by calculating the cosine similarity of these word vectors, resulting in a 56 × 56 semantic similarity matrix. To obtain the abstract semantic similarity matrix, we performed a stepwise regression in which we regressed the similarity matrices of each DNN layer from the semantic similarity matrix. Note that the semantic similarity was not reversely regressed out from the DNN. This is because the DNN is specifically designed to provide labels to pictures, regressing the semantic representations out from the DNN could effectively deconstruct its organization.

To create the neural similarity matrix, we correlated all pairs of pictures across frequencies and channels, using Spearman's correlations. This was done in sliding time windows of 200 ms, with a step size of 10 ms. In this analysis, data across repetitions of the same pictures were first averaged. To remove potential confounds of commonly evoked power across all trials on neural activation pattern, we normalized the power spectral data across trials during each time window for each frequency and each channel. Notably, this normalization does not significantly change the item-specific RSA values ($P > 0.25$).

Finally, the DNN similarity matrices, the semantic similarity matrix, and the abstract semantic matrix were correlated with the neural similarity matrix via Spearman's correlation in each encoding time window. The correlation values were then Fisher Z-transformed for further statistical analysis.

To determine the significance of the correlations, the neural similarity matrices were recalculated after randomly shuffling the labels of pictures. The surrogate neural similarity matrices were then correlated with the similarity matrix in each layer of the DNN, the semantic matrix as well as the abstract semantic matrix separately. This was done 1,000 times. 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.

Analysis of the Stability of Representational Formats Across Time. Previous work argued that if neural representations are transformed from one time point (t_1) to another time point (t_2), the off-diagonal similarity would be significantly reduced (13). Accordingly, we quantified the dynamicity of this transformation between t_1 and t_2 by comparing the on-diagonal similarity values ($r(t_1, t_1)$, $r(t_2, t_2)$, etc.) with the off-diagonal similarity values ($r(t_1, t_2)$). The dynamicity score (*dyna*) was obtained using the following equation:

$$dyna(t_1, t_2) = \begin{cases} 1, & if\ r(t_1, t_2) < r(t_1, t_1) \cap r(t_1, t_2) < r(t_2, t_2) \\ 0, & otherwise \end{cases}.$$

If $r(t_1, t_1)$ and $r(t_2, t_2)$ are both significantly greater than $r(t_1, t_2)$ across participants, then the *dyna* between t_1 and t_2 is 1, indicating that the neural representational formats were transformed from t_1 to t_2. Otherwise the *dyna* is 0. It should be noted that the correlation was not conducted between different time points within the same trial, in which $r(t_1, t_1)$ or $r(t_2, t_2)$ is always 1. Instead, the similarity values reflect correlations between neural activities of trials from two independent runs that share the same picture in the current study, so that the off-diagonal reduction is nontrivial. Clusters showing significant dynamicity (with *dyna* = 1) were corrected for multiple comparisons using a cluster-based permutation test, in which the null distribution was obtained by randomly shuffling the on-diagonal similarity values (e.g., $r(t_1, t_1)$, $r(t_2, t_2)$) with the off-diagonal similarity values (e.g., $r(t_1, t_2)$) for 1,000 times. To further characterize the representational dynamicity over time, we averaged the *dyna* across the two temporal dimensions, resulting in the dynamicism index ($di(t)$) over time by using the equation

$$di(t) = \frac{1}{2T}\left(\sum_{t_1}[dyna\ (t_1, t)] + \sum_{t_2}[dyna\ (t_2, t)]\right),$$

where T denotes the number of the time windows during the encoding or maintenance period. The larger the *di* is, the greater the representational transformation.

Assessing Rhythmicity of VSTM Representations. IRASA is a well-established method for disentangling the oscillatory component from the fractal component (i.e., 1/f background activity) of a signal (38). We applied IRASA to examine the rhythmicity of encoding–maintenance similarity during maintenance using the following steps. First, data were segmented into a continuous period of 5,250 ms (i.e., 75% of a trial duration, using sliding windows with 500 ms as a step). Second, the auto-spectral density was calculated by performing the fast Fourier transformation (FFT) on each of the segments with frequencies ranging from 0.01 to 25 Hz. Third, for each segment, we irregularly resampled the data with resampling rates h and 1/h (h = 1.1:0.05:1.9) to up-sample and down-sample the data, respectively. The auto-spectral density was computed again for the resampled data by FFT. The fractal component distribution (1/f) across frequencies was estimated by extracting the median from the resampled auto-spectral density separately for each time point of a segment. Fourth, the oscillatory components were obtained by subtracting the fractal component from the original auto-spectral density of each segment. Then, we averaged the oscillatory components and the fractal components across segments and trials. Notably, to define the critical values for the significance test of oscillatory components, we set the values of averaged fractal component plus three SDs as the threshold, following a previous study (71). Distinctive peaks at any frequency with oscillatory power above the threshold were defined as evidence for the presence of oscillatory activities.

Analysis of Clustering of VSTM Representations to the Phase of Hippocampal Low-Frequency Activity. The hippocampal low-frequency bands for the extraction of phase values were determined by a data-driven analysis, MODAL (40), which allowed us to detect oscillatory components in single trials. Specifically, for each trial, the power at each frequency was obtained by time-frequency transformation using Morlet wavelets (cycle number = 6). Frequency bands with power surpassing the 1/f function were defined as oscillatory components. Next, MODAL filtered the raw iEEG data using the identified frequency band and applied frequency sliding to each time point (72). We then summarized dominant frequencies and the duration when they occurred across trials and participants. This analysis showed that oscillatory activities were detected in a broad frequency range between 1 and 10 Hz across all trials and participants, with peaks at 3 and 9 Hz (*SI Appendix*, Fig. S6). The overall duration of oscillatory activity in a single frequency bin (1 Hz) was no more than 30% of time points on average across all participants. The selection of this band was further confirmed by increased neural activities in this frequency band during the maintenance period compared to that during the prestimulus baseline period (300 to 500 ms prior to stimulus onset) (*SI Appendix*, Fig. S6). Then, a bandpass filter (1 to 10 Hz) was applied to the iEEG data of hippocampal channels. We performed a Hilbert transformation on the filtered data to extract phase series.

The WI EMS between two trials with the same item was averaged across the time windows in the early or the late encoding cluster, resulting in a time course of WI EMS across the maintenance period of one trial. Coupling between WI EMS and concurrent phases of hippocampal low-frequency activity (1 to 10 Hz) during the maintenance period was examined by the Moore–Rayleigh test (43). As a nonparametric extension of the Rayleigh test, the Moore–Rayleigh test ranks and weighs the phase according to the magnitude of the WI EMS. It then outputs a clustering value (r*) and a preferred phase separately for the early and the late cluster of each trial. We obtained the distribution of surrogate r* values by circularly shifting the amplitude values with regard to the concurrent phases and recomputing the clustering value. This was done 100 times. We then averaged the 100 surrogate r* values and obtained one averaged surrogate r* value for each cluster of a trial. These clustering values were averaged across trials. We then performed a paired *t* test between empirical r* values and surrogate r* values across participants.

Data Availability. Intracranial EEG data and materials are available at the Open Science Framework: https://osf.io/yqftv/.

ACKNOWLEDGMENTS. We thank Elkan Akyürek, Mark Stokes, Bryan Strange, and Johannes Sarnthein for reviewing and providing insightful feedback on our paper and Yuntao Zhou and Youyan Li for help with the data collection. G.X. received grants from the National Science Foundation of China (31730038), the China–Israel collaborative research grant (NSFC 31861143040), and the Guangdong Pearl River Talents Plan Innovative and Entrepreneurial Team Grant 2016ZT06S220. N.A. received funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), Projektnummer 316803389–SFB 1280, via Projektnummer 122679504–SFB 874, and via DFG Grant AX 82/3. H.Z. received funding by DFG Projektnummer 429281110.

1. A. Baddeley, Working memory: Looking back and looking forward. *Nat. Rev. Neurosci.* **4**, 829–839 (2003).
2. M. D'Esposito, From cognitive to neural models of working memory. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **362**, 761–772 (2007).
3. J. M. Fuster, G. E. Alexander, Neuron activity related to short-term memory. *Science* **173**, 652–654 (1971).
4. S. Funahashi, C. J. Bruce, P. S. Goldman-Rakic, Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *J. Neurophysiol.* **61**, 331–349 (1989).
5. E. M. Meyers, D. J. Freedman, G. Kreiman, E. K. Miller, T. Poggio, Dynamic population coding of category information in inferior temporal and prefrontal cortex. *J. Neurophysiol.* **100**, 1407–1419 (2008).
6. M. G. Stokes *et al*., Dynamic coding for cognitive control in prefrontal cortex. *Neuron* **78**, 364–375 (2013).
7. M. Lundqvist *et al*., Gamma and beta bursts underlie working memory. *Neuron* **90**, 152–164 (2016).
8. K. K. Sreenivasan, M. D'Esposito, The what, where and how of delay activity. *Nat. Rev. Neurosci.* **20**, 466–481 (2019).
9. R. M. Cichy, D. Pantazis, A. Oliva, Resolving human object recognition in space and time. *Nat. Neurosci.* **17**, 455–462 (2014).
10. M. A. Spiegel, D. Koester, T. Schack, The functional role of working memory in the (re-)planning and execution of grasping movements. *J. Exp. Psychol. Hum. Percept. Perform.* **39**, 1326–1339 (2013).
11. M. G. Stokes, 'Activity-silent' working memory in prefrontal cortex: A dynamic coding framework. *Trends Cogn. Sci.* **19**, 394–405 (2015).
12. J. D. Murray *et al*., Stable population coding for working memory coexists with heterogeneous neural dynamics in prefrontal cortex. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 394–399 (2017).
13. E. Spaak, K. Watanabe, S. Funahashi, M. G. Stokes, Stable and dynamic coding for working memory in primate prefrontal cortex. *J. Neurosci.* **37**, 6503–6516 (2017).
14. S. A. Harrison, F. Tong, Decoding reveals the contents of visual working memory in early visual areas. *Nature* **458**, 632–635 (2009).
15. J. T. Serences, E. F. Ester, E. K. Vogel, E. Awh, Stimulus-specific delay activity in human primary visual cortex. *Psychol. Sci.* **20**, 207–214 (2009).
16. K. K. Sreenivasan, J. Vytlacil, M. D'Esposito, Distributed and dynamic storage of working memory stimulus information in extrastriate cortex. *J. Cogn. Neurosci.* **26**, 1141–1153 (2014).
17. N. S. Rose *et al*., Reactivation of latent working memories with transcranial magnetic stimulation. *Science* **354**, 1136–1139 (2016).
18. N. Axmacher *et al*., Cross-frequency coupling supports multi-item working memory in the human hippocampus. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 3228–3233 (2010).
19. O. Jensen, J. E. Lisman, Hippocampal sequence-encoding driven by a cortical multi-item working memory buffer. *Trends Neurosci.* **28**, 67–72 (2005).
20. C. Poch, L. Fuentemilla, G. R. Barnes, E. Düzel, Hippocampal theta-phase modulation of replay correlates with configural-relational short-term memory performance. *J. Neurosci.* **31**, 7038–7042 (2011).
21. L. Kunz *et al*., Hippocampal theta phases organize the reactivation of large-scale electrophysiological representations during goal-directed navigation. *Sci. Adv.* **5**, eaav8192 (2019).
22. A. J. Watrous, J. Fell, A. D. Ekstrom, N. Axmacher, More than spikes: Common oscillatory mechanisms for content specific neural representations during perception and memory. *Curr. Opin. Neurobiol.* **31**, 33–39 (2015).
23. A. Bahramisharif, O. Jensen, J. Jacobs, J. Lisman, Serial representation of items during working memory maintenance at letter-selective cortical sites. *PLoS Biol.* **16**, e2003805 (2018).
24. S. Ten Oever, P. De Weerd, A. T. Sack, Phase-dependent amplification of working memory content and performance. *Nat. Commun.* **11**, 1832 (2020).
25. E. F. Ester, T. C. Sprague, J. T. Serences, Parietal and frontal cortex encode stimulus-specific mnemonic representations during visual working memory. *Neuron* **87**, 893–905 (2015).
26. Q. Yu, W. M. Shim, Occipital, parietal, and frontal cortices selectively maintain task-relevant features of multi-feature objects in visual working memory. *Neuroimage* **157**, 97–107 (2017).
27. B.-I. Oh, Y.-J. Kim, M.-S. Kang, Ensemble representations reveal distinct neural coding of visual working memory. *Nat. Commun.* **10**, 5665 (2019).
28. M. Villena-González, V. López, E. Rodríguez, Orienting attention to visual or verbal/auditory imagery differentially impairs the processing of visual stimuli. *Neuroimage* **132**, 71–78 (2016).
29. A. Clarke, B. J. Devereux, L. K. Tyler, Oscillatory dynamics of perceptual to conceptual transformations in the ventral visual pathway. *J. Cogn. Neurosci.* **30**, 1590–1605 (2018).

30. R. M. Cichy, A. Khosla, D. Pantazis, A. Torralba, A. Oliva, Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Sci. Rep.* **6**, 27755 (2016).
31. P. Bao, L. She, M. McGill, D. Y. Tsao, A map of object space in primate inferotemporal cortex. *Nature* **583**, 103–108 (2020).
32. U. Güçlü, M. A. J. van Gerven, Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *J. Neurosci.* **35**, 10005–10014 (2015).
33. S. Xie, D. Kaiser, R. M. Cichy, Visual imagery and perception share neural representations in the alpha frequency band. *Curr. Biol.* **30**, 2621–2627.e5 (2020).
34. B. P. Staresina *et al*., Hippocampal pattern completion is linked to gamma power increases and alpha power decreases during recollection. *eLife* **5**, e17397 (2016).
35. A. Krizhevsky, I. Sutskever, G. E. Hinton, ImageNet classification with deep convolutional neural networks. *Commun. ACM* **60**, 84–90 (2012).
36. Y. Song, S. Shi, J. Li, H. Zhang, "Directional skip-gram: Explicitly distinguishing left and right context for word embeddings" in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (Short Papers)*, M. Walker, H. Ji, A. Stent, Eds. (Association for Computational Linguistics, 2018), pp. 175–180, vol. 2.
37. M. E. van de Nieuwenhuijzen, E. W. P. van den Borne, O. Jensen, M. A. J. van Gerven, Spatiotemporal dynamics of cortical representations during and after stimulus presentation. *Front. Syst. Neurosci.* **10**, 42 (2016).
38. H. Wen, Z. Liu, Separating fractal and oscillatory components in the power spectrum of neurophysiological signal. *Brain Topogr.* **29**, 13–26 (2016).
39. J. E. Lisman, O. Jensen, The θ-γ neural code. *Neuron* **77**, 1002–1016 (2013).
40. A. J. Watrous, J. Miller, S. E. Qasim, I. Fried, J. Jacobs, Phase-tuned neuronal firing encodes human contextual representations for navigational goals. *eLife* **7**, e32554 (2018).
41. J. Aru *et al*., Untangling cross-frequency coupling in neuroscience. *Curr. Opin. Neurobiol.* **31**, 51–61 (2015).
42. A. C. Heusser, D. Poeppel, Y. Ezzyat, L. Davachi, Episodic sequence memory is supported by a theta-gamma phase code. *Nat. Neurosci.* **19**, 1374–1380 (2016).
43. B. R. Moore, A modification of the Rayleigh test for vector data. *Biometrika* **67**, 175–180 (1980).
44. T. B. Christophel, P. C. Klink, B. Spitzer, P. R. Roelfsema, J.-D. Haynes, The distributed nature of working memory. *Trends Cogn. Sci.* **21**, 111–124 (2017).
45. L. Kunz *et al*., Mesoscopic neural representations in spatial navigation. *Trends Cogn. Sci.* **23**, 615–630 (2019).
46. J. J. DiCarlo, D. Zoccolan, N. C. Rust, How does the brain solve visual object recognition? *Neuron* **73**, 415–434 (2012).
47. Z. Kourtzi, C. E. Connor, Neural representations for object perception: Structure, category, and adaptive coding. *Annu. Rev. Neurosci.* **34**, 45–67 (2011).
48. I. Kuzovkin *et al*., Activations of deep convolutional neural networks are aligned with gamma band activity of human visual cortex. *Commun. Biol.* **1**, 107 (2018).
49. T. C. Kietzmann *et al*., Recurrence is required to capture the representational dynamics of the human visual system. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 21854–21863 (2019).
50. M. Botvinick, J. X. Wang, W. Dabney, K. J. Miller, Z. Kurth-Nelson, Deep reinforcement learning and its neuroscientific implications. *Neuron* **107**, 603–616 (2020).
51. E. A. Hirshorn *et al*., Decoding and disrupting left midfusiform gyrus activity during word reading. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 8162–8167 (2016).
52. R. E. O'Donnell, A. Clement, J. R. Brockmole, Semantic and functional relationships among objects increase the capacity of visual working memory. *J. Exp. Psychol. Learn. Mem. Cogn.* **44**, 1151–1158 (2018).
53. L. L. W. Owen *et al*., A Gaussian process model of human electrocorticographic data. *Cereb. Cortex* **30**, 5333–5345 (2020).
54. K. Feng *et al*., Spaced learning enhances episodic memory by increasing neural pattern similarity across repetitions. *J. Neurosci.* **39**, 5351–5360 (2019).
55. Y. Lu, C. Wang, C. Chen, G. Xue, Spatiotemporal neural pattern similarity supports episodic memory. *Curr. Biol.* **25**, 780–785 (2015).
56. R. B. Yaffe *et al*., Reinstatement of distributed cortical oscillations occurs with precise spatiotemporal dynamics during successful memory retrieval. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 18727–18732 (2014).
57. H. Zhang, J. Fell, N. Axmacher, Electrophysiological mechanisms of human memory consolidation. *Nat. Commun.* **9**, 4103 (2018).
58. X. Xiao *et al*., Transformed neural pattern reinstatement during episodic memory retrieval. *J. Neurosci.* **37**, 2986–2998 (2017).
59. R. Naud, H. Sprekeler, Sparse bursts optimize information transmission in a multiplexed neural code. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E6329–E6338 (2018).
60. G. Mongillo, O. Barak, M. Tsodyks, Synaptic theory of working memory. *Science* **319**, 1543–1546 (2008).
61. J. E. Lisman, M. A. Idiart, Storage of 7 +/- 2 short-term memories in oscillatory subcycles. *Science* **267**, 1512–1515 (1995).
62. E. K. Miller, M. Lundqvist, A. M. Bastos, Working memory 2.0. *Neuron* **100**, 463–475 (2018).
63. J. Kamiński *et al*., Persistently active neurons in human medial frontal and medial temporal lobe support working memory. *Nat. Neurosci.* **20**, 590–601 (2017).
64. C. Constantinidis *et al*., Persistent spiking activity underlies working memory. *J. Neurosci.* **38**, 7020–7028 (2018).
65. O. Jensen, Information transfer between rhythmically coupled networks: Reading the hippocampal phase code. *Neural Comput.* **13**, 2743–2761 (2001).
66. T. Eliav *et al*., Nonoscillatory phase coding and synchronization in the bat hippocampal formation. *Cell* **175**, 1119–1130.e15 (2018).
67. N. Axmacher, S. Lenz, S. Haupt, C. E. Elger, J. Fell, Electrophysiological signature of working and long-term memory interaction in the human hippocampus. *Eur. J. Neurosci.* **31**, 177–188 (2010).
68. R. Oostenveld, P. Fries, E. Maris, J.-M. Schoffelen, FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Comput. Intell. Neurosci.* **2011**, 156869 (2011).
69. E. Maris, R. Oostenveld, Nonparametric statistical testing of EEG- and MEG-data. *J. Neurosci. Methods* **164**, 177–190 (2007).
70. J. Deng *et al*., "ImageNet: A large-scale hierarchical image database" in *2009 IEEE Conference on Computer Vision and Pattern Recognition* (Institute of Electrical and Electronics Engineers, 2009), pp. 248–255.
71. R. F. Helfrich *et al*., Neural mechanisms of sustained attention are rhythmic. *Neuron* **99**, 854–865.e5 (2018).
72. M. X. Cohen, Fluctuations in oscillation frequency control spike timing and coordinate neural networks. *J. Neurosci.* **34**, 8988–8998 (2014).