

Supplementary Information for Sensory encoding and memory retrieval are coordinated with propagating waves in the human brain

Yifan Yang¹, David A. Leopold^{3,4}, Jeff H. Duyn⁵, Xiao Liu^{1,2*}

¹Department of Biomedical Engineering, The Pennsylvania State University, University Park, 16802, PA, USA.

²Institute for Computational and Data Sciences, The Pennsylvania State University, University Park, 16802, PA, USA.

³Neurophysiology Imaging Facility, National Institute of Mental Health, National Institute of Neurological Disorders and Stroke, and National Eye Institute, National Institutes of Health, Bethesda, 20892, MD, USA.

⁴Section on Cognitive Neurophysiology and Imaging, Laboratory of Neuropsychology, National Institute of Mental Health, National Institutes of Health, Bethesda, 20892, MD, USA.

⁵Advanced MRI Section, Laboratory of Functional and Molecular Imaging, National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, 20892, MD, USA.

*Corresponding author(s). E-mail(s): xx1213@psu.edu;
Contributing authors: zy161@psu.edu; leopoldd@mail.nih.gov;
jeff.duyn@nih.gov;

Contents

	Page
1 Materials and Methods	3
1.1 Semantic Decoding Model	3
1.1.1 fMRI Encoder	3
1.1.2 Caption Decoder	3
1.1.3 Representation Similarity Analysis	3
1.1.4 Semantic Similarity Metrics	4
1.1.5 Region-wise Decoding	4
1.2 Infra-slow Propagating Waves	4
1.2.1 Delay-profile Decomposition	4
1.2.2 fMRI SM-to-DMN Propagating Waves	5
1.2.3 Spiking Cascade	5
2 Supplementary Figures	6
2.1 Figure S1	6
2.2 Figure S2	7
2.3 Figure S3	8
2.4 Figure S4	9

38	2.5	Figure S5	10
39	2.6	Figure S6	11
40	2.7	Figure S7	12
41	2.8	Figure S8	13
42	2.9	Figure S9	14
43	2.10	Figure S10	15
44	2.11	Figure S11	16
45	2.12	Figure S12	17
46	3	Supplementary Tables	18
47	3.1	Table S1	18
48	3.2	Table S2	19
49	3.3	Table S3	19

1 Materials and Methods

1.1 Semantic Decoding Model

The semantic decoding model is designed to evaluate the semantic information contained in the stimuli-evoked BOLD responses, generating text captions that describe the image stimuli presented to the subject. The decoding model consists of two main components: an fMRI encoder and a caption decoder (Figure 2A).

1.1.1 fMRI Encoder

The fMRI encoder is used to extract latent representation from the BOLD response. The detailed architecture of the fMRI encoder, as shown in Figure S6A, utilizing convolutional layers and residual connections, aims to transform the high-dimensional BOLD response into 512-dimensional fMRI embeddings. To address the challenge of fMRI data scarcity, the fMRI encoder was trained to align the fMRI embeddings with the CLIP embedding space which has been extensively pre-trained using 400 million (image, text) pairs, thus offering a rich, 512-dimensional target. Therefore, we trained the fMRI encoder in contrastive learning paradigm [31], aiming to maximize the alignment between the fMRI embeddings and the corresponding CLIP text embeddings, while minimizing the alignment with mismatched pairing. To achieve this, we use the contrastive training loss with loss function defined for i^{th} fMRI embedding $\mathbf{Z}_i$ and j^{th} CLIP text embedding $\mathbf{T}_j$ within a batch B as:

$$\mathcal{L}_{\text{contrast}}(\mathbf{Z}_i, \mathbf{T}_j) = -\log \frac{\exp(\text{sim}(\mathbf{Z}_i, \mathbf{T}_j)/\tau)}{\sum_{i,j \in B, i \neq j} \exp(\text{sim}(\mathbf{Z}_i, \mathbf{T}_j)/\tau)} \quad (1)$$

Here, τ represents the temperature, a hyperparameter, and $\text{sim}(\cdot, \cdot)$ computes vector similarity, with cosine similarity being applied in this instance:

$$\mathcal{L}_{\text{contrast}}(\mathbf{Z}_i, \mathbf{T}_j) = -\log \frac{\exp(\cos(\mathbf{Z}_i, \mathbf{T}_j)/\tau)}{\sum_{i,j \in B, i \neq j} \exp(\cos(\mathbf{Z}_i, \mathbf{T}_j)/\tau)}. \quad (2)$$

In addition, we also maximize the alignment between fMRI embeddings and CLIP embeddings by incorporating the cosine loss defined as

$$\mathcal{L}_{\text{align}}(\mathbf{Z}_i, \mathbf{T}_i) = \lambda_1(1 - \cos(\mathbf{Z}_i, \mathbf{T}_i)) \quad (3)$$

Therefore, the total loss is defined as the summation of alignment loss and contrastive loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{align}} + \lambda_2 \mathcal{L}_{\text{contrast}} \quad (4)$$

where λ_1 and λ_2 are tuning hyperparameters. In our training regimen, we set $\lambda_1 = 0.35$, $\lambda_2 = 0.65$ and $\tau = 0.45$. We also incorporate dropout before the final layers, with a dropout ratio of 0.3.

1.1.2 Caption Decoder

We utilized a pre-trained caption decoder, DeCap [29], to generate captions from fMRI embeddings. DeCap was initially trained to produce captions using CLIP text embeddings based on large text corpus. Since the fMRI embeddings were aligned with the CLIP space through the fMRI encoder, we directly employed DeCap to decode these fMRI embeddings, for the generation of captions that describe the content of image stimuli shown to the subjects.

1.1.3 Representation Similarity Analysis

We analyzed the alignment between CLIP text embeddings and fMRI embeddings through representation similarity analysis (RSA). For each validation fold, we constructed representation dissimilarity matrices (RDMs) for both the fMRI and CLIP text embeddings. To facilitate visualization and interpretation of the RDMs' structure, we employed t-SNE techniques [33] to project CLIP text embeddings into a 2-dimensional space. Subsequently, we applied k-means

clustering to group stimuli with minimal Euclidean distance in this 2-dimensional representation into the same cluster. The RDMs for both fMRI and CLIP text embeddings, as depicted in Figure 2A, are organized and sorted according to this clustering scheme. The similarity between two RDMs is quantified using Pearson’s correlation coefficient.

1.1.4 Semantic Similarity Metrics

To evaluate the fidelity of generated captions to their semantic content, we computed semantic similarity score between the predicted captions and ground truth captions labeled by human. This evaluation employs several established metrics widely used in computer vision and natural language processing research: Bleu.1, Bleu.2 [101], CIDEr [102], METEOR [103], and ROUGE-L [104], with each of these metrics offering a different perspective on the semantic alignment between generated and ground truth text. Given the unique limitation inherent to each metric, we additionally defined a composite semantic metric (CSS) derived by averaging the scores from the aforementioned metrics, thereby providing a more holistic evaluation of caption fidelity.

To assess the statistical significance of the model-generated captions, for each trial/image, the semantic similarity score, for example, CSS score, is compared against a null distribution. This null distribution is constructed from the CSS scores between the ground truth caption and 1,000 randomly generated captions, which are generated by randomly sampling from the CLIP embedding space [31]. Notably, this null distribution exhibits variability across different trials, reflecting the diverse complexity levels associated with the semantic content of each image stimulus. Such approach ensures the evaluations accurately reflecting the model’s performance in generating semantically coherent captions and invariant to the varying degrees of semantic complexity present across trials.

The same evaluative framework is applied across all aforementioned metrics to assess the accuracy of the generated captions. A trial is considered correctly decoded if the metric score is significant compared to the null distribution, using a significance level of 0.95; otherwise, it is considered incorrectly decoded. Encoding accuracy is quantified as the proportion of trials that are correctly decoded.

1.1.5 Region-wise Decoding

In our study, unless otherwise specified, we use the decoding model to decode the response of voxels within the "nsdgeneral" ROI which includes occipital regions that are generally responsive in the NSD experiment [30].

For regionwise decoding analysis as shown in Figure 4 and 12, we adopted the regions of interest (ROIs) as delineated by the Yeo-17 network [38]. For each specific ROI, only the voxels that are defined by the corresponding ROI mask are considered. The responses from these selected voxels are then utilized as inputs to the decoding model to generate captions.

1.2 Infra-slow Propagating Waves

1.2.1 Delay-profile Decomposition

Following prior studies [7, 13] for analyzing the temporal dynamics of neural signals, we apply delay-profile decomposition method to identify the activation phase of particular regions or neuron that are activated in relation to specific events. Briefly, for each candidate event k , we derived a **delay profile** vector $d_k \in \mathbb{R}^N$ representing the activation phases of each region or neuron within that event:

$$\mathbf{d}_k = (t_{1k} \ t_{2k} \ \cdots \ t_{Nk})^T \quad (5)$$

where $t_{i,k}$ represents the temporal centroid of the signal within the time segment k for the i -th region/neuron and N represents the total number of regions/neurons.

To represent the entire set of temporal relationships among the regions/neurons relative to each event segment analyzed, we combined the delay profiles for all M segments into a delay profile matrix $\mathbf{D} \in \mathbb{R}^{M \times N}$:

$$\mathbf{D} = (\mathbf{d}_1 \ \mathbf{d}_2 \ \cdots \ \mathbf{d}_M) \quad (6)$$

$$= \begin{pmatrix} t_{11} & t_{12} & \cdots & t_{1N} \\ t_{21} & t_{22} & \cdots & t_{2N} \\ \vdots & \vdots & & \vdots \\ t_{M1} & t_{M2} & \cdots & t_{MN} \end{pmatrix} \quad (7)$$

Singular value decomposition (SVD) was then applied to the delay matrix $\mathbf{D}$ with the resulting principle component $\mathbf{u}^* \in \mathbb{R}^N$ defined as the **principal delay profile (PD)** representing the major sequential organization among the collective regions or neurons.

1.2.2 fMRI SM-to-DMN Propagating Waves

To detect SM-to-DMN propagating waves, we adopt a template-matching approach [7]. The propagations, which typically involve a majority of cortical regions, are assumed entrained within the fMRI global signal fluctuations. Accordingly, we defined set of candidate events based on the low-frequency ($<0.15\text{Hz}$) components of the global signal, with the event boundaries defined by the adjacent troughs. This segmentation ensures that each candidate event contains a peak of global neural activity flanked by periods of relative quiescence.

For each candidate event k , we derive a delay profile as described previously. A candidate event k is then considered as SM-to-DMN propagation if its delay profile closely matches the principal gradient profile (PG) $\mathbf{u}_{PG} \in \mathbb{R}^N$ [22]. The degree of similarity is quantitatively assessed using Pearson's correlation coefficient:

$$r_k = \text{PearsonCorr}(\mathbf{d}_k, \mathbf{u}_{PG}) \quad (8)$$

The significance thus can be evaluated from the correlation r_k and we consider a candidate event as SM-to-DMN propagation if the corresponding p-value less than 0.001.

1.2.3 Spiking Cascade

Processed neural spike data sorted with Kilosort2 pipeline [43] from the Allen Institute Visual Coding Mice Neuropixel dataset were utilized in this analysis. In line with prior studies [13, 15], we first computed neural spike rates using 200ms time bins and identified candidate infra-slow neural events by segmenting the spiking rate data based on the troughs of the filtered global mean spike rate (low-pass, 0.25Hz). We then, using the delay-profile decomposition method, extract delay profile for each candidate event and the principal delay profile representing the predominant sequential pattern within infra-slow brain activity. A candidate event k was considered as a valid cascade event when the principal delay (PD) profile $\mathbf{u}^* \in \mathbb{R}^N$ matched with its delay profile $\mathbf{d}_k \in \mathbb{R}^N$, as quantified by Pearson's correlation coefficient:

$$r_k = \text{PearsonCorr}(\mathbf{d}_k, \mathbf{u}^*) \quad (9)$$

A candidate event that has a significant r_k with $p < 0.001$ was considered a cascade event in our study.

Adhering to the analysis in previous studies [13, 15], we categorized neurons into two distinct groups based on their positioning within the PD profile: **positive-delay neurons** exhibiting significant PD values and **negative-delay neurons** characterized by significantly negative PD values ($p < 0.001$, two-sided one-sample t-test).

The slow cascade features a sharp increase in the spiking activity of positive-delay neurons in the middle, following the definition in [13], we thus identified and defined these time points as the local peak of the first-order temporal derivative of the mean spiking time course of the positive-delay neurons.

188 2 Supplementary Figures

189 2.1 Figure S1

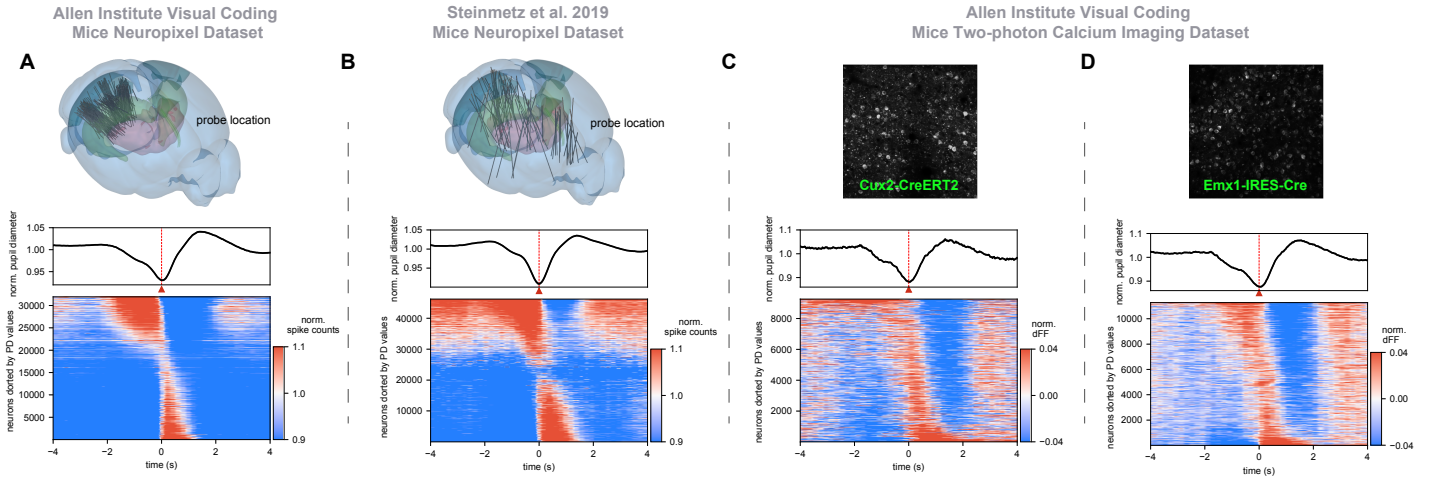


Figure S1. Pupil dilation is accompanied by stereotypical infra-slow, brain-wide sequential neural activation, evident across different cell types and extending through various brain regions in mice. (A) Analysis of Allen Institute Visual Coding Neuropixel dataset. Top row: visualization of 3-dimensional probes location in Allen Mouse Brain Common Coordinate Framework (version 3). Second row: The normalized pupil diameter (Top) and neuronal spiking activities (Bottom), both averaged around the onset of pupil dilation over an 8-second time window from all N mice, with neurons sorted by the principal delay profile identified during periods of pupil dilation. Similar analytical approaches are applied to another mice neuropixel dataset with probes recording covering the whole brain (B), and to two-photon calcium imaging datasets of two types of excitatory neurons in mice visual regions, i.e. Cux2-CreERT2 (C) and Emx1-IRES-Cre (D).

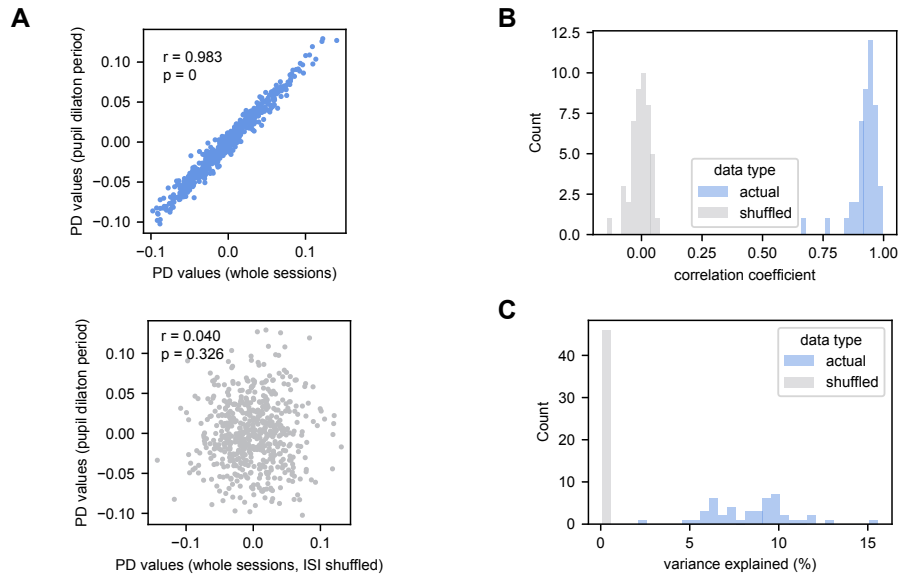


Figure S2. The sequential activation during pupil dilation is the principal infra-slow dynamics in mice brain. (A) High degree of similarity between the principal delay profile derived from periods of pupil dilation and that from whole recording sessions in an example mouse (Top). This correlation significantly diminishes when neural activities are randomized by inter-spike interval (ISI) permutation (Bottom). (B) The distribution of similarity, quantified by Pearson's correlation coefficient, between the principal delay profiles derived from pupil dilation periods and from whole recording sessions across all studied mice, showing high degree of similarity generalizable to all the mice. (C) Distribution of the variance explained by the principal delay profile derived from whole session data for neural activities during pupil dilation periods. This analysis validates the high degree of similarity between the infra-slow dynamics characterizing the entire recording session and those specifically associated with pupil dilation periods. Source data are provided as a Source Data file.

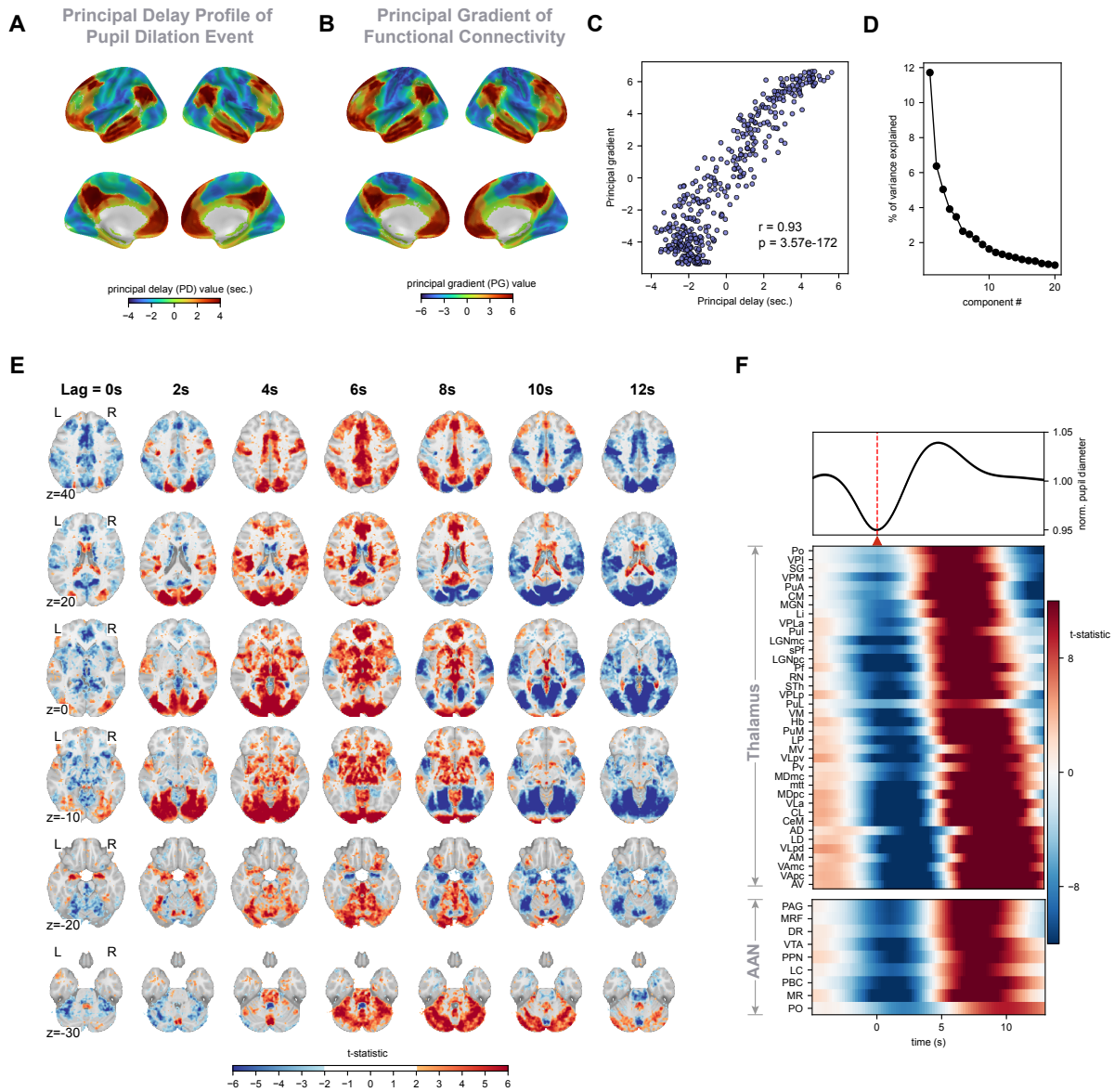


Figure S3. Resting-state pupil dilation is accompanied by hierarchical propagating waves in human brain. (A) The principal delay (PD) profile derived for the period of pupil dilation quantifying the temporal shift of each brain region in relation to the onset of pupil dilation. (B) The principal gradient (PG) of functional connectivity, characterizing the topographical organization of the cortex with regions at one end serving sensory/motor (SM) functions and the other end known as the default-mode network (DMN). (C) Strong similarity between the PD profile (shown in A) and PG (shown in B) and quantified using Pearson's correlation coefficient. (D) The scree plot showing the percentage of variance explained by the first 20 principal components derived during periods of pupil dilation. (E) The BOLD signal statistical map shown for 7 evenly spaced time lags, from 0 to 12 seconds following the onset of pupil dilation, with each row showing a distinct brain slice. (F) The temporal dynamics of pupil size (Top) and subcortical regions within the temporal window surrounding pupil dilation onset. The thalamic regions of interest (ROIs) are defined according to the Morel's atlas (Middle) and 9 brainstem ROIs are defined by the Harvard AAS atlas (Bottom). The full name of the abbreviation can be found in Table S1 and Table S2. Source data are provided as a Source Data file.

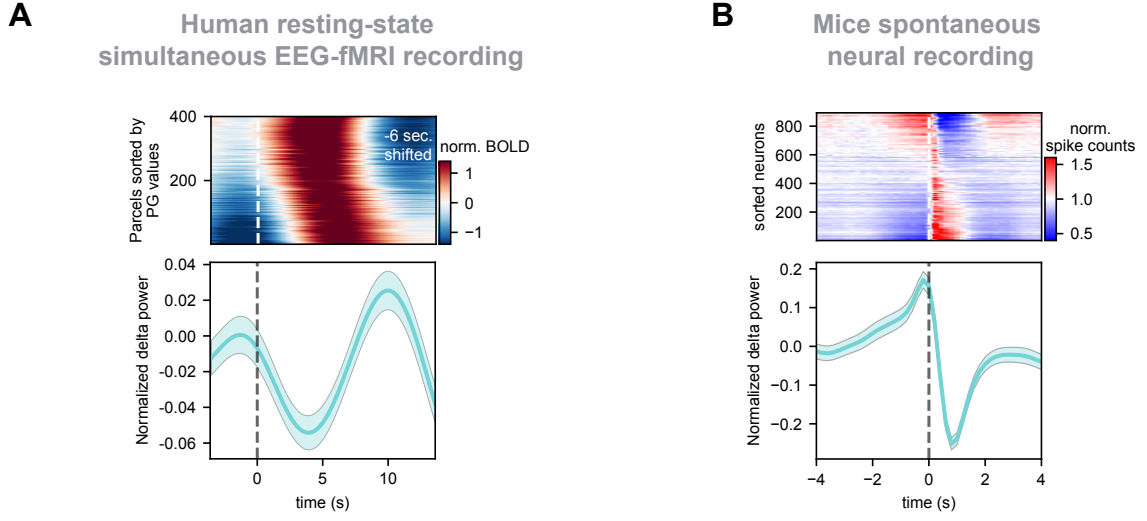


Figure S4. Association between brain-wide sequential activity and arousal modulation in humans and mice. (A) Cortical fMRI ROI signals (Top) and EEG delta-band power (Bottom) are averaged according to the detected SM-DMN propagation center across all resting-state simultaneous EEG-fMRI recording sessions. The averaged propagation pattern is adjusted by a 6-second left shift to account for hemodynamic delay, with a white dashed line marking time zero to indicate the onset of an increasing global signal. (B) Spiking activity (Top) and delta-band power (Bottom) are averaged according to the time of positive-delay neuron onset within the detected spiking cascade across all mice from Allen Institute Visual Coding Neuropixel dataset. The neuronal spiking data are organized in the form of spiking cascade by sorting the neurons according to the principal delay profile. Together, these findings highlight the unified coordination between stereotypical brain-wide neural activity and arousal modulation across species. Source data are provided as a Source Data file.

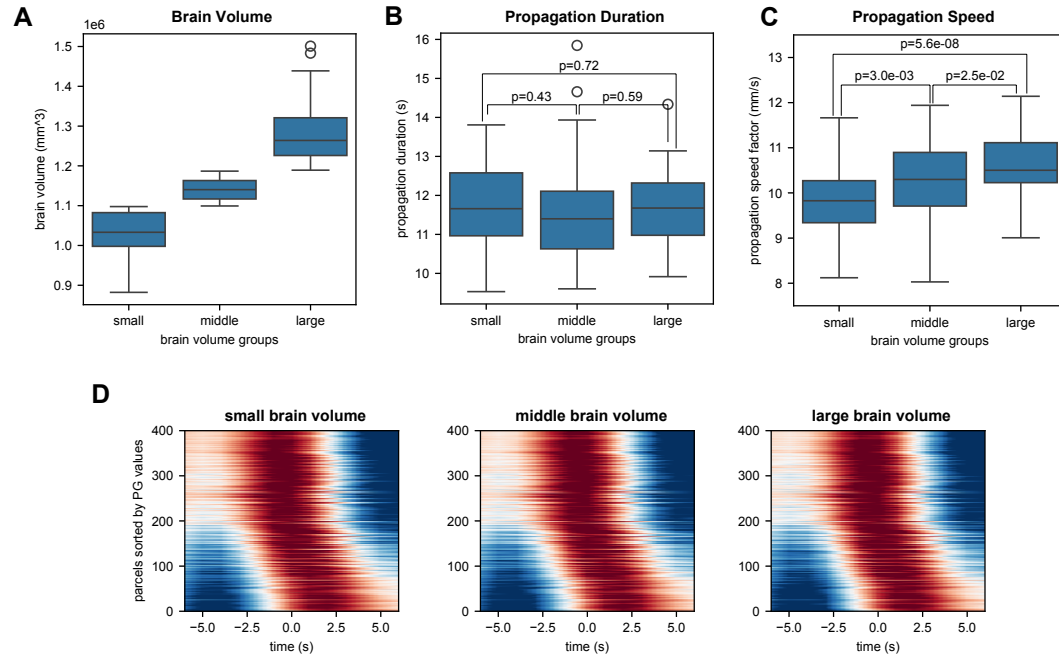


Figure S5. Relation between propagating wave duration/speed and brain volume. (A) 183 participants from HCP-7T dataset were categorized into three groups based on brain volume: small (lower 1/3), medium (middle 1/3), and large (upper 1/3). One subject was excluded due to insufficient number of propagating wave events detected. The box plot shows brain volumes across these groups. Box plots showing propagation durations (B) and propagation speed (C) across the different brain volume groups. Two-sided two-sample t-test was used to assess differences between groups ($N_1 = 61$, $N_2=60$, $N_3=62$). Box plot show the interquartile range (boxes), minima and maxima (whiskers), and the median (center line). (D) Average pattern of detected SM-to-DMN waves across the brain volume groups. Source data are provided as a Source Data file.

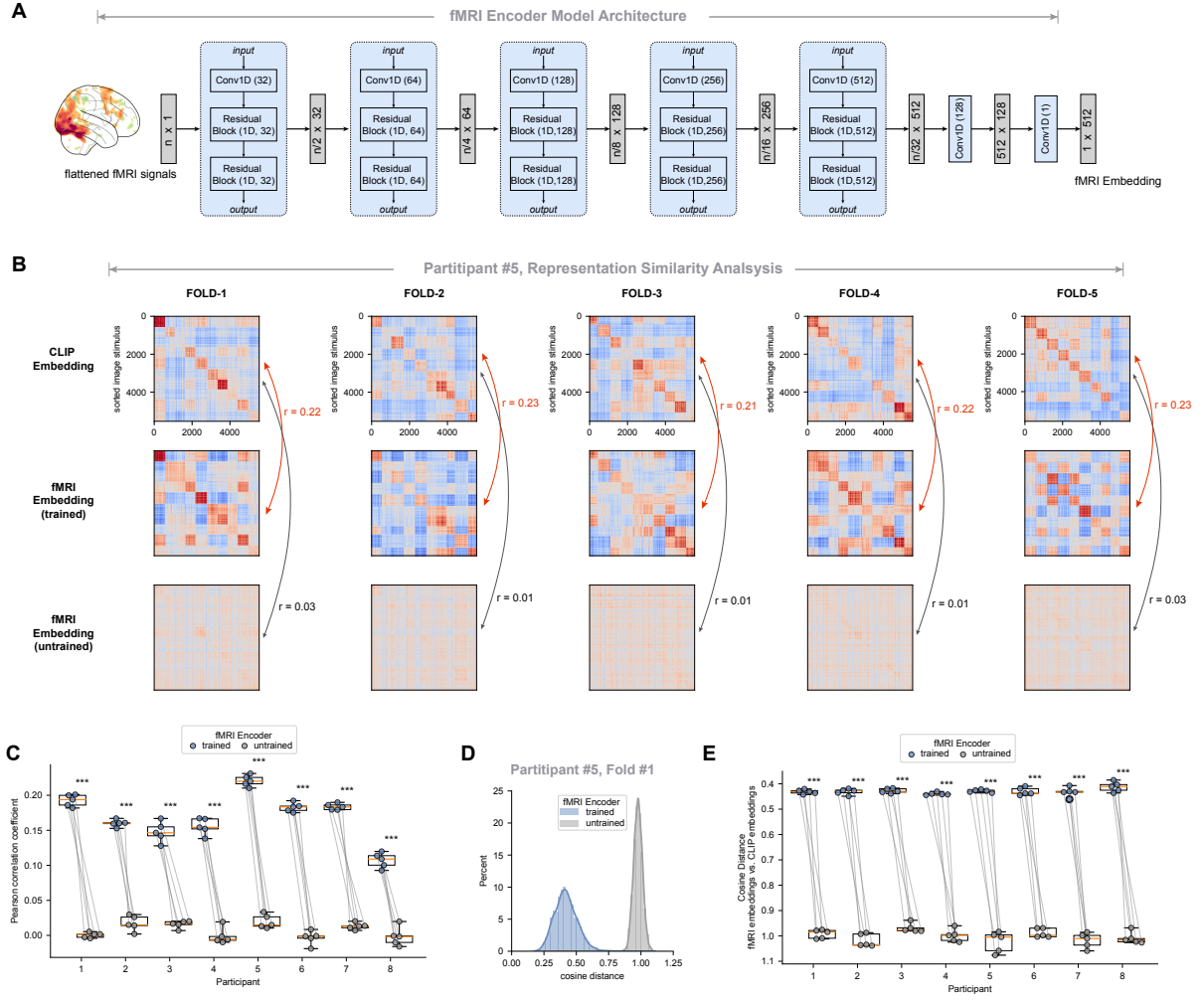
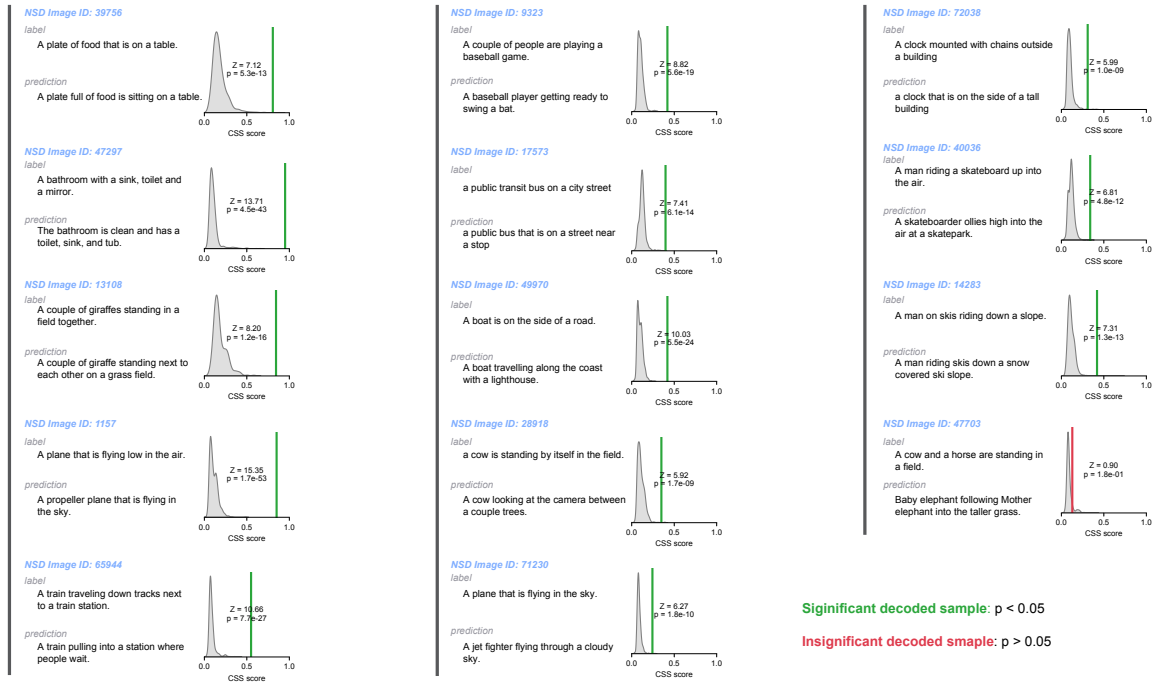
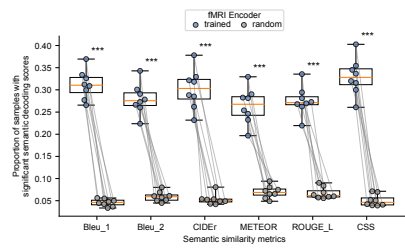


Figure S6. Bridging fMRI and CLIP embedding spaces through a deep learning-based fMRI encoder. (A) The model architecture of fMRI encoder, containing several convolutional layers to transform the fMRI BOLD response to 512-dimensional fMRI embedding vector, i.e., the latent representation of fMRI signal. (B) Representation similarity analysis for an example participant, with fMRI encoders are trained and evaluated with 5-fold cross-validation. For each validation fold (a column in the panel), representation dissimilarity matrices (RDMs) are constructed for semantic CLIP embeddings (Top), fMRI embeddings extracted with trained fMRI encoder (Middle), and fMRI embeddings extracted with untrained fMRI encoder (Bottom). For the RDMs within each fold, image stimuli are sorted identically according to CLIP-derived RDMs. Notably, RDMs based on fMRI embeddings show a significant improvement in similarity to CLIP-derived RDMs when extracted by the trained encoder (indicated by red arrows), compared to the untrained encoder (black arrows). The similarities are all significant with $p=0$. (C) Boxplots comparing the correlation differences between CLIP RDMs and RDMs from trained versus untrained fMRI embeddings across subjects. Each dot represents a validation fold from an individual subject, with pair-wise t-test (two-sided, $N=5$) used for statistical significance test. (D) Distribution of cosine distance, measuring the discrepancy between CLIP embeddings and fMRI embeddings, summarized separately for trained and untrained fMRI encoders from an example validation fold. (E) Boxplots showing the significant decrease in cosine distance (i.e., discrepancy) between CLIP and fMRI embeddings following the training of the fMRI encoder. Box plots show the interquartile range (boxes), minima and maxima (whiskers), and the median (center line). Source data are provided as a Source Data file.

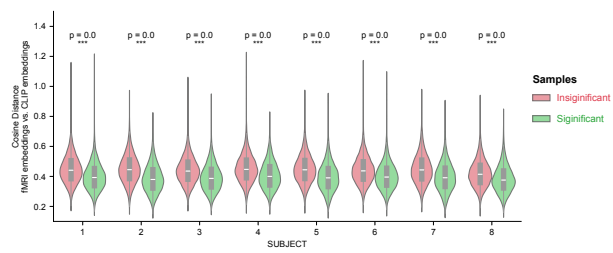
A



B



C

**Figure S7. Decoding semantic information from stimuli-evoked BOLD response.**

(A) Examples of correct and incorrect decoded trials. Each trial includes the model-generated caption and the human-labeled ground truth caption, with their similarity quantified using composite semantic similarity (CSS) scores. To assess the statistical significance of the decoding, for each trial, the actual CSS score is compared against a null distribution, constructed based on CSS scores between the ground truth caption and 1,000 randomly generated captions by sampling the CLIP embedding space. Notably, the null distribution varies across trials, reflecting differences in decoding complexity tied to the semantic content of the stimuli. Trials are considered correctly decoded if their p-value is below 0.05, indicated by green.

(B) Box plot comparing encoding accuracy—quantified as the proportion of correctly decoded trials—between the trained and untrained (control) fMRI encoder. Results are presented for various semantic similarity metrics, with each dot representing an individual participant. Statistical significance is evaluated using a pairwise t-test (two-sided, $N=8$). Box plot show the interquartile range (boxes), minima and maxima (whiskers), and the median (center line).

(C) Violin plot showing the distribution of the cosine distance between the fMRI embeddings and CLIP embeddings, comparing between the incorrect (red) and correct (green) decoded trials, summarized for each of the participants. Aligning with expectations, trials that are correctly decoded exhibit a closer alignment (reduced cosine distance) between the fMRI embeddings and the CLIP embeddings. Source data are provided as a Source Data file.

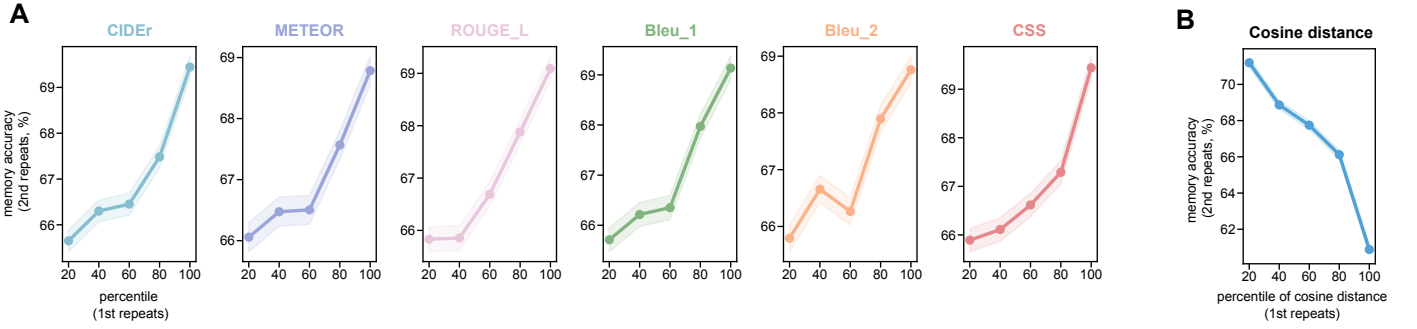


Figure S8. Semantic encoding accuracy of first-time image presentation predicts subsequent memory recall. (A) Influence of first-time encoding accuracy on subsequent memory task performance. Stimuli are categorized based on the percentile of their initial encoding accuracy, with increments of 20%, and the average memory recall accuracy during their second presentation is calculated for each bin. Each column shows the results for specific metrics to evaluate the semantic similarity score, from left to right showing CIDEr, METEOR, ROUGE_L, Bleu_1, Bleu_2, and CSS scores. (B) Parallel analysis to (A) but with a focus on the effect of the initial alignment between fMRI and CLIP embeddings (measured by cosine distance) on subsequent memory performance. A closer initial alignment (reduced cosine distance), indicative of better sensory encoding within the brain signal, is shown to enhance memory recall. Source data are provided as a Source Data file.

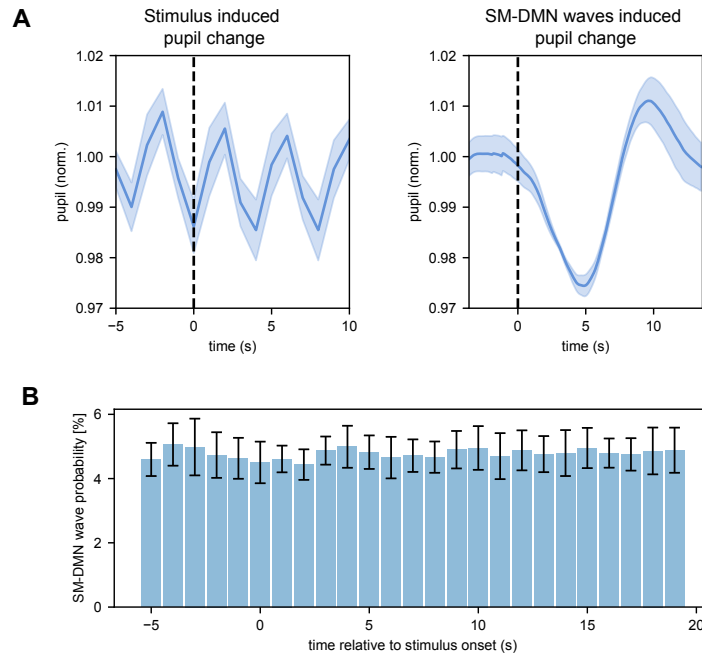


Figure S9. Influence of external stimulation to pupil dynamics and SM-to-DMN wave occurrences. (A) Left: The averaged pupil fluctuation across the onsets of stimuli with time zero defined as the timing of stimulus onset. Right: The averaged pupil size across detected SM-to-DMN waves with time zero defined as the onset of global mean signal increase. The shaded area indicates the standard error of the mean (SEM). (B) The histogram showing the probability of occurrence for SM-to-DMN waves, with time zero set at the onset of the stimulus. To avoid overlapping between segments for averaging, stimuli were selected with a minimum interval of 40 seconds between each. We also excluded the stimuli that within first 40 seconds and last 40 seconds of fMRI scans to mitigate edge effects. Error bars represent the standard deviation.

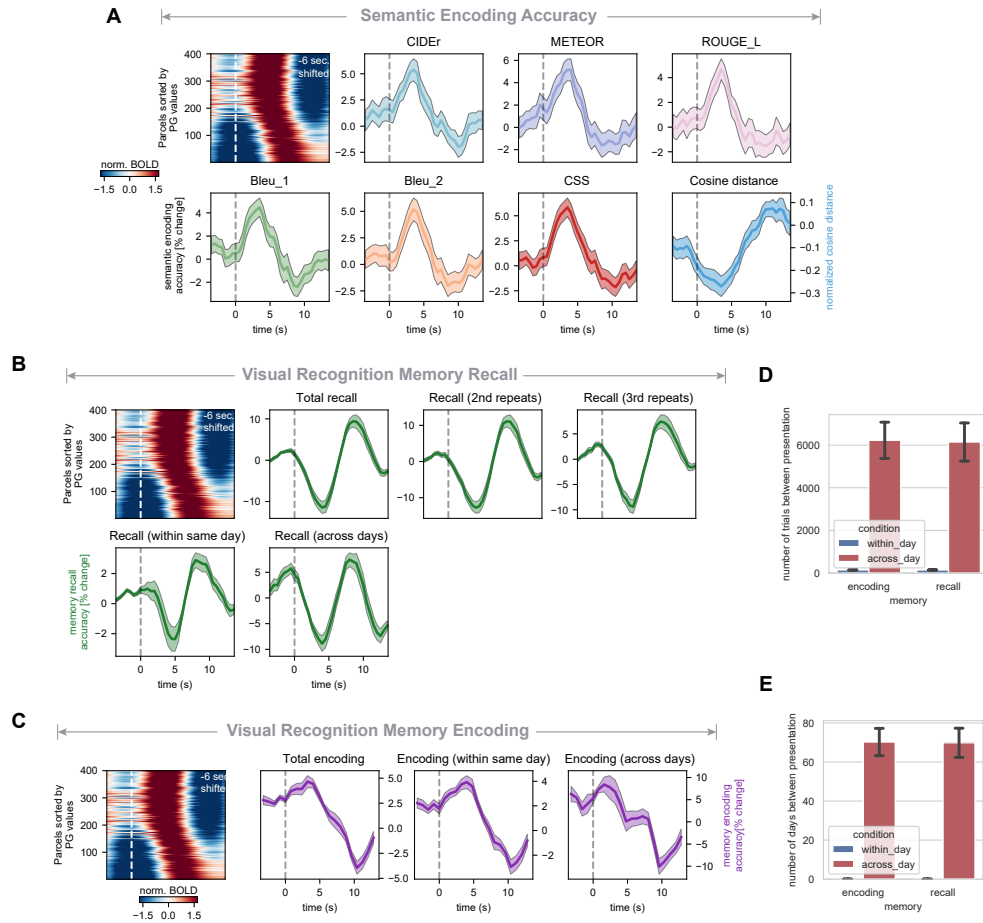


Figure S10. Sensory encoding and memory recall are oppositely modulated across the cycle of the SM-to-DMN propagating wave. (A) Phase-dependent modulation of sensory processing efficacy, quantified through semantic encoding accuracy using various semantic similarity metrics. The cosine distance measures the alignment between fMRI and CLIP embeddings, with closer alignment (reduced cosine distance) indicating better sensory encoding quality within the brain signal. The sensory processing shows enhanced encoding efficacy during the early phase of propagation cycle. (B) Phase-dependent modulation of memory recall. Various recall metrics are incorporated, including total recall (the combined recall from second and third image presentations), recalls specific to the second and third presentations, short-term recall (recalls derived for the scenario where first and second image presented within the same day), and long-term recall (recall across different days). Consistently, memory recalls show significant enhancement during the latter phase of the propagation cycle. (C) Phase-dependent modulation of memory encoding. Various encoding metrics are considered, including total encoding accuracy (across all first and second presentation pairs), short-term encoding (all encoding within the same day), and encoding across different days. Memory encoding are significantly improved during the early phase of the propagation cycle. The difference in the range of accuracy changes between within-day and across-day conditions likely reflects trial difficulty: within-day trials are easier to remember, showing smaller changes, while across-day trials are more challenging, resulting encoding effects more pronounced and sensitive. The averaged pattern of the detected SM-DMN propagation left shifted by 6 seconds, with time zero marked by white dashed line indicating the onset of increasing global signal. (D) Bar plot comparing the number of intervening trials between repeated presentations of the same stimulus for within-day (short-term) and across-day (long-term) conditions. (E) Bar plot showing the corresponding number of intervening days between stimulus repetitions. Source data are provided as a Source Data file.

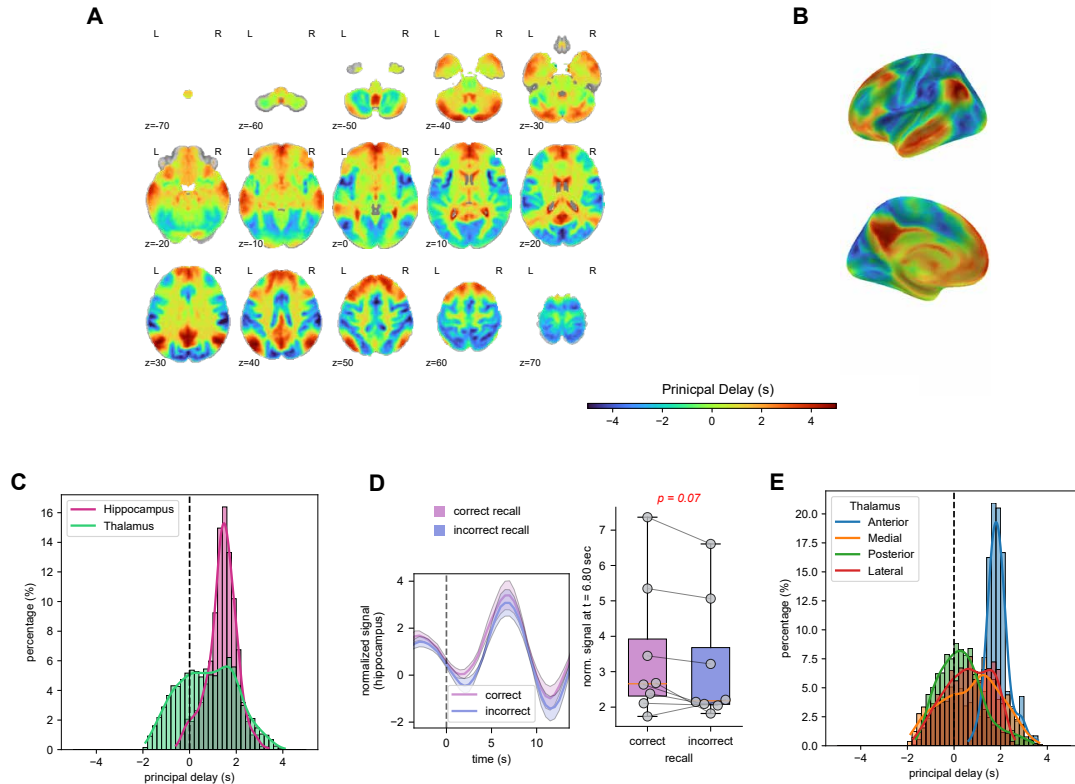


Figure S11. Principal delays and their distribution in the hippocampus and thalamus. (A) The principal delay (PD) profile quantifies the temporal relationships between brain regions during fMRI SM-to-DMN propagating waves. SM regions show early activations (negative PD value, blue), while DMN regions are activated in the late phase of the wave (positive PD value, red). (B) Principal delay profile mapped onto the brain surface. (C) Comparison of the principal delays distributions between the hippocampus and thalamus, showing hippocampal activation occurring in the later phase of the wave (positive PD value). (D) Association between hippocampal activity and memory recall performance. Correct recall is associated with higher hippocampal activity. Left: Time courses of hippocampal activity for correct (purple) and incorrect (blue) recall conditions. Shaded regions denote the standard error of the mean (SEM). Right: Boxplot comparing peak hippocampal activity between correct and incorrect recall conditions, with a trend toward significance ($p = 0.07$). (E) Distribution of principal delays within the thalamus, where anterior regions (associated with memory functions, blue) exhibit preference for the late phase, and posterior regions (associated with sensory processing, green) preferentially activate in the earlier phase. This findings aligns with the alternating modulation of sensory processing and memory recall during the SM-to-DMN waves. Source data are provided as a Source Data file.

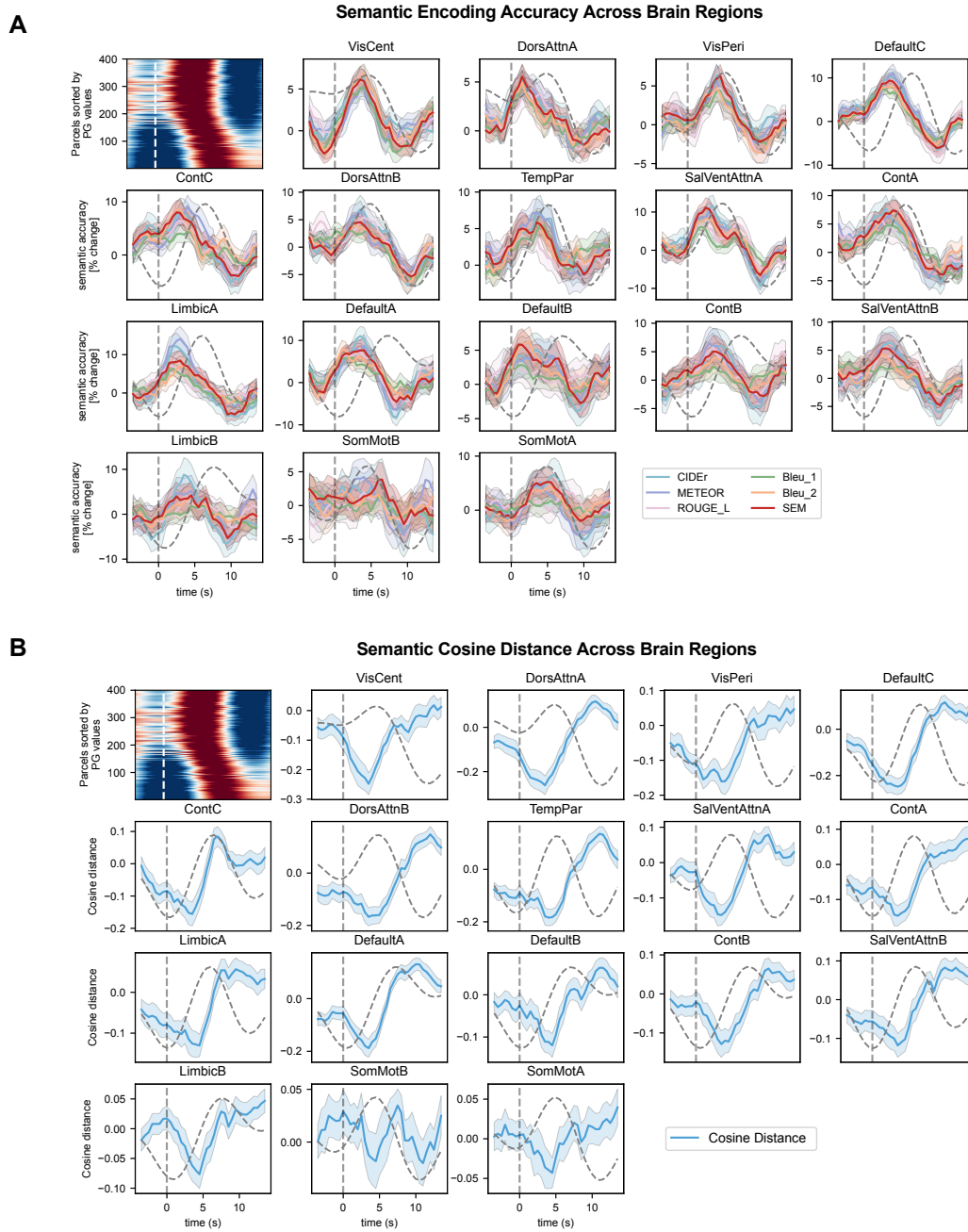


Figure S12. Regional dynamics of sensory encoding efficacy during SM-to-DMN propagation. (A) Semantic encoding accuracy evaluated with various semantic similarity metrics (CIDEr, METEOR, ROUGE_L, Bleu_1, Bleu_2, and CSS). (B) Modulation of cosine distance, which measures the alignment between fMRI and CLIP embeddings (reduced cosine distance indicating better sensory encoding quality), consistently aligns with the modulation of encoding accuracy in (A). For both panels, the regions are organized in descending order according to the significance of their semantic encoding accuracy, revealing a gradient in the modulation effect. The averaged propagation pattern is adjusted by a 6-second left shift to account for the hemodynamic delay, with a white dashed line marking time zero to indicate the onset of an increasing global signal. For comparison, the average activation for each region is marked with gray dotted line, which is shifted ahead of time by 6 seconds to account for the hemodynamic response delay. Source data are provided as a Source Data file.

201 3 Supplementary Tables

202 3.1 Table S1

Table S1. List of thalamic nuclei and their abbreviation as in the Morel atlas [96].

Abbreviation	Full Name
AD	anterior dorsal nucleus
AM	anterior medial nucleus
AV	anterior ventral nucleus
LD	lateral dorsal nucleus
VAmc	ventral anterior nucleus magnocellular part
VApC	ventral anterior nucleus parvocellular part
VLa	ventral lateral anterior nucleus
VLpd	ventral lateral posterior nucleus dorsal part
VLpv	ventral lateral posterior nucleus ventral part
VM	ventral medial nucleus
VPI	ventral posterior inferior nucleus
VPLa	ventral posterior lateral nucleus anterior part
VPLp	ventral posterior lateral nucleus posterior part
VPM	ventral posterior medial nucleus
LGNmc	Lateral geniculate nucleus magnocellular part
LGNpc	Lateral geniculate nucleus parvocellular part
MGN	Medial geniculate nucleus
Po	posterior nucleus
SG	suprageniculate nucleus
LP	lateral posterior nucleus
PuA	anterior pulvinar nucleus
PuI	inferior pulvinar nucleus
PuL	lateral pulvinar nucleus
PuM	medial pulvinar nucleus
Hb	habenular nucleus
CeM	central median nucleus
MV	medioventral nucleus
Pv	paraventricular nucleus
CL	central lateral nucleus
CM	centre median nucleus
Pf	parafascicular nucleus
sPf	subparafascicular nucleus
MDmc	mediodorsal nucleus magnocellular part
MDpc	mediodorsal nucleus parvocellular part

203 **3.2 Table S2**

Table S2. List of brain regions and their abbreviation as in the AAN atlas [97].

Abbreviation	Full Name
DR	dorsal raphe
LC	locus coeruleus
MRF	midbrain reticular formation
MR	median raphe
PAG	periaqueductal gray
PBC	parabrachial complex
PO	pontis oralis
PPN	pedunculopontine nucleus
VTA	ventral tegmental area

204 **3.3 Table S3**

Table S3. The significance of encoding accuracy across different brain regions

Brain region	p value	t score	N	Significance
VisCent	1.5e-05	10.54	8	***
VisPeri	3.9e-05	9.14	8	***
DorsAttnA	2.1e-05	10.05	8	***
DorsAttnB	4.3e-04	6.23	8	***
ContA	1.5e-03	5.05	8	**
ContB	7.5e-03	3.72	8	**
ContC	4.1e-04	6.28	8	***
SomMotA	5.0e-02	2.37	8	*
SomMotB	6.9e-03	3.78	8	**
TempPar	3.8e-04	6.37	8	***
DefaultA	7.7e-03	3.69	8	**
DefaultB	6.2e-03	3.87	8	**
DefaultC	2.6e-04	6.77	8	***
SalVentAttnA	2.6e-04	6.77	8	***
SalVentAttnB	3.0e-02	2.71	8	*
LimbicA	2.0e-03	4.79	8	**
LimbicB	5.6e-01	0.61	8	ns