nature communications

Article | Article in Press

https://doi.org/10.1038/s41467-026-78895-x

# Neural subspaces encode sequential working memory without detectable neural sequences

Received: 10 November 2025

Accepted: 7 October 2026

Published online: 23 October 2026

Check for updates

Aniol Santo-Angles[1,2] ✉, Jiarui Yang[1], Ying Zhou[1], Wing Kwan Hannah Chu[1], Grace W. Lindsay[3] & Kartik K. Sreenivasan[1,4]

The neural mechanisms of multiple-item working memory are not well understood. In the current study, we address two competing hypotheses about the neural basis of sequential working memory: neural subspaces versus neural sequences. Using broadband MEG data from human participants, we applied dimensionality reduction and multivariate decoding techniques to test whether a sequence of items is maintained through the reactivation of individual items in sequence (neural sequences) or by organizing them into distinct low-dimensional subspaces (neural subspaces). Our results revealed behaviorally relevant, low-dimensional neural subspaces that organized memory representations according to ordinal position in the sequence, supporting the neural subspaces hypothesis. In contrast, we found no evidence of sequential neural replay during the retention period, contrary to the neural sequences hypothesis. In this work, we show that sequential working memory is maintained through structured geometric organization in low-dimensional representational space, rather than through the sequential reactivation of individual items, thus clarifying the mechanisms of multi-item working memory.

The neural mechanisms underlying working memory—the ability to hold and manipulate information over short periods—remain incompletely understood[1]. Classical theories based on single-neuron recordings propose that working memory relies on persistent firing[2,3], while more recent theories hypothesize that working memory information is represented via transient bursts of activity[4,5] or an interplay between activity-dependent and activity-silent mechanisms, such as short-term synaptic plasticity[6–9]. A critical challenge for neural theories of working memory storage is how they account for the ability to simultaneously represent multiple items without interference between the items. This challenge becomes particularly thorny when considering working memory for sequences—for example, when remembering a phone number—where not only the individual items, but their temporal order must be maintained with high fidelity. An emerging view has reframed working memory as an emergent property of dynamic activity patterns across neural ensembles, where heterogeneous neural dynamics at the single neuron level coexist with stable representations of memory contents in a low-dimensional space[10–12]. Recent work using population-level analyses of multiple neuron recordings has focused on characterizing the representational geometry of this low-dimensional subspace. These studies find that in multi-item working memory tasks, individual items are maintained during the retention period in distinct subspaces that help minimize interference and preserve behaviorally relevant information[13–15]. These low-dimensional subspaces not only capture stable memory representations but also reflect ongoing cognitive operations. When two items are held in memory simultaneously, their representations occupy quasi-orthogonal subspaces. Once one item is prioritized and

[1]Division of Science and Mathematics, New York University Abu Dhabi, Abu Dhabi, United Arab Emirates. [2]Neuroscience Center, Helsinki Institute of Life Science HiLIFE, University of Helsinki, Helsinki, Finland. [3]Department of Psychology and Center for Data Science, New York University, New York, NY, USA. [4]Center for Brain and Health, New York University Abu Dhabi, Abu Dhabi, United Arab Emirates. ✉e-mail: asantoangles@gmail.com

the other becomes irrelevant, the subspaces reorganize, becoming more aligned or parallel[13]. Similarly, in sequential working memory tasks, the identity of each item is encoded in a distinct subspace according to its ordinal position in the sequence[15]. If subjects are cued to recall the sequence in reverse order, memory representations swap between subspaces during the retention period to reflect the updated sequence[14]. Human neuroimaging research has similarly shown that large-scale brain dynamics evolve within a low-dimensional subspace[16,17]. Representational geometry has been extensively studied through representational similarity analysis[18,19], which demonstrates that the distances between high-dimensional multivariate neural activity patterns elicited by different sensory inputs reflect information processing[20]. Thus, the idea of stable subspaces represents an intriguing possibility for how sequences of information are maintained in working memory. However, studying neural geometry by leveraging low-dimensional subspaces remains a relatively underexplored direction in human neuroimaging, particularly in the domain of working memory[21]. In an fMRI study with a single-item working memory task[22], spatial stimuli were encoded within low-dimensional subspaces that showed a topological organization in visual and parietal cortices, but not in prefrontal areas. In another fMRI study[23], posterior cortical areas were found to orthogonalize memory representations of targets and distractors to reduce interference. These studies suggest that the human brain may leverage low-dimensional neural geometry to separate and maintain memory representations. Nevertheless, it remains unclear whether multiple working memory contents—whether presented simultaneously or sequentially—are also represented in orthogonal subspaces, whether this neural geometry is behaviorally relevant, and whether such neural coding can be detected using non-invasive, large-scale brain imaging techniques like MEG.

Another possibility is that multiple memory contents are embedded within the dynamic patterns of neural activity. This notion is particularly relevant in the context of sequential working memory, where the ordinal position of each item may be encoded through the temporal structure of transient neural reactivations. One compelling example of such temporally structured activity is hippocampal replay, an offline process occurring during sleep or restful wakefulness, in which previously experienced sequences are spontaneously reactivated in either forward or reverse order[24]. Hippocampal replay has also been implicated in sequence learning, where it is thought to support the extraction and consolidation of temporal structure across events[25]. Although neural replay has usually been studied with intracranial neuronal recordings, recent studies demonstrated that neural replay could also be studied in MEG data with the combination of decoding analysis and linear modeling[26]. These studies revealed human neural replay of task-relevant sequences deploying at compressed timescales during resting-state periods[27,28]. However, this approach has not yet been extended to examine online replay, defined as the rapid reactivation of memory sequences during working memory maintenance. In this context, the theta–gamma coding hypothesis of sequential working memory proposes that temporal organization relies on oscillatory phase coding mechanisms, particularly through interactions between theta and gamma rhythms[29]. According to this model, each item in a sequence is represented at a distinct phase of a theta cycle, with item-specific gamma bursts occurring at consistent theta phases, allowing the entire sequence to be maintained within a single theta cycle. Empirical support for this mechanism comes from M/EEG studies showing that gamma activity is systematically modulated by theta phase during the encoding and maintenance of sequential items in working memory, particularly in the hippocampus and prefrontal areas[30,31]. Moreover, the temporal phase patterns observed during memory retrieval and memory encoding of sequences are similar, suggesting that temporal information is preserved during the retention period[32]. However, the evidence for online neural replay during the maintenance period of sequential working memory is mixed. An EEG study using a temporal response function (TRF) approach reported rapid backward reactivation of memory items during the delay[33]. In contrast, an MEG decoding study found no evidence of sequential replay; instead, neural activity was dominated by a single prioritized item, typically the one that had been more weakly encoded[34]. These findings suggest that while replay may support memory maintenance, its occurrence during the delay period remains uncertain. Hence, evidence of online neural replay during the retention interval of sequential working memory tasks remains scarce.

In the current study, we tested these two complementary (but not mutually exclusive) hypotheses regarding the neural representation of sequential working memory (Fig. 1A). The neural subspaces hypothesis proposes that memory items are encoded and maintained in distinct, orthogonal neural subspaces, each representing a specific ordinal position. In contrast, the neural sequences hypothesis suggests that sequential information is encoded through the temporal dynamics of neural reactivations, reflecting online replay of memory items. Testing both hypotheses in the same dataset enables a direct, controlled comparison of competing neural coding models for sequential working memory, reducing confounds and revealing potential interactions. Here, we show that sequential working memory is maintained through low-dimensional neural subspaces that organize memory representations according to ordinal position in the sequence, rather than through the sequential reactivation of individual items, suggesting that the brain transforms temporal sequences into geometric patterns in the representational space to support efficient, stable, and behaviorally relevant maintenance of information in working memory.

## Results

### Task description and behavioral performance

Seventeen healthy participants performed a sequential working memory task (Fig. 1D) while their brain activity was recorded using magnetoencephalography (MEG). On each trial, participants observed a sequence of either 3 or 4 spatial locations and, following a blank memory retention interval, were required to reproduce the sequence accurately. Participants performed well, successfully reproducing the presented sequence on 83 ± 19% of trials across sessions, significantly above chance according to a binomial test ($p < 1e\text{-}15$ for all subjects). A linear regression analysis showed no significant effects of sequence length, session, or their interaction on task performance (all $p > 0.05$), suggesting consistent performance across different experimental conditions and sessions. This consistency justified pooling trials across sessions for subsequent analyses. Following MEG preprocessing (see Methods for details), an average of 267 ± 41 trials per subject were included in the analysis (221 ± 50 correct trials; 126 ± 15 correct trials for length 3 sequences; 96 ± 14 correct trials for length 4 sequences.

### Geometric analysis of neural subspaces

To test the hypothesis that sequential working memory contents were encoded in neural subspaces, we adapted the analysis pipeline described in previous work[13,15] for application to MEG data. Because signals from a common neural source can be detected by multiple distant sensors[35] potentially introducing spurious correlations that artifactually align neural subspaces in sensor space, we performed all analyses in source space. Specifically, we examined the geometry of neural subspaces using source-reconstructed data from a subset of 15 participants for whom individual MRI scans were available. The source data were parcellated into 200 cortical regions[36], and corrected for signal leakage using symmetric orthogonalization[37]. Neural subspaces for each ordinal position in the sequence were defined by projecting subject-specific condition-by-channel matrices of neural activity—where each condition represented a specific stimulus location at a given ordinal position in the sequence—onto their leading three principal components. Hence, each subspace was represented by the corresponding matrix of PC scores (Fig. S1 in Supplementary

https://doi.org/10.1038/s41467-026-78895-x

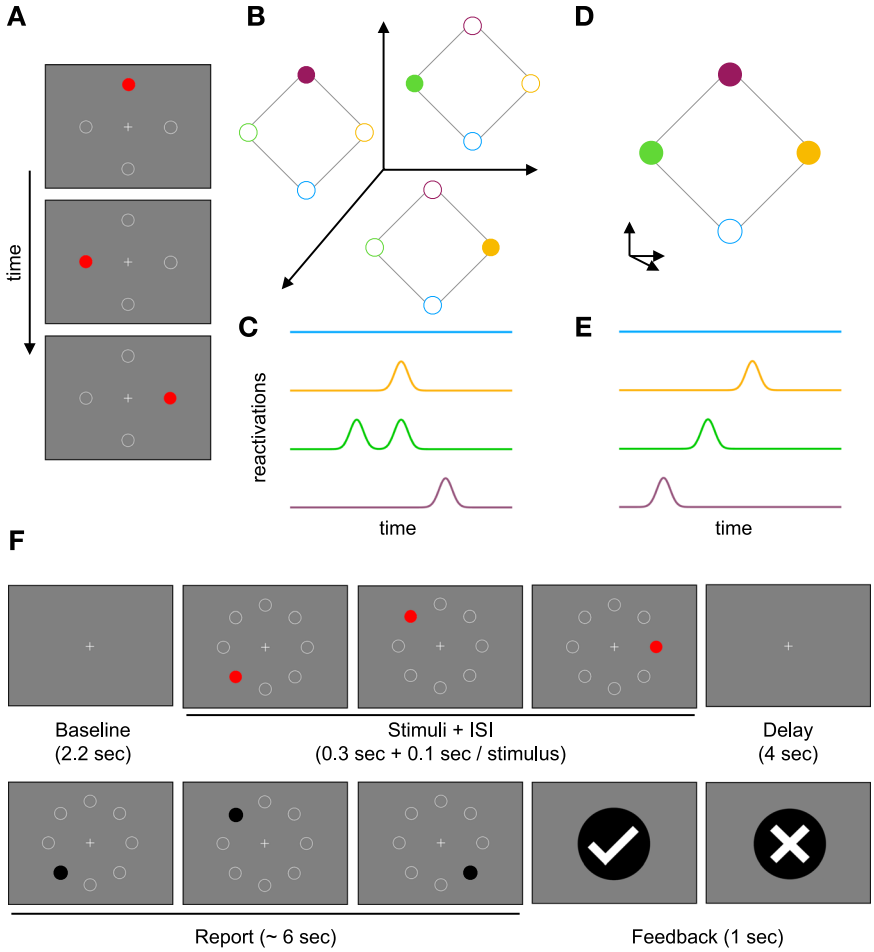



**Fig. 1 | Conceptualization of the study hypotheses regarding the neural code underlying sequential working memory. A** Example stimulus sequence in which three red dots are presented sequentially at one of four possible spatial locations, indicated by white placeholder circles. **B** Neural subspaces hypothesis: memory representations are encoded in rank-specific ring attractors, such that ordinal position is represented explicitly in distinct neural subspaces. Different stimuli can therefore be maintained simultaneously without interference because ring attractors corresponding to different ranks occupy orthogonal subspaces. **C** Under the Neural subspaces hypothesis, memory reactivations during the delay period—illustrated here as the time-resolved accuracy of linear classifiers—show no systematic temporal structure. Multiple representations can also be reactivated simultaneously (e.g., yellow and green stimuli) without interference because their corresponding ring attractors occupy orthogonal subspaces. **D** Neural sequence hypothesis: memory representations are encoded within a ring attractor embedded in an n-dimensional space. **E** Under the Neural sequences hypothesis, ordinal position is represented dynamically, such that reactivations during the delay period unfold sequentially according to the original presentation order, with a constant temporal lag between successive reactivations. Because the representations are embedded in a shared representational space, simultaneous reactivations would lead to interference, and therefore cannot occur without disrupting the sequential structure. **F** Design of the sequential working memory task.

Information). The degree of alignment between subspaces was quantified using two complementary metrics: the principal angle (PA) and the variance accounted for (VAF) between subspaces (see Methods for details). During both stimulus and delay periods, subspaces corresponding to different ordinal positions were significantly more orthogonal —that is, less overlapping or more discriminable— than expected by chance (Fig. 2A; summary statistics are provided in Supplementary Table S1). This effect was observed for both sequence lengths (3 and 4 items); the pattern was consistent across different alignment measures, and held relative to two baselines: a within-subspace comparison and a surrogate-based null distribution generated from geometric surrogates (see Methods for details). These findings suggest that the discriminability of memory representations was supported by a rank-specific subspace organization, in which representations were grouped according to their ordinal position. To assess the effect of memory load, we compared alignment metrics between trials of length 3 and length 4 (Fig. S2C and Table S1). Subspaces were generally more orthogonal (less aligned) for length-3 trials, though this difference was not consistent across time and only reached significance in isolated time windows. In order to assess whether these neural subspaces were behaviorally relevant, we compared trials with correct and incorrect responses after controlling for the imbalanced number of trials (see Methods for details). Incorrect trials showed significantly less orthogonal (i.e., more overlapping) neural subspaces (Fig. 2B and Table S1). For length 4 trials, these differences appeared during the stimulus period and persisted throughout the entire delay. In contrast, for length 3 trials, the differences were more transient, appearing briefly after stimulus offset and not maintained across the full delay. To further characterize the geometry of neural subspaces, we quantified the separability of memory representations within the subspace of each ordinal position using two metrics: the Euclidean distance between stimulus representations, and the volume spanned by those representations in the subspace. We observed no effect of memory load on separability metrics (Fig. S2D). However, the relationship between separability and performance depended on memory load: in length-4 trials, subspace volume was significantly greater for correct than for incorrect trials (Fig. S2A, B and Table S1). In summary, subspace geometry was behaviorally relevant,

https://doi.org/10.1038/s41467-026-78895-x

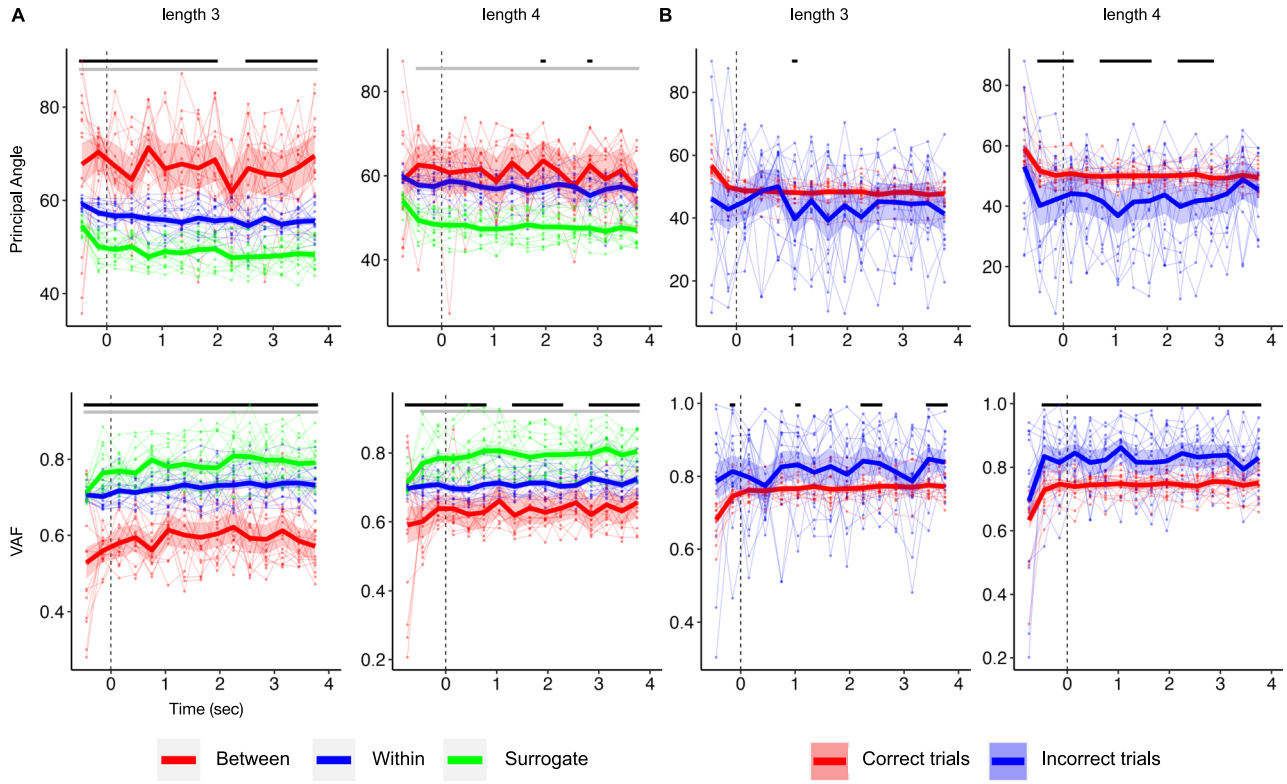



**Fig. 2 | Geometric analysis of neural subspaces. Alignment between rank-specific subspaces, quantified using the principal angle (PA) and variance accounted for (VAF), separately for trials of length 3 and length 4.** **A** Comparison of alignment measures for correct trials: empirical between-subspaces (red), geometric surrogate between-subspaces (green), and within-subspaces (blue). **B** Comparison of empirical between-subspace alignment in correct (red) versus incorrect trials (blue). X-axis shows time locked at delay onset (vertical dashed line). Black (gray) horizontal lines on top indicate time segments with FDR-corrected significant differences between the variables represented by red vs blue (red vs green) colors. Shaded areas indicate the standard error of the mean across subjects. Individual dots indicate subject-level observations. Summary statistics are provided in Supplementary Table S1. Source data are provided as a Source Data file.

with correct responses associated with more orthogonal subspaces and, under higher cognitive demand, with increased separability of memory representations within each subspace.

The observed orthogonality between subspaces suggests that memory representations occupy rank-specific subspaces. However, orthogonality alone does not establish whether subspace geometry is systematically organized according to ordinal position. To address this question, we examined whether subspace similarity varied as a function of rank separation and whether ordinal position could be decoded directly from subspace geometry. First, we quantified pairwise subspace similarity using two complementary metrics: the Frobenius norm, which captures element-wise differences between normalized matrices, and the Procrustes distance, a shape metric that quantifies differences in the global geometric configuration of the subspaces after optimal alignment[38]. We then tested whether subspace similarity varied as a function of rank separation and found a significant positive relationship between dissimilarity and rank separation (Fig. S3C, D). In contrast, alignment metrics (PA, VAF) showed no association with rank separation, indicating that subspaces remained similarly aligned regardless of ordinal distance and were approximately evenly spaced within the representational space (Fig. S3A, B). Second, we next asked whether ordinal position could be decoded directly from subspace geometry. To enable decoding despite having a single rank-specific subspace per subject and time window, we implemented a two-fold cross-validation procedure in which trials were split prior to subspace estimation, and bootstrap resampling was used to generate pseudo-independent subspace representations within each fold. A linear classifier trained on the training-fold subspaces successfully decoded ordinal position from the test-fold subspaces, revealing significantly above-chance decoding accuracy across time windows (Fig. S3E, F). To determine whether this effect reflected low-dimensional subspace geometry rather than information trivially present in the full neural activity matrices, we repeated the same analysis using the original full-data matrices instead of the low-dimensional PC score subspaces. Despite these matrices constituting the basis for subspace estimation and containing the underlying organization of neural responses across item identities, ordinal position could not be decoded from the full-data representations (Fig. S3E, F). These findings indicate that rank-related information was specifically captured by the low-dimensional geometry of the neural subspaces rather than by the high-dimensional activity patterns themselves.

Although neural representations occupied distinct rank-specific subspaces, it remained unclear whether the geometry of these subspaces was preserved over time or dynamically reconfigured throughout stimulus processing and memory maintenance. To test this, we recomputed subspaces (PC scores) by projecting time-resolved X matrices of neural activity onto common eigenvectors derived from specific reference windows. When eigenvectors were obtained from a single time window—either during the stimulus period (final stimulus presentation) or a restricted delay period window—subspaces were significantly more orthogonal than chance only within that same window, with no significant orthogonality at other time points (Fig. S4A, B and Table S2). This indicates that subspace geometry is highly dynamic across both stimulus and delay periods. In contrast, when common eigenvectors were derived from the entire delay period, subspaces remained

https://doi.org/10.1038/s41467-026-78895-x

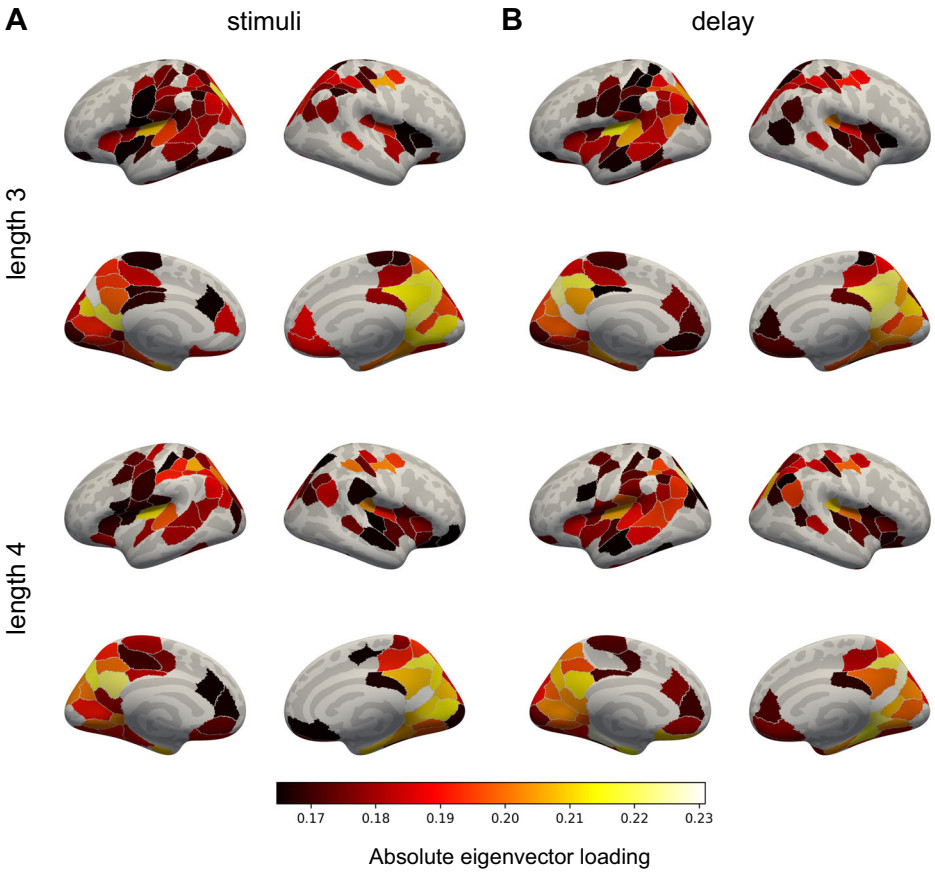



**Fig. 3** | Brain maps of the eigenvectors defining the neural subspaces, averaged across time windows and subjects, during **A** stimuli period and **B** delay period, separately for trials of length 3 and length 4. For visualization, maps are thresholded at the top 50% of values to highlight regions with the strongest contributions. Source data are provided as a Source Data file.

significantly more orthogonal than chance throughout the delay, but not during stimulus presentation, particularly when alignment was quantified using VAF (Figure S4C and Table S2). Similarly, when eigenvectors were computed from the second half of the delay period, subspaces remained significantly orthogonal during portions of the first half of the delay (Fig. S4D and Table S2). Together, these results suggest that although subspace geometry evolves over time, it exhibits greater temporal stability during the delay period than during stimulus presentation. To further assess the temporal stability of subspace geometry, we quantified subspace similarity over time using Frobenius norm and Procrustes distance. Subspaces corresponding to the same rank were consistently more similar to each other than to subspaces of different ranks, indicating stable rank-specific geometry over time. This effect was strongest during the delay period and was most robustly captured by the Procrustes distance. Importantly, subspace similarity did not decrease with increasing temporal separation within the delay period, providing no evidence for progressive drift. Together, these findings suggest that neural representations occupy stable, rank-specific subspaces during memory maintenance, whereas subspaces during stimulus processing are more dynamic (see Supplementary Results, Fig. S5, and Table S2).

Although rank specificity is expressed in low-dimensional neural subspaces, it remained unclear whether this low-dimensional structure accounts for a substantial portion of the neural variance and how it is distributed across brain networks. We therefore quantified subspace dimensionality using the eigenvalue spectrum and characterized the spatial distribution of subspace geometry across brain networks. The dimensionality of neural subspaces was evaluated using the eigenvalue spectrum of the data covariance (Fig. S7). In trials of length 3 (4), the leading three principal components explained 61 ± 5% (66 ± 5%) of the variance during the delay period, and 67 ± 8% (72 ± 8%) during the stimulus period. In contrast, when PCA was applied to eigenspectrum surrogates, constructed by independently shuffling rows within each channel, the variance explained by the leading three components dropped to 31 ± 0.6% for both periods, indicating that the observed low-dimensional structure reflects meaningful organization in the neural data rather than noise-driven covariance structure. The contributions of different brain networks to the neural subspaces were evaluated by projecting the leading three eigenvectors onto brain maps (Fig. 3). These maps were then averaged across time windows and subjects to identify consistent spatial patterns. To quantify network-specific contributions, eigenvector values were averaged across parcels within each network. Results showed that the stimulus and delay periods exhibited a similar distribution of parcel and network engagement. Among the networks, the visual network contributed the most, accounting for 31% of the total magnitude across all parcels, followed by the dorsal attentional network (20%), default mode network (16%), limbic network (13%), somatomotor network (10%), control network (7%), and salience/ventral attentional network (3%).

Supplementary analyses confirmed the robustness of the main findings across alternative subspace dimensionalities, control procedures for shared task-related activity, different definitions of neural activity, sensor- and source-space analyses, and approaches for handling trial-count imbalances (see Supplementary Results, Figs. S7–S13, and Tables S3–S8).

https://doi.org/10.1038/s41467-026-78895-x

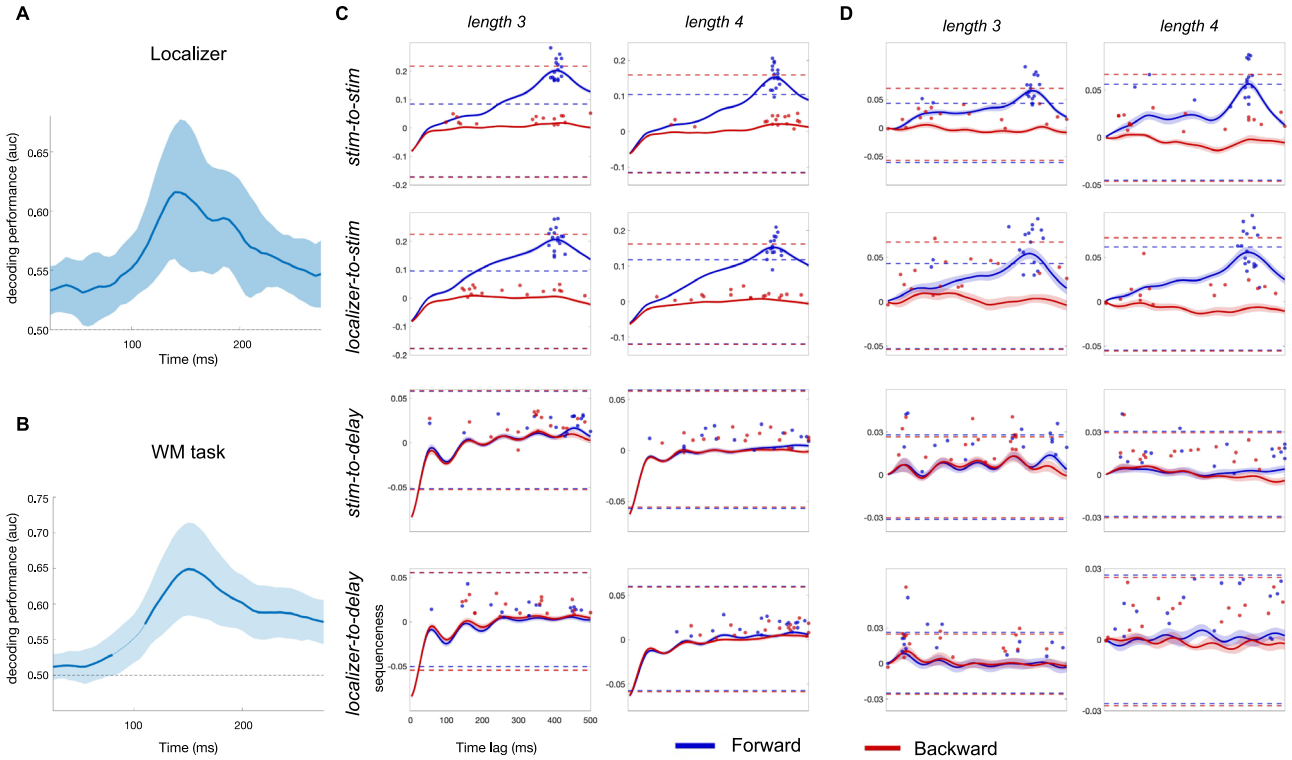



**Fig. 4 | Analysis of neural sequences. A,B** Decoding performance for stimulus location during stimulus presentation in the localizer task (**A**) and sequential WM task (**B**), time-locked to stimulus onset. The solid lines and shaded areas represent the mean and standard deviation of the area under the curve (AUC) across subjects. Thicker lines indicate decoding performance significantly above chance (dashed line at AUC = 0.5). **C,D** Sequenceness effect using **C** the trial-averaged TDLM approach, and D the across-trial concatenated TDLM approach. Each row corresponds to a different method for deriving the decoded state space used as input to TDLM: a) stim-to-stim: classifier trained on WM-stimuli data and applied on WM-stimuli data to compute decoded state space, b) localizer-to-stim: classifier trained on localizer data and applied on WM-stimuli data, c) stim-to-delay: classifier trained on WM-stimuli data and applied on WM-delay data, and d) localizer-to-delay: classifier trained on localizer data and applied on WM-delay data. Solid lines represent forward (blue) or backward (red) sequenceness effects averaged across subjects, shaded areas indicate the standard error of the mean, dots indicate the peak of sequenceness for each subject, and dashed lines indicate the upper and lower thresholds for statistical significance (two-sided non-parametric permutation test with maxT correction across time lags, $p < 0.05$). Source data are provided as a Source Data file.

## Sequence effects of memory representations

To test the hypothesis that sequential information was encoded in the dynamics of memory reactivations, we applied temporally delayed linear modeling (TDLM)[26]. We trained linear classifiers on sensor MEG data ($n = 17$) and source-reconstructed data ($n = 15$) to identify spatial locations during stimulus presentation in both the localizer and working memory tasks. Classifiers achieved significant cross-validated classification accuracy spanning through most of stimulus period, peaking around 150 ms after stimulus onset during localizer (sensor: peak = 160 ± 125, AUC = 0.56 ± 0.02, $p < 1e\text{-}5$; source: peak = 140 ± 50 ms, AUC = 0.63 ± 0.05, $p < 1e\text{-}5$) and WM task (sensor: peak=150 ± 90 ms, AUC = 0.66 ± 0.05, $p < 1e\text{-}5$; source: peak=155 ± 85, AUC = 0.59 ± 0.03, $p < 1e\text{-}5$) (Fig. 4 and S14A, B). Since sensor-level classifiers outperformed source-level classifiers, we performed the analysis of sequence effects on sensor data. The most accurate classifiers were then used to construct a decoded state space of memory representations by applying the classifier weights to delay period data (see Methods section for the group-level selection of the most accurate classifier). This process generated a matrix representing the decoded state space (time by spatial locations), encoding the probability of memory reactivations throughout the delay period. We also computed the decoded state space of sensory representations by repeating the same process on WM task stimuli period data. To quantify sequence effects at the subject level, we employed two complementary approaches. The first approach involved calculating sequence effects at the trial level and averaging these effects across all trials (trial-averaged TDLM approach) (Fig. 4B)[39]. The second approach concatenated the decoded state space across trials, allowing sequence effects to be computed collectively for all trials (across-trials concatenated TDLM approach) (Fig. 4C). This latter method aligns with the longer timescales for which TDLM was originally developed, specifically to investigate neural replay during resting state[27,28], and mitigates finite sample effects inherent to trial-level analyses. Both approaches yielded similar results. During the stimulus presentation, TDLM identified a significant forward sequence effect, peaking at a time lag of 400 ms. This timing matched the expected dynamics given the 300 ms stimulus presentation and 100 ms interstimulus interval (ISI) (Fig. 4B-C, first and second rows). These results were consistent across decoded state spaces computed using classifiers trained on either WM stimuli period data (stim-to-stim) or localizer data (localizer-to-stim). The successful recovery of forward sequence effects during stimulus presentation supports the sensitivity of both the decoding and TDLM analyses to detect sequence structure under the available trial numbers, signal-to-noise ratio, and temporal resolution. In contrast, during the delay period, we observed no evidence of forward or backward sequence effects, regardless of whether classifiers were trained on WM stimuli period data (stim-to-delay) or localizer data (localizer-to-delay) (Fig. 4B, C, third and fourth rows). Since classifier performance was reduced for discriminating between nearby locations (Fig. S14C, D), we repeated the analysis of sequence effects, excluding trials in which the sequence contained locations separated by only one step (e.g., the sequence [1, 2, 4] was excluded, but not [1, 3, 5]). This analysis

confirmed our previous findings (Fig. S15). In summary, we found no evidence of neural sequences of memory reactivations during the delay period, either at trial- or subject-level. Although linear classifiers did not consistently decode stimulus identity above chance during the delay period, analyses of classifier decision values indicated that they nevertheless carried stimulus-specific information during the delay period (see Supplementary Information, Fig. S14E–G).

To assess whether limited SNR could account for the absence of delay period sequenceness, we performed a sensitivity analysis in which simulated sequences were injected into the empirical delay period decision values. This approach preserved the empirical noise structure while allowing us to systematically vary the strength of the sequenceness signal (see Methods). We varied the injected sequence amplitude, temporal lag, and onset across trials, and quantified the TDLM sensitivity as the probability of recovering the injected sequence with TDLM. Sensitivity was high (>90%) at amplitudes comparable to the stronger empirical in-versus-out differences, while sensitivity was more moderate (50–90%) for more conservative estimates of the empirical amplitude (Fig. S17). Thus, although limited SNR may have constrained sensitivity to weaker sequential reactivation, the sensitivity analysis does not suggest that insufficient SNR alone can explain the absence of significant delay period sequenceness.

To further assess sequenceness effects and quantify evidence for the null hypothesis of no sequenceness, we conducted Bayesian analyses on TDLM results (see Supplementary Information, Fig. S16). During stimulus presentation, there was strong and consistent evidence for forward sequenceness across TDLM variants and sequence lengths (mean log Bayes factor across lags, logBF > 10), peaking in the long temporal lags (~370–420 ms), with no reliable evidence for backward sequences. In contrast, during the delay period there was no consistent evidence for either forward or backward sequenceness. Evidence for the null hypothesis of no delay period sequenceness depended on the analysis approach: it was strong and consistent for short time lags in trial-averaged TDLM (logBF = [−4.3 − 1.2]) but weak and inconsistent in across-trial concatenated TDLM analyses (logBF = [−1.5 1.8]). In summary, Bayesian analyses broadly supported the conclusions from the TDLM analyses, providing strong evidence for forward sequenceness during stimulus presentation and no evidence for reliable sequenceness during the delay period.

## Discussion

In this study, we tested two competing—but not mutually exclusive—hypotheses about the neural mechanisms underlying sequential working memory: neural subspaces versus neural sequences. Our results support the neural subspace hypothesis, showing that memory items are maintained in low-dimensional, behaviorally relevant neural subspaces. In contrast, we found no evidence supporting the sequential reactivation of memory contents.

Our findings on neural subspaces replicate previous results from neuronal recordings in non-human primates, demonstrating quasi-orthogonal subspaces during the retention period. We observed principal angle values of approximately 70 degrees for trials of length 3 and 60 degrees for trials of length 4—comparable to the 70–80 degree range reported in non-human primates[13–15]. Notably, subspace geometry differed between the stimulus and delay periods. Using subspace alignment measures (principal angle and VAF), we found that rank-specific subspaces were significantly more orthogonal than expected by chance, but this structure was strongly dependent on the temporal reference used to define the eigenvectors. When subspaces were defined using eigenvectors derived from a single time window (e.g., the stimulus or a restricted delay period), significant orthogonality was largely confined to that same window, with little generalization across time, indicating that subspace structure is highly dynamic[22]. In contrast, eigenvectors derived from the entire delay period yielded sustained orthogonality throughout the delay, whereas no comparable stability was observed during stimulus presentation. These findings were further supported by shape-based analyses of global subspace geometry (Procrustes distance), which showed that the overall geometric configuration of rank-specific subspaces remained stable across the delay period but not during stimulus processing. Together, these results suggest a transition from a highly dynamic representational geometry during stimulus encoding to a more stable, rank-specific subspace organization during the delay period, consistent with a stable maintenance format for sequential working memory. Interestingly, the brain maps reflecting cortical parcel contributions to the neural subspaces were remarkably similar across stimuli and retention periods (Fig. 3). These maps revealed a consistent spatial pattern involving occipital and parietal cortices—recruiting visual and dorsal attention networks—as well as left-lateralized prefrontal and temporal regions, brain areas widely associated with working memory[40]. Taken together, these findings suggest that encoding and maintenance engage overlapping brain networks, in line with the sensory recruitment hypothesis[41–43]. However, the computations supported by these regions differ across phases: neural subspaces are present in both stimulus and retention periods, but become stable and temporally consistent only after stimulus offset, whereas they remain more dynamic during stimulus encoding. Consistent with this idea, previous studies report that sensory representations are transformed into memory representations via orthogonalization through rotational dynamics, allowing perceptual and mnemonic codes to be segregated within overlapping neural populations[44].

The geometry of subspaces in our data was behaviorally relevant, consistent with previous studies[13,15]. Specifically, subspace geometry has been shown to predict systematic memory errors: greater overlap between adjacent rank subspaces leads to transposition errors, while ring-like structures within subspaces account for confusions between nearby spatial locations[15]. In addition, error trials tend to exhibit more overlapping subspaces[13]. We replicated this pattern, observing increased overlap between subspaces in incorrect trials. This suggests that orthogonal—or at least sufficiently distinct—neural subspaces are necessary to keep memory representations discriminable. In contrast, excessive alignment between subspaces reduces this discriminability, making representations more confusable and thus more prone to behavioral errors. These results support the idea that distortions in neural subspace geometry may underlie failures in working memory performance. Regarding the separability of stimulus representations, when subspaces were computed using 3 PCs, we initially found no differences between correct and incorrect trials, except for the volume metric in trials of length 4, where correct trials exhibited greater volume than incorrect trials. The selective effect on volume may suggest that subspaces under higher memory load exhibit greater nonlinearity in their geometry. Unlike pairwise distances, volume reflects the overall shape and spread of the representation, capturing more complex spatial configurations. For example, a flat and a curved (e.g., bent) ring structure with equally spaced item representations yields the same Euclidean distances between points, but the curved structure would enclose a larger volume when measured using the convex hull. The fact that this effect emerged only in trials of length 4 may reflect a memory load–dependent distortion of the representational geometry. Consistent with this interpretation, principal angles were smaller in length-4 than length-3 correct trials. Extending the analysis to subspaces computed with 8 PCs revealed additional effects: volume differences also appeared in length-3 trials, and pairwise distance differences emerged in length-4 trials. These findings indicate that higher-dimensional subspaces reveal aspects of the representational geometry that are less apparent in lower-dimensional projections. This could reflect nonlinear distortions, but it could also indicate that memory representations are fundamentally linear yet occupy a higher-

dimensional space that is missed when projecting onto only 3 PCs. Together, these findings suggest that memory load may influence both the dimensionality of subspaces and the degree of nonlinearity in their geometry, consistent with prior work linking greater task complexity to more nonlinear representational structures[45], although the proposed link between memory load and nonlinear distortions in subspace geometry remains speculative and should be tested experimentally.

Contrary to predictions from the neural sequences hypothesis, we did not observe evidence of sequential reactivation during the retention period. We applied a validated method, Temporally Delayed Linear Modeling (TDLM)[26], to estimate neural sequences in broadband MEG data. This approach has reliably detected neural sequence replay during resting-state periods[27,28] as well as task-related sequential reactivation[39], demonstrating its sensitivity and robustness in identifying temporally structured neural activity. The absence of evidence was consistent across multiple analysis strategies and was supported by Bayesian analyses. We trained classifiers both on a localizer task, which contained no inherent sequence structure, and on stimulus presentations during the working memory task, achieving significant decoding performance in both cases. The decoded state space successfully identified expected sequences of reactivation during stimulus presentation, confirming the validity of our classifiers—even when trained on the localizer data and tested on WM task data. To further validate our results, we adapted the TDLM pipeline to control for finite sample effects and concatenated across trials to increase statistical power. Despite these efforts, no evidence of sequential reactivation emerged during the retention period. Our results align with previous findings[34], which also reported no evidence of neural replay during the retention period in an MEG study of sequential working memory. Instead, their decoding analysis revealed that neural activity during the delay was dominated by a single item, regardless of its original position in the sequence or category. This prioritized item tended to be one that had been more weakly encoded during the stimulus period, suggesting that attention may selectively enhance the maintenance of less robust memory traces, rather than supporting uniform sequential replay. In contrast, our findings diverge from another study that reported evidence of neural replay during the retention period in a sequential working memory task[33]. Using a temporal response function (TRF) approach, they tracked item-specific reactivations by presenting task-irrelevant visual probes—each uniquely colored and previously associated with a memory item—and modeling alpha-band responses to these flickering stimuli. Their results revealed fast, reverse-order replay of items within brief (200–400 ms) time windows. While this approach offers valuable insights, its reliance on rhythmic luminance introduces interpretive challenges, as the observed reactivations may reflect stimulus-evoked responses rather than spontaneous memory processes. Indeed, prior work has shown that flicker-based "pinging" can trigger retrieval of latent memory representations[46,47], reorganizing working memory contents by decreasing neural dynamics[48], raising the possibility that the reported sequential reactivations were externally driven. In this context, decoding-based approaches, which track spontaneous reactivations of memory content during the delay period, may offer a more direct test of whether sequential replay underlies working memory maintenance[26,49]. Our findings regarding neural sequences also resonate with behavioral and neuroimaging studies suggesting that working memory compresses sequential information using abstract geometric and numerical primitives[50,51]. According to this view, rather than storing and replaying individual items in sequence, the brain encodes higher-order structure—such as symmetry, rotation, or numerical rules—that generates the entire sequence. This form of mental compression could reduce the need for item-by-item reactivation during the retention period, potentially explaining the absence of detectable neural replay in our data. However, if rule-based compression were the only mechanism supporting sequential working memory, the neural subspace patterns we observed would likely not emerge.

The absence of evidence for neural sequences during the delay period is somewhat unexpected in light of the theta–gamma coding model of sequential working memory, which proposes that individual memory items are represented by gamma-band activity nested within distinct phases of the theta cycle[29]. This model is supported by evidence of cross-frequency coupling modulation during both the encoding and maintenance periods[30,31], showing that the temporal structure of theta–gamma interactions plays a critical role in organizing sequential information in working memory. However, a recent study challenged the main prediction of sequential reactivation of memory representations during the maintenance period[52]. Spikes and LFPs recorded in the human medial temporal lobe showed increased theta power and spike-phase locking for remembered items. While firing phase reflected item position in correct trials, the sequence of firing phases did not match item order, consistent with our finding on neural sequences. Taken together with previous work, an alternative interpretation is that theta–gamma cross-frequency coupling observed during sequential working memory tasks, traditionally interpreted as evidence of sequential reactivation of memory representations phase-locked to theta oscillations, may instead reflect that ordinal position information is encoded in subspaces via the phase of low-frequency activity, without necessitating sequential reactivations. Previous work, including our own, has shown that multi-item working memory contents are maintained in quasi-orthogonal subspaces to minimize interference. However, it remains unclear how ordinal position is geometrically encoded within this framework. The position of each subspace in the ambient neural space might carry this information[14]; alternatively, subspaces may be dynamically indexed by low-frequency oscillations, consistent with a recent study in macaque PFC in which theta power transiently encoded each stimulus in the sequence irrespective of ordinal position, while theta–spike interactions selectively tracked memory subspaces to bind items to their ranks[53]. Future studies combining geometric analysis of neural subspaces with oscillatory brain activity could more directly test how interactions between neural rhythms shape subspace geometry.

Several factors in our study may also account for the absence of evidence for neural sequences. The use of broadband time series to train classifiers within the TDLM pipeline may have limited our ability to detect neural sequences that are expressed in specific frequency bands, such as theta or gamma. Frequency-specific dynamics are thought to play a critical role in organizing sequential information, and collapsing across frequencies might obscure these patterns. However, previous MEG studies have successfully identified neural replay during resting-state periods using similar decoding approaches[27,28], despite the replay having well-defined spectral characteristics, suggesting that frequency specificity alone may not fully account for the absence of detectable replay in our data. Another limitation for testing the neural sequences hypothesis is the potential mismatch between the classifiers' training data and the neural format of memory representations during the retention period. In our study, classifiers were trained on neural activity recorded during stimulus presentation, which may not capture the distinct dynamics or coding schemes that characterize mnemonic representations during delay. This misalignment could have reduced the sensitivity of the decoded state space to detect memory-related neural sequences. Future studies could address this limitation by using a memory-related functional localizer—specifically, by training classifiers on neural activity recorded during the delay period of single-item delayed match-to-sample tasks involving the same stimuli used in the main sequential working memory task. This approach may produce classifiers that are better tuned to the neural code used during memory maintenance. Another limitation of our study, concerning the behavioral relevance of neural subspaces, is the imbalance in trial numbers between correct and incorrect trials, which

can bias estimates of subspace geometry. To mitigate this, we implemented a two-step strategy combining projection onto common eigenvectors and controlled resampling of correct trials to match the number of incorrect trials. Our control analyses show that subspace metrics are strongly influenced by trial count and by the choice of representational space, and that this combined approach reduces these confounds and yields more interpretable comparisons. However, this strategy is conservative: by projecting both trial types into a shared space, it may reduce sensitivity to genuine differences, potentially increasing false negatives. In addition, this approach may obscure differences in representational dimensionality between correct and incorrect trials. For instance, incorrect trials may arise from representations that are either too high- or too low-dimensional for effective readout. Given the limited number of incorrect trials in our dataset, we are not well positioned to directly test this possibility. Future studies with more balanced trial counts will be necessary to assess differences in representational dimensionality.

In conclusion, we tested two hypotheses about the neural mechanisms underlying sequential working memory: the neural subspaces hypothesis and the neural sequences hypothesis. Our findings revealed behaviorally relevant, low-dimensional neural subspaces that organized memory representations during the retention period. In contrast, we found no evidence of sequential neural replay, arguing against the idea that maintenance relies on item-by-item reactivation. Together, these results suggest that the brain transforms temporal sequences into geometric patterns in the representational space to support efficient, stable, and behaviorally relevant maintenance of information in working memory.

## Methods

### Sample

We recruited 29 healthy adults from the NYUAD community to participate in this study, which included two MEG sessions and one MRI session. Five participants were excluded for not completing all MEG sessions, four were excluded due to excessive motion, and three were excluded for not adhering to the fixation task (details provided below). Ultimately, 17 participants successfully completed both MEG sessions (age: 23 ± 5 [19 40] years; 7 females). Sex was self-reported by participants at enrollment. Time between MEG sessions: 8 ± 5 [1 18] days. Additionally, MRI data for anatomical structure were collected from 15 (6 females) of these 17 participants (two participants could not complete the MRI session due to claustrophobia). The final sample of 17 (15) participants was used for MEG analysis at the sensor (source) level. All subjects provided informed written consent in accordance with procedures approved by NYUAD's IRB.

### Task design

For each session, subjects performed a functional localizer and a sequential working memory task (Fig. 1D). Both tasks used the same set of stimuli: eight dots with a diameter of 60 pixels (1.88°), positioned at an eccentricity of 350 pixels (11°) and evenly spaced in 45° intervals around a 360° circle. During the presentation of a location, its corresponding dot turned red, while the placeholders for the other locations remained visible throughout both the stimulus presentation and the inter-stimulus interval.

In the functional localizer, each location was presented for 0.4 seconds, with an interstimulus interval of 0.2 seconds. A total of 560 stimuli were displayed in pseudorandom order, with 70 presentations per location, ensuring that no location was repeated consecutively. Participants were instructed to maintain their gaze on a fixation point. To enforce this, subjects pressed a button whenever the fixation point flickered (a 10% change in luminance), occurring randomly twelve times (fixation task). Additionally, to maintain engagement, participants pressed a button whenever the color of a dot was orange instead of red, which also occurred twelve times (color task). All participants completed both tasks successfully (fixation task: 11.2 ± 1.5 [7 12] responses; color task: 11.8 ± 0.4 [11 12]). The primary goal of the functional localizer was to collect sensory representation data for training classifiers used in the decoding analysis, independently of the working memory task.

The sequential working memory task consisted of 210 trials, divided into three blocks of 70 trials each. Unique sequences were created at random, with no consecutive repeated locations. Trials of length-3 and 4 were intertwined. The first (second) session had 115 (110) trials of length-3, and 95 (100) trials of length-4. Every trial began with a 2.2-second baseline, followed by the presentation of a sequence of spatial locations (0.3 s per item, with an interstimulus interval of 0.1 s). The sequence length varied between 3 and 4 items. After a 4-second delay, participants were prompted to reproduce the sequence using a discrete circular response controller (Current Designs, Philadelphia, PA, USA). At the end of their response, participants received dichotomous qualitative feedback. After the feedback, the next trial started only when the participant indicated they were ready by pressing a button, allowing them time to blink. Participants were instructed to keep their gaze fixed on a central point throughout the entire trial. During the delay period, they performed the same fixation task described previously, with the fixation point flickering in 10 randomly selected trials across the session, and these trials were excluded for subsequent analyses. Participants who did not successfully complete the fixation task in at least 60% of the trials in both sessions were excluded from the study (fixation task: 8.6 ± 2 [6 10] correct responses in the final sample).

### Behavioral analysis

To determine whether participants' performance on the working memory task was significantly above chance, we conducted a binomial test for each subject and sequence length, comparing the observed number of correct responses to the expected number based on the chance levels. The probability of achieving correct responses by chance was calculated separately for sequences of length-3 ($1/n^3$) and length-4 ($1/n^4$), where $n = 8$ possible locations. Additionally, we performed a linear regression analysis to assess behavioral performance. In this model, the percentage of correct responses served as the response variable, while sequence length, session, and their interaction were included as predictor variables.

### MEG and MRI acquisition

MEG data were recorded continuously using a 208-channel axial gradiometer Yokogawa system (Kanazawa Institute of Technology, Kanazawa, Japan) with a sampling rate of 1000 Hz and an online low-pass filter of 200 Hz. During the MEG session, visual stimulation was presented in a Vpixx system (PROPixx 120 Hz). An image of 1920×1080 pixels was projected on a screen of 65 ×36.5 cm, and the viewing distance was 56 cm, resulting in a viewable image of 60.26° x 36.1° of visual angle in width and height. Before MEG acquisition, each subject's head shape was digitized using a Polhemus dual source handheld FastSCAN-II. High-resolution anatomical MRI volume was collected during a separate session to constrain the transformation from sensor to source space. The anatomical scan was acquired on a Siemens Magnetom Prisma 3 T MRI scanner using a 6.5 min MPRAGE-3D T1-weighted, gradient-echo acquisition sequence (TR: 2,400 ms, TE = 2.22 ms, flip angle: 8 degrees, voxel size: 0.8 mm3, 208 slices, FOV: 256 mm).

### Preprocessing

The continuous MEG data was first noise-reduced using eight magnetometer reference channels located away from the participant's head and using the Time Shifted Principle Component Analysis (TSPCA, block width of 5,000 and 30 shifts) as implemented in MEG160 software (Yokogawa Electric Corporation and Eagle

Technology Corporation, Tokyo, Japan). All subsequent preprocessing steps were performed using FieldTrip[54], according to the FLUX pipeline[55]. Localizer data were segmented into trials based on stimulus presentation. For the working memory task, data were epoched from 1 second before stimulus onset to 4 seconds after stimulus offset, comprising baseline, stimulus presentation, and the memory delay period. Automatic artifact detection was used to exclude trials with jump, muscle, and blink artifacts (z-value = 25). We kept 91 ± 3 % of trials in the localizer (valid trials: 1023 ± 30 [943 1053] out of 1120 trials across sessions), and 80 ± 10% of trials in the working memory task (valid trials: 336 ± 40 [233 380] out of 420 trials across sessions). To further clean up the data, we used Independent Component Analysis (ICA) to attenuate blink, muscle, and cardiac artifacts. The noise ICs, identified by visual inspection based on their topography and time courses, were regressed out from the data (localizer: 18 ± 4 [10 25] noise ICs; WM task: 23 ± 5 [15 30]).

Structural images were skull-stripped (FSL-bet) and registered to the MNI space with linear (FSL-flirt) and non-linear (FSL-fnirt) transformations with FMRIB Software Library[56]. Then, these transformation matrices were used to compute the subject-specific parcellations of the Schaefer atlas with 200 parcellations and 7 networks[36]. We translated sensor-level data to source space using a beamforming technique[57]. The head shape and structural MRI images were aligned to the Neuromag coordinate system, facilitating accurate coregistration between anatomical and functional data. A realistic single-shell volume conduction model was computed based on the subject's anatomy[58], and the source model was generated using a 1.25 cm isotropic 3D grid inside the head model. These source and head models were subsequently used to construct a forward (leadfield) model in the native anatomical space. The resulting spatial filters were used to localize sources in the time domain with the Linearly Constrained Minimum Variance method, projecting the dipole time series in the direction of maximum variance, and computing the absolute value of the reconstructed signal to deal with sign ambiguity. Finally, we pooled sources within each atlas parcel using an eigenvariate method, and removed zero-lag correlations between parcels' time series in order to address spatial leakage with symmetric orthogonalization using the MEG-ROI-nets toolbox[37].

### Geometric analysis

The geometric analysis of neural subspaces was adapted from previous work[13,15], but applied here to time series of broadband neural activity from 200 cortical parcels defined by Schaefer atlas[36]. We selected 200 parcels to approximately match the dimensionality of the denoised sensor data (207 sensors minus the number of components rejected during ICA-based artifact attenuation), as increasing the spatial resolution with finer-grained cortical parcellation would lead to greater linear mixing of source time series. An overview of the geometry analysis pipeline is provided in Supplementary Fig. S1. First, we constructed neural activity matrices X of size M×N, where M is the number of conditions—defined by the stimulus locations at each ordinal position in the sequence (rank)—and N is the number of channels (brain parcels). Each column of X captures the responsiveness of a given channel to the sequences presented across trials, estimated using ridge-regularized linear regression ($\lambda = 1$) applied separately to each channel. The response variable was the channel activity averaged over a predefined time window (see below), baseline-corrected by subtracting the mean activity during a 1-second pre-stimulus baseline period. The design matrix included n×r binary predictor variables, where $n = 8$ is the number of stimulus locations and r is the rank (sequence length), indicating which stimuli were presented on each trial. The resulting beta weights for the M conditions served as the entries in that channel's column of X. Finally, each column of the X matrix was demeaned prior to geometric analysis. We computed subject-specific X matrices separately for correct and incorrect trials. For each response type, we constructed one X matrix per sequence length (three or four items) and for each time window of interest. These time windows included the stimulus presentation period (0.3 seconds) as well as the delay period, which was divided into 13 non-overlapping windows of 0.3 seconds each. Then, we reduced the dimensionality of the X matrix along the channel dimension (columns of X) using Principal Component Analysis (using the svd.m MATLAB function applied to the covariance matrix of X). We then projected X onto the subspace defined by the top three principal components, resulting in a matrix Z of size M×3, where each row corresponds to a condition and each column to a principal component (i.e., PC scores). Then, we subset the Z matrix by sequence rank, creating $Z_r$ matrices (8 locations by r components), with r = [1 2 3] for length-3 and r = [1 2 3 4] for length-4, representing the rank-specific neural subspaces. Finally, we computed the subspace plane for each rank-specific subspace by taking the two leading components of a PCA on $Z_r$. To visualize the spatial distribution of neural subspaces, we plotted brain maps of the eigenvectors corresponding to the top three principal components. For each component, we computed the absolute value of its eigenvector and averaged these across the three leading components, across time windows and across subjects. Taking the absolute value addresses the inherent sign ambiguity of PCA, preventing cancellation due to arbitrary polarity. This approach highlights the relative contribution of each brain parcel to the dominant dimensions of the neural subspaces.

To quantify the alignment between rank-specific subspaces, we computed the principal angle (PA) and variance accounted for (VAF) between subspaces[13,15]. PA was defined as the angle between the normal vectors of the planes corresponding to each subspace, following the normal vector approach outlined in ref. 13. VAF was calculated as the proportion of variance explained by the principal components of one rank-subspace when neural data from another rank-subspace is projected onto it, providing a measure of the similarity in neural geometry across subspaces, as described in ref. 15. We calculated the lower (upper) bound of PA (VAF), referred to as within-subspace PA/VAF (blue curves in Fig. 2), with a bootstrapping approach. We randomly half-split trials to create two X matrices, and computed the PA and VAF between the same ranks across the two splits. We repeated the process 1,000 times and averaged across iterations. To quantify the discriminability of locations for each rank, we computed two separability metrics in the neural subspaces $Z_r$. First, the Euclidean distance between locations nearby in the physical space as a measure of how well the subspace captures fine-grained differences between adjacent or closely located points. Second, the volume of the set of points defining all rank-specific locations (MATLAB's convhulln function).

We used linear regression models to perform group-level statistical analyses. Separate models were fitted for each sequence length (three or four items) and across different time windows. To test whether the alignment between rank subspaces exceeded chance levels, we fit linear models with PA or VAF as the response variable and a predictor indicating whether the comparison was between different subspaces (between-subspace PA/VAF, red curves in Fig. 2) or within the same subspace (within-subspace PA/VAF, blue curves in Fig. 2). To further assess whether observed alignment exceeded what would be expected from structureless data, we fit a second model with the same response variable (PA or VAF) but with a predictor distinguishing empirical data from geometric surrogate data (surrogate PA/VAF, green curves in Fig. 2; see below for a detailed description). To assess the behavioral relevance of neural geometry, we fitted several distinct models. Each model included trial accuracy (correct vs. incorrect) as the predictor variable and one of the following as the response variable: alignment measures (PA or VAF) or separability measures (distance or volume). To examine the modulation of neural geometry by cognitive load, we applied the same set of models, but compared geometric variables between trials of length 3 vs. length 4. All *p*-values

were corrected for multiple comparisons using False Discovery Rate (FDR) correction. Effects were considered significant at $p < 0.05$ (two-tailed).

To address potential bias arising from the reduced number of incorrect trials—typically around 20% of the number of correct trials—we adopted a two-step strategy. First, we computed the neural subspace for incorrect trials by projecting their X matrices onto the leading principal components (eigenvectors) derived from correct trials ('common eigenvectors' step). This yielded a Z matrix (PC scores) for incorrect trials within the same low-dimensional space as correct trials. This approach ensures that comparisons between correct and incorrect trials are made within a shared representational subspace, enabling direct assessments of geometric properties such as distances and structure. It also assumes that the neural geometry of correct trials reflects a more stable or canonical representation of task-relevant information, allowing us to quantify how much the representations during incorrect trials deviate from this normative structure. Importantly, this avoids estimating an unstable or noisy subspace from the smaller and potentially more variable set of incorrect trials, which could confound interpretation of geometric differences. Second, we implemented a controlled resampling approach to address the imbalance in trial counts ('controlled resampling' step). For each comparison of geometric variables, we randomly subsampled correct trials to match the number of incorrect trials, then recomputed the X matrix using only those correct trials. We projected both the subsampled correct-trial and the full incorrect-trial X matrices onto the leading principal components derived from the original correct-trial data. Geometric variables were then calculated for each iteration, and results were averaged across 1000 resampling iterations to ensure robustness. As a supplementary analysis to evaluate the impact of these steps, we conducted additional analyses in which each step of the procedure was selectively omitted (Figure S12; Table S8), and we assessed the effect of trial number on subspace geometry by varying the proportion of trials used to estimate $X$ in correct trials (Figure S13).

We generated two distinct null distributions (1,000 iterations), each targeting a different aspect of the data's underlying structure: geometric surrogates, which disrupt stimulus arrangement in the subspaces, and eigenspectrum surrogates, which disrupt the global covariance structure of the data. In the geometric surrogates, we constructed a null distribution of subspace alignment metrics (PA and VAF) by independently shuffling the rows and columns of the X matrices (green curves in Fig. 2). This procedure preserves the overall covariance structure of the data, reflected in an unchanged eigenvalue spectrum, while disrupting the specific stimulus arrangement within the low-dimensional manifold. In the eigenspectrum surrogates, designed to assess the dimensionality of the neural manifold, we generated a null distribution of eigenvalues by independently shuffling the rows (conditions) within each column (channel) of the X matrix (see Fig. S6). This permutation destroys covariance structure between conditions and channels, yielding covariance matrices whose spectra are dominated by noise.

The stability of neural subspaces was evaluated as follows. First, we recomputed subspaces (PC scores) by projecting time-resolved neural activity matrices (X) onto a common set of eigenvectors derived from different reference time windows: (A) presentation of the final stimulus in the sequence; (B) 1.8–2.1 s after delay onset; (C) the full delay period; and (D) the second half of the delay period (2–4 s after delay onset). Then, we evaluated whether those subspaces were significantly more orthogonal than chance level, comparing the alignment metrics (PA and VAF) with both within-subspace and geometric surrogate-based estimates (see Fig. S4; Table S2). Persistence of significant orthogonality when using common eigenvectors was taken as evidence for stable subspace geometry (i.e., consistent alignment between subspaces), whereas its absence indicated a more dynamic representational code. Second, we assessed the temporal consistency of subspace geometry itself (independent of alignment between subspaces). To do this, we quantified the similarity between subspaces (PC-score matrices) obtained from different time windows with two distance metrics: Frobenius norm and procrustes distance. (A) The Frobenius norm distance quantifies the overall element-wise difference between subspaces. Each PC-score matrix was first normalized to unit Frobenius norm, allowing comparisons that are independent of absolute scale. (B) Procrustes distance captures differences in the overall geometric shape of the subspaces. Each PC-score matrix is optimally aligned to the other through translation, rotation, and uniform scaling, and the residual mismatch after this alignment is taken as the Procrustes distance. For each distance metric, we computed two time-generalization distance matrices at subject-level. For every pair of time windows, we calculated the distance between the corresponding PC-score matrices $Z_i(t1)$ and $Z_j(t2)$, where i and j denote the rank of the subspace (i.e., ordinal position) and t denotes time (see Fig. S1 in the manuscript). When $i = j$, the resulting matrix reflects the temporal distance of the same subspace across time ("within-rank" distances). When $i \neq j$, the matrix reflects the distance between different subspaces ("between-rank" distances), which quantifies how similar or distinct the representational geometries of different ranks are over time. We then averaged across the appropriate rank combinations to obtain a single within-rank and a single between-rank time-generalization matrix for each subject. For example, for length 3 trials, the within-rank matrix is the average of the three diagonal rank pairs (1,1), (2,2), and (3,3), while the between-rank matrix is the average of all six off-diagonal rank pairs (1,2), (1,3), (2,1), (2,3), (3,1), (3,2). To quantify whether subspace geometry is more stable within-rank than between-ranks, and whether this stability differs between stimulus and delay periods, we applied a two-level hierarchical statistical model. In the first level, we tested for each time-window pair whether within-rank distances were significantly smaller (i.e., more stable) than between-rank distances. For each matrix element (i,j), we computed a paired t-test comparing distances across subjects (within vs. between), yielding a time-by-time map of t-values. In the second level, we tested (a) whether within-rank distances were reliably lower than between-rank distances across the entire temporal generalization space (one-sample t-test), (b) whether subspaces were more stable during the delay than during the stimulus period (dependent samples t-test between the stimulus-to-stimulus pairs versus delay-to-delay pairs), and (c) whether representational dissimilarity increased with temporal separation during the delay period by means of a regression analysis on the delay–delay section of the time-generalization distance matrix. For each pair of delay period time points t1 and t2, we extracted the corresponding distance value and computed their temporal separation $\Delta t = |t1 - t2|$. We then fit a linear model of the form $D = \beta 0 + \beta 1\, \Delta t + \varepsilon$, where D denotes the distance between the two time points. A positive slope ($\beta 1 > 0$) indicates that subspaces become more dissimilar as the temporal gap between them increases. These analyses were performed separately for Procrustes and Frobenius distances.

Rank-specificity of neural subspaces was evaluated as follows. First, we tested whether relationships between subspaces preserve ordinal structure (e.g., whether rank 1 is more similar to rank 2 than to rank 3). To do so, we quantified pairwise subspace relationships using measures of subspace alignment (PA/VAF) and global geometry (Procrustes distance and Frobenius norm). For each geometric measure, we fitted a linear model separately across all pairs of subspaces (ranks), subjects, and time windows, with the geometric variable as the response and rank separation as the predictor. Subject identity was included as a covariate to account for repeated measurements, and observations from all subjects and time windows were concatenated within the same model. This analysis allowed us to test whether

geometric variables between subspaces increase systematically with ordinal rank separation, consistent with preservation of rank structure.

Second, we tested whether ordinal position (rank) could be decoded from subspace geometry. Subspaces were defined at the subject level using all available trials, which precludes standard decoding approaches that rely on independent samples, as this yields only one rank-specific subspace (i.e., PC scores) per subject and time window. To overcome this limitation, we implemented a 2-fold cross-validation procedure at the level of subspace estimation. Specifically, instead of splitting independent subspace samples (which are not available), we split the underlying trials prior to computing the subspaces. For each subject and time window, trials were randomly divided into two halves, defining training and test folds. Crucially, subspaces (PC scores) were computed independently within each fold, ensuring that training and test representations were derived from non-overlapping data. Within each fold, we applied bootstrap resampling (sampling with replacement) to generate 500 pseudo-independent X matrices of neural activity. Rank-specific subspaces were recomputed for each resample, yielding a distribution of subspace representations per rank, separately for the training and test folds. A linear classifier was then trained on the training-fold subspaces to decode ordinal position and evaluated on the test-fold subspaces. Statistical significance was assessed using a permutation test in which class labels in the training fold were randomly shuffled, and the full decoding procedure was repeated 500 times to construct a null distribution of accuracy. Finally, the entire pipeline—including trial splitting, resampling, subspace estimation, decoding, and permutation testing—was repeated 500 times. Reported accuracy corresponds to the mean empirical decoding accuracy across iterations, while variability (SEM) was computed by first estimating subject-level variability within each iteration and then averaging these SEM estimates across iterations. Statistical significance was assessed using a pooled permutation approach, performed separately for each time window. Permutation-based decoding accuracies were aggregated across all iterations to construct a single null distribution per time window. The *p*-value was computed as the proportion of permutation samples exceeding the mean empirical accuracy. Multiple comparisons across time windows were controlled using false discovery rate (FDR) correction. Importantly, by performing the split prior to subspace estimation, this procedure constitutes a form of 2-fold cross-validation that preserves independence between training and test data at the trial level, thereby preventing information leakage while enabling reliable decoding from subspace geometry.

We conducted several supplementary analyses to assess the robustness of our findings. First, neural subspaces were estimated by retaining the leading six and eight Principal Components instead of three, explaining 82 ± 3% and 88 ± 2% of the variance, respectively. For subspaces of dimensionality greater than three, the normal vector approach to computing the principal angle (PA) is not applicable, as there is no single vector orthogonal to the entire subspace. Instead, we computed principal angles between subspaces using a method based on singular value decomposition (SVD). Specifically, we first obtained an orthonormal basis for each subspace using QR decomposition. We then calculated the matrix of inner products between the basis vectors of the two subspaces and computed the singular values of this matrix, corresponding to the cosines of the principal angles between the subspaces. Because each subspace is defined as a two-dimensional plane in $R^n$ (being $n > 3$), this method returns two principal angles per pair of subspaces. We report both the minimum and maximum of these two angles as summary measures of subspace alignment. The choice of six and eight principal components was motivated by geometric constraints: in $R^6$ and $R^8$, respectively, it becomes possible for three and four two-dimensional subspaces to be mutually orthogonal—that is, for all pairs of subspaces to have both principal angles equal to 90 degrees. This level of orthogonality is not achievable in lower-dimensional spaces due to the overlap required between subspaces. While increasing the number of principal components expands the ambient space and theoretically allows for fully orthogonal subspaces (i.e., two principal angles of 90° between each pair of planes), this comes with an important cost. Principal angles are computed between two-dimensional planes, each defined by two vectors in $R^n$, where n is the number of retained principal components. These vectors are obtained as the leading two eigenvectors from a second PCA applied to the PC scores corresponding to each ordinal position (i.e., the $Z_r$ matrix described earlier). In lower-dimensional spaces such as $R^3$, the first two eigenvectors typically explain a large proportion of the variance (85 ± 1%), making the plane a faithful representation of the underlying subspace geometry. However, as the dimensionality increases, the variance captured by the first two eigenvectors decreases (e.g., 65 ± 2% in $R^6$ and 58 ± 2% in $R^8$), meaning that the 2D planes used to compute principal angles become progressively less representative of the full subspace structure. As a second supplementary analysis, we addressed whether our results on low-dimensional rank-specific orthogonal subspaces were not driven by shared task-epoch components. We recomputed subspaces with the leading PCs 2-4 and 3-5, hence not considering the first components, which might reflect shared global components. Moreover, we also regressed out global components at the level of source time series. Specifically, we concatenated source time series across trials, computed PCA on the source time series (time by channels), reconstructed the time courses of the leading PCs (first, and first and second jointly), and regressed out those global time courses from the original time series. Then, we applied the geometric analysis of neural subspaces to the residualized source time series. Third, to mitigate the impact of limited sample size on the estimation of the X matrices, we reduced the number of stimulus locations. Rather than using all 8 locations for each ordinal position, we recomputed the X matrices using only 4 locations by grouping pairs of neighboring positions around the cardinal axes. Because each location could be grouped with either its left-side or right-side neighbor, we applied both grouping schemes and averaged the resulting matrices. This procedure yielded X matrices centered on the four cardinal locations. This analysis was motivated by the limited number of observations available to estimate the X matrix for incorrect trials. For sequences of length 3 (4), the number of trials per condition was 3 ± 2 [range: 0–9] (5 ± 3 [range: 0–17]), with 54% (25%) of conditions having fewer than 3 trials and 10% (3%) having no trials at all. After grouping locations, the number of trials per condition increased to 5 ± 3 [range: 1–16] (9 ± 5 [range: 1–29]), with only 18% (5%) of conditions having fewer than 3 trials, and no conditions with zero trials. The fourth supplementary analysis involved repeating the entire geometric analysis pipeline using MEG sensors (N = 207) as channels, instead of cortical parcels in source space.

### Decoding analysis and sequence effects

The analysis of neural sequences of memory representations during the delay period was performed with the temporally delayed linear modeling (TDLM)[26], a domain-general method for finding neural sequences. TDLM has been validated for identifying neural replay in MEG recordings of humans both at rest and task[28,39,59]. The TDLM process involves, first, constructing the decoded state space, and then analyzing the neural sequences within that space.

To construct the decoded state space of memory representations, we trained binomial logistic regression classifiers, one for each specific stimulus location. Classifiers were trained separately for the localizer and WM task, using the 200 source-based parcels (or the 207 MEG sensors) as data features. Training was performed at each time point of stimulus presentation. To improve the signal-to-noise ratio, we downsampled the time series into 50 ms bins with 5 ms steps. The training set included: (a) positive examples, when the target stimulus was displayed on the screen, and (b) negative examples, when any

other stimulus was displayed. To balance the sample sizes, negative examples were created by randomly selecting stimuli from all other locations, ensuring an equal number of negative examples for each location. The logistic regression models were L1- and L2-regularized, with lambda values of [1, 0.1, 0.01, 0.001], and classifier accuracy was evaluated using a 5-fold cross-validation approach with 10 repetitions. For group-level statistical analysis, we calculated the area under the curve (AUC) values for each subject at every time point. To determine if the AUC values were significantly different from chance (AUC = 0.5), we performed a cluster-based permutation test. This non-parametric method was chosen for its ability to handle multiple comparisons over time points while controlling for false positives[60,61]. We applied a correction for multiple comparisons using a cluster-based approach, with 5000 permutations, 'maxsum' cluster statistic, a critical z-value for identifying significant clusters was set at 2.58, and the significance threshold (alpha) was set to 0.01 (two-sided). At the group level, we selected the timepoint and the regularization (L1 or L2, and the lambda parameter) with the highest performance (AUC). Then, we used the subject-specific weights of the winning classifiers to predict memory reactivations during the delay period, getting the decoded state space. Since the sequences presented never included all 8 locations, we defined the state space with the locations presented in each trial, resulting in a decoded state space of dimensions time by sequence length. Delay period time series were also downsampled, averaging across bins of 50 msec in steps of 5 msec, to increase the signal-to-noise ratio. Decoding analysis was performed with MVPA-Light[62].

The temporal structure of working memory representations was assessed on the decoded state space in a two-step linear modeling approach, as described in ref. 26. In the first-level sequence analysis, we quantified the state-to-state transitions across time lags, defined as the time lapse between neural reactivations in the sequence. For each time lag $\Delta t$, ranging from 1 to 500 ms in steps of 5 ms, we fitted the regression model

$$X_j(t+\Delta t) = \sum_{k=1}^{n} X_k(t)\beta_{kj} \quad (1)$$

where the time-lagged copy of one state $X_j(t+\Delta t)$ was regressed onto the state space $X_k(t)$ of all the states. When the process is completed for all states, we get a matrix of regression coefficients beta (states by states), named the empirical transition matrix, defining the state-to-state transitions for the time lag $\Delta t$. This process was applied separately for each trial, and the regression coefficients were averaged across trials (trial-averaged TDLM approach). In the second-level sequence analysis, we tested the strength of particular sequences, defined by the transition matrices $T_r$, with a linear regression model that relates them to $\beta$. Specifically,

$$\boldsymbol{\beta} = \sum_{r=1}^{r} Z(r) * T_r \quad (2)$$

where $T_r = [T_F, T_B, T_{auto}, T_{const}]$ was a set of predefined template matrices, including the hypothesized sequence of events running forward $T_F$ and backward $T_B$, and another two matrices controlling for state autocorrelation $T_{auto}$ and the average of all transitions $T_{const}$. The regression coefficients $Z(r)$ reflect the strength of the sequences, or sequence effect. Statistical significance of forward and backward effects was tested with a non-parametric permutation-based method, creating null distributions by reshuffling the order of the states and recalculating the second-level analysis. Multiple comparisons over time lags were addressed with extreme-value or maxT correction, where the null distribution was formed by taking the maximum of sequenceness across all computed time lags for each permutation. The significance threshold across time lags was computed as the 2.5 and 97.5 percentiles of the null distribution. See[26] for a detailed description of TDLM steps.

We tested sequence effects during the delay period using decoded state spaces generated by classifiers from both the localizer (localizer-to-delay) and the stimulus period of the working memory task (stim-to-delay). To further validate the accuracy of the analysis pipeline, we also applied the TDLM to the stimulus presentation with classifiers trained on the WM stimulus period (stim-to-stim) and on localizer data (localizer-to-stim). Sequence effects were computed at trial level, then averaged across trials and subsequently across subjects. To control for finite sample effects during the first-level sequence analysis, we alternatively concatenated the trial-specific decoded state space and performed the analysis at the subject level (across-trials concatenated TDLM approach). This approach increased the number of observations in the initial regression, aligning more closely with seminal studies on neural replay during rest[27,28]. To account for inter-trial effects, we used a linear mixed-effects model in the first-level sequence analysis to include random intercepts for trial and session. To prevent cross-trial contamination in the time-lagged decoded state, we set to zero the timepoints (rows of X) where trial boundaries were crossed. As a further step, we also tested eliminating these timepoints (rows of X) entirely, which produced equivalent results.

To further validate our decoding approach for detecting memory reactivations during the delay period, we compared classifier outputs (decision values) at the single-trial level for stimuli that were in-sequence versus out-of-sequence. As described above, each binary classifier was trained to distinguish one stimulus identity (spatial location) from a balanced mixture of all other identities. Thus, decoding accuracy reflects the ability to detect the presence of a given stimulus identity, rather than to discriminate between specific stimulus pairs. Given the high dimensionality of the decision value data—organized as tensors over subjects, sessions, stimuli, trials, and time, separately for in-sequence and out-of-sequence conditions—we implemented a hierarchical analysis. At the first level (within-subject), decision values were pooled across sessions and stimulus sets (length 3 and length 4), and concatenated across all trials and time points. We then compared in-sequence and out-of-sequence decision values using an independent-samples t-test, yielding one t-statistic per subject. At the second level (group level), these subject-specific t-values were entered into a one-sample t-test against zero to assess whether, across subjects, classifier evidence was systematically stronger for in-sequence than out-of-sequence stimuli. This hierarchical approach allowed us to account for within-subject variability while enabling population-level inference on memory reactivation signals.

To assess whether the absence of delay period sequencing could be explained by insufficient SNR, we performed an injection-based sensitivity analysis using the empirical delay period classifier decision values. This approach preserved the empirical noise structure while allowing the amplitude of a known sequential signal to be systematically manipulated. We first estimated the empirical magnitude of the in-sequence versus out-of-sequence decision-value difference separately for each subject and session. For each trial and time point, we averaged the decision values across out-of-sequence stimuli and subtracted this value from the decision value of each in-sequence stimulus. This resulted in a trial × time × stimulus distribution of in-versus-out decision-value differences. Because memory reactivations are expected to be transient, rather than continuously expressed throughout the delay, we characterized this distribution using its 85th, 90th, and 95th percentiles. These percentiles capture the magnitude of the in-versus-out difference among the upper 15%, 10%, and 5% of trial × time × stimulus observations, respectively. Given the 4-s delay period, these

correspond approximately to 600, 400, and 200 ms of delay period observations. The resulting percentile-specific values were used as estimates of the empirical reactivation amplitude. For the injection analysis, we added the estimated empirical amplitude to the decision values of the in-sequence stimuli according to predefined sequences separated by a fixed temporal lag. Sequence onsets were randomized independently across trials, and the temporal separation between successive injected sequences was also randomized to avoid systematic alignment of injections with multiples of the tested lag. The injected sequences were added to the empirical decision values without otherwise modifying the data, thereby preserving the original noise and temporal structure. We then applied the complete trial-averaged TDLM analysis to each simulated dataset and assessed whether the injected sequence was recovered significantly at the expected temporal lag. Sensitivity was evaluated across injection amplitudes ranging from 0.5 to 1.5 times the corresponding empirical amplitude, in increments of 0.1, and across temporal lags of 100, 200, and 300 ms. For each combination of percentile, injection amplitude, lag, and sequence length, the procedure was repeated for 50 iterations, with sequence onsets independently randomized across trials in each iteration. Sensitivity was defined as the proportion of iterations in which the injected sequence produced a significant TDLM effect at the expected lag. Sensitivity curves therefore represent the probability of detecting a known sequential signal as a function of its amplitude relative to the empirically observed in-versus-out difference. The analysis was performed using the trial-averaged TDLM approach, which showed the most consistent sequenceness effects during stimulus presentation. This analysis allows us to determine the amplitude at which TDLM becomes sensitive to sequential reactivation under the empirical noise structure. If TDLM reliably detects injected sequences at amplitudes corresponding to, or below, the empirically observed in-versus-out difference ($\leq 1$ on the x-axis of the sensitivity plots), this would indicate that the null TDLM result is unlikely to be explained simply by insufficient SNR. Conversely, if reliable detection requires injection amplitudes substantially larger than the empirically observed difference ($>1$ on the x-axis), insufficient SNR would remain a plausible explanation for the absence of a significant TDLM effect.

We further quantified evidence for sequenceness using Bayesian one-sample t-tests with default JZS priors[63,64]. For each TDLM variant and sequence length, separately for forward and backward sequenceness, Bayes factors were computed independently at each time lag across subjects. The prior scale parameter was defined using the standard deviation of a global null distribution constructed across all subjects, sequence lengths, and lags. Using a single global null distribution provides a common reference scale for all Bayesian tests, ensuring that Bayes factors are directly comparable across TDLM approaches, sequence directions, sequence lengths, and temporal lags. We report the transformed statistic $2*\log_e(BF)$, referred as logBF from now on, where positive values indicate evidence for non-zero sequenceness ($BF_{10}$), and negative values indicate evidence for the null hypothesis of zero sequenceness ($BF_{01} = 1/BF_{10}$). Following the interpretation criteria described in ref. 65, logBF values of 0–2 correspond to evidence not worth more than a bare mention (BF = 1–3), logBF values of 2–6 to positive evidence (BF = 3–20), logBF values of 6–10 to strong evidence (BF = 20–150), and logBF values > 10 to very strong evidence (BF > 150). Negative values follow the same scale in favor of the null hypothesis.

Analyses performed locally used MATLAB R2022a and R version 4.2.1 (2022-06-23), with ggplot2 version 3.4.4. Analyses performed on the JUBAIL High-Performance Computing cluster at New York University Abu Dhabi (NYUAD) used MATLAB R2022a.

### Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

## Data availability

The data generated in this study are available from the Open Science Framework (OSF) repository https://osf.io/hvq8p and in the Source Data file[66]. The OSF repository contains the intermediate input data used for the analyses. For the neural geometry pipeline, these include the neural activity matrices (X) used to compute the neural subspaces. For the neural sequences pipeline, these include the decoded state spaces (classifier prediction outputs) used as inputs to the Temporal Delayed Linear Modeling (TDLM) analyses. The corresponding analysis outputs are provided in the Source Data file. Together, the data provided in the OSF repository and Source Data file allow reproduction of all statistical analyses and figures reported in the study. The raw MEG recordings and the preprocessed data from which these intermediate inputs were derived are not available due to data privacy laws. Source data are provided with this paper.

## Code availability

The code used for analyses in this study is publicly available at github.com/asantoangles/geometry_working_memory[67].

## References


1. D'Esposito, M. & Postle, B. R. The cognitive neuroscience of working memory. *Annu. Rev. Psychol.* **66**, 115–142 (2015).
2. Goldman-Rakic, P. S. Cellular basis of working memory. *Neuron* **14**, 477–485 (1995).
3. Riley, M. R. & Constantinidis, C. Role of prefrontal persistent activity in working memory. *Front. Syst. Neurosci.* **9**, 181 (2015).
4. Lundqvist, M. et al. Gamma and beta bursts underlie working memory. *Neuron* **90**, 152–164 (2016).
5. Lundqvist, M., Herman, P., Warden, M. R., Brincat, S. L. & Miller, E. K. Gamma and beta bursts during working memory readout suggest roles in its volitional control. *Nat. Commun.* **9**, 1–12 (2018).
6. Rose, N. S. et al. Reactivation of latent working memories with transcranial magnetic stimulation. *Science* **354**, 1136–1139 (2016).
7. Barbosa, J. et al. Interplay between persistent activity and activity-silent dynamics in the prefrontal cortex underlies serial biases in working memory. *Nat. Neurosci.* **23**, 1016–1024 (2020).
8. Stokes, M. G. "Activity-silent" working memory in prefrontal cortex: a dynamic coding framework. *Trends Cogn. Sci.* **19**, 394–405 (2015).
9. Panichello, M. F. et al. Intermittent rate coding and cue-specific ensembles support working memory. *Nature* **636**, 422–429 (2024).
10. Cueva, C. J. et al. Low-dimensional dynamics for working memory and time encoding. *Proc. Natl. Acad. Sci. Usa.* **117**, 23021–23032 (2020).
11. Murray, J. D. et al. Stable population coding for working memory coexists with heterogeneous neural dynamics in prefrontal cortex. *Proc. Natl. Acad. Sci. USA* **114**, 394–399 (2017).
12. Spaak, E., Watanabe, K., Funahashi, S. & Stokes, M. G. Stable and dynamic coding for working memory in primate prefrontal cortex. *J. Neurosci.* **37**, 6503–6516 (2017).
13. Panichello, M. F. & Buschman, T. J. Shared mechanisms underlie the control of working memory and attention. *Nature* **592**, 601–605 (2021).
14. Tian, Z. et al. Mental programming of spatial sequences in working memory in the macaque frontal cortex. *Science* **385**, eadp6091 (2024).
15. Xie, Y. et al. Geometry of sequence working memory in macaque prefrontal cortex. *Science* **375**, 632–639 (2022).

16. Shine, J. M. et al. Human cognition involves the dynamic integration of neural activity and neuromodulatory systems. *Nat. Neurosci.* **22**, 289–296 (2019).
17. Gao, S., Mishne, G. & Scheinost, D. Nonlinear manifold learning in functional magnetic resonance imaging uncovers a low-dimensional space of brain dynamics. *Hum. Brain Mapp.* **42**, 4510–4524 (2021).
18. Kriegeskorte, N., Mur, M. & Bandettini, P. Representational similarity analysis - connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* **2**, 1–28 (2008).
19. Kriegeskorte, N. & Kievit, R. A. Representational geometry: integrating cognition, computation, and the brain. *Trends Cogn. Sci.* **17**, 401–412 (2013).
20. Diedrichsen, J. & Kriegeskorte, N. Representational models: a common framework for understanding encoding, pattern-component, and representational-similarity analysis. *PLoS Comput. Biol.* **13**, e1005508 (2017).
21. Kriegeskorte, N. & Wei, X.-X. Neural tuning and representational geometry. *Nat. Rev. Neurosci.* **22**, 703–718 (2021).
22. Li, H.-H. & Curtis, C. E. Neural population dynamics of human working memory. *Curr. Biol.* **33**, 3775–3784.e4 (2023).
23. Xu, Y. The human posterior parietal cortices orthogonalize the representation of different streams of information concurrently coded in visual working memory. *PLoS Biol.* **22**, e3002915 (2024).
24. Foster, D. J. & Wilson, M. A. Reverse replay of behavioural sequences in hippocampal place cells during the awake state. *Nature* **440**, 680–683 (2006).
25. Tacikowski, P., Kalender, G., Ciliberti, D. & Fried, I. Human hippocampal and entorhinal neurons encode the temporal structure of experience. *Nature* **635**, 160–167 (2024).
26. Liu, Y. et al. Temporally delayed linear modelling (TDLM) measures replay in both animals and humans. *Elife* **10**, 66917 (2021).
27. Kurth-Nelson, Z., Economides, M., Dolan, R. J. & Dayan, P. Fast sequences of non-spatial state representations in humans. *Neuron* **91**, 194–204 (2016).
28. Liu, Y., Dolan, R. J., Kurth-Nelson, Z. & Behrens, T. E. J. Human replay spontaneously reorganizes experience. *Cell* **178**, 640–652.e14 (2019).
29. Lisman, J. E. & Jensen, O. The Theta-Gamma neural code. *Neuron* **77**, 1002–1016 (2013).
30. Axmacher, N. et al. Cross-frequency coupling supports multi-item working memory in the human hippocampus. *Proc. Natl. Acad. Sci. Usa.* **107**, 3228–3233 (2010).
31. Heusser, A. C., Poeppel, D., Ezzyat, Y. & Davachi, L. Episodic sequence memory is supported by a theta-gamma phase code. *Nat. Neurosci.* **19**, 1374–1380 (2016).
32. Michelmann, S., Bowman, H. & Hanslmayr, S. The temporal signature of memories: Identification of a general mechanism for dynamic memory replay in humans. *PLoS Biol.* **14**, e1002528 (2016).
33. Huang, Q., Jia, J., Han, Q. & Luo, H. Fast-backward replay of sequentially memorized items in humans. *Elife* **7**, 376202 (2018).
34. Jafarpour, A., Penny, W., Barnes, G., Knight, R. T. & Duzel, E. Working memory replay prioritizes weakly attended events. *eNeuro* **4**, 171 (2017).
35. Hari, R., Puce, A., Hari, R. & Puce, A. 3 Introduction. In *MEG-EEG Primer* (Oxford University Press, 2017).
36. Schaefer, A. et al. Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. *Cereb. Cortex* **28**, 3095–3114 (2018).
37. Colclough, G. L., Brookes, M. J., Smith, S. M. & Woolrich, M. W. A symmetric multivariate leakage correction for MEG connectomes. *Neuroimage* **117**, 439–448 (2015).
38. Barbosa, J. et al. Quantifying differences in neural population activity with shape metrics. *bioRxiv* 2025.01. 10.632411 (2025) https://doi.org/10.1101/2025.01.10.632411.
39. Schwartenbeck, P. et al. Generative replay underlies compositional inference in the hippocampal-prefrontal circuit. *Cell* **186**, 4885–4897.e14 (2023).
40. Sreenivasan, K. K. & D'Esposito, M. The what, where and how of delay activity. *Nat. Rev. Neurosci.* **20**, 466–481 (2019).
41. Pasternak, T. & Greenlee, M. W. Working memory in primate sensory systems. *Nat. Rev. Neurosci.* **6**, 97–107 (2005).
42. D'Esposito, M. From cognitive to neural models of working memory. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **362**, 761–772 (2007).
43. Sreenivasan, K. K., Gratton, C., Vytlacil, J. & D'Esposito, M. Evidence for working memory storage operations in perceptual cortex. *Cogn. Affect. Behav. Neurosci.* **14**, 117–128 (2014).
44. Libby, A. & Buschman, T. J. Rotational dynamics reduce interference between sensory and memory representations. *Nat. Neurosci.* **24**, 715–726 (2021).
45. Shine, J. M. et al. The low-dimensional neural architecture of cognitive complexity is related to activity in medial thalamic nuclei. *Neuron* **104**, 849–855.e3 (2019).
46. Barbosa, J., Lozano-Soldevilla, D. & Compte, A. Pinging the brain with visual impulses reveals electrically active, not activity-silent, working memories. *PLoS Biol.* **19**, e3001436 (2021).
47. Wolff, M. J., Jochim, J., Akyürek, E. G. & Stokes, M. G. Dynamic hidden states underlying working-memory-guided behavior. *Nat. Neurosci.* **20**, 864–871 (2017).
48. Yang, C., He, X. & Cai, Y. Reactivating and reorganizing activity-silent working memory: two distinct mechanisms underlying pinging the brain. *Cereb. Cortex* **35**, 494 (2025).
49. Liu, Y., Nour, M. M., Schuck, N. W., Behrens, T. E. J. & Dolan, R. J. Decoding cognition from spontaneous neural activity. *Nat. Rev. Neurosci.* **23**, 204–214 (2022).
50. Wang, L. et al. Representation of spatial sequences using nested rules in human prefrontal cortex. *Neuroimage* **186**, 245–255 (2019).
51. Al Roumi, F., Marti, S., Wang, L., Amalric, M. & Dehaene, S. Mental compression of spatial sequences in human working memory using numerical and geometrical primitives. *Neuron* **109**, 2627–2639.e4 (2021).
52. Liebe, S. et al. Phase of firing does not reflect temporal order in sequence memory of humans and recurrent neural networks. *Nat. Neurosci.* **28**, 873–882 (2025).
53. Fang, W., Jiang, X., Chen, J., Zhang, C. & Wang, L. Oscillatory control over representational geometry of sequence working memory in macaque frontal cortex. *Curr. Biol.* **35**, 1495–1507.e5 (2025).
54. Oostenveld, R., Fries, P., Maris, E. & Schoffelen, J. M. FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Comput. Intell. Neurosci.* **2011**, 156869 (2011).
55. Ferrante, O. et al. FLUX: a pipeline for MEG analysis. *Neuroimage* **253**, 119047 (2022).
56. Smith, S. M. et al. Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage* **23**, 208–219 (2004).
57. Westner, B. U. et al. A unified view on beamformers for M/EEG source reconstruction. *Neuroimage* **246**, 118789 (2022).
58. Nolte, G. The magnetic lead field theorem in the quasi-static approximation and its use for magnetoenchephalography forward calculation in realistic volume conductors. *Phys. Med. Biol.* **48**, 3637–3652 (2003).
59. Liu, Y., Mattar, M. G., Behrens, T. E. J., Daw, N. D. & Dolan, R. J. Experience replay is associated with efficient nonlocal learning. *Science* **372**, 1357 (2021).
60. Maris, E. & Oostenveld, R. Nonparametric statistical testing of EEG- and MEG-data. *J. Neurosci. Methods* **164**, 177–190 (2007).
61. Sassenhagen, J. & Draschkow, D. Cluster-based permutation tests of MEG/EEG data do not establish significance of effect latency or location. *Psychophysiology* **56**, 1–8 (2019).

62. Treder, M. S. MVPA-Light: a classification and regression toolbox for multi-dimensional data. *Front. Neurosci.* **14**, 289 (2020).
63. Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D. & Iverson, G. Bayesian t tests for accepting and rejecting the null hypothesis. *Psychon. Bull. Rev.* **16**, 225–237 (2009).
64. Wagenmakers, E.-J. et al. Bayesian inference for psychology. Part I: theoretical advantages and practical ramifications. *Psychon. Bull. Rev.* **25**, 35–57 (2018).
65. Kass, R. E. & Raftery, A. E. Bayes factors. *J. Am. Stat. Assoc.* **90**, 773–795 (1995).
66. Santo-Angles A., et al. (2026). Source Data file. Neural Subspaces Encode Sequential Working Memory Without Detectable Neural Sequences. Nature Communications. Zenodo https://doi.org/10.5281/ZENODO.22670599
67. Santo-Angles, A. *Neural Subspaces Encode Sequential Working Memory Without Detectable Neural Sequences. Geometry_working_memory*. (Zenodo, 2026). https://zenodo.org/records/22655738

## Acknowledgements

This research was conducted using resources provided by the High-Performance Computing Cluster and the Core Technology Platform at NYUAD.

## Author contributions

A.S.A. designed the study, collected, preprocessed, and analyzed the data, and wrote the manuscript. K.K.S. co-designed the study and revised the manuscript. G.W.L. contributed to the study design and revised the manuscript. J.Y., Y.Z. and W.K.H.C. contributed to data collection and preprocessing.

## Funding

This work was supported by the NYU Research Enhancement Fund and the NYUAD Center for Brain and Health, funded by Tamkeen under NYUAD Research Institute grant CG012 (KS).

## Competing interests

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Additional information

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41467-026-78895-x.

**Correspondence** and requests for materials should be addressed to Aniol Santo-Angles.

**Peer review information** *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available

**Reprints and permissions information** is available at http://www.nature.com/reprints

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2026

This Article in Press is shared early to give you faster access to new research. It is citable and carries a permanent DOI. The final edited version will replace it automatically.