nature communications

Article

https://doi.org/10.1038/s41467-026-69041-8

Geometry of neural dynamics along the cortical attractor landscape reflects changes in attention

Received: 26 August 2025

Accepted: 23 January 2026

Published online: 12 February 2026

Check for updates

Hayoung Song [1], Ruiqi Chen [2], Thomas L. Botch [3], Todd S. Braver [4], Monica D. Rosenberg [5,6,7], Jeffrey M. Zacks [4] & ShiNung Ching [8]

Large-scale brain activity reflects changes in attention. To understand how, we tested a hypothesis that the geometry of neural dynamics on the cortical attractor landscape, or movement along its "hills and valleys", reflects attentional states. A dynamical systems model separating intrinsic dynamics from stimulus-driven influences was fit to fMRI data collected during rest, tasks, and movie-watching. Model simulations revealed a set of attractors aligned with canonical functional brain networks. The speed and direction of neural trajectories toward these attractors varied systematically with attentional states over time and across contexts. When participants were paying attention to effortful tasks, neural dynamics converged fast and directly toward a task-relevant attractor. In contrast, when participants were engaged in sitcom episodes, neural dynamics occupied a flatter region of the landscape, directed away from attractors. These findings demonstrate that while attractor locations are largely determined by cortical organization, the geometry of neural dynamics changes systematically across attentional states and contexts.

The brain is a multiscale system that dynamically evolves over time to produce behavior. Dynamical systems modeling, which formalizes how neural activity changes over time using differential equations, has played a central role in neuroscience[1,2]. Such models have uncovered numerous biophysical mechanisms of the brain, from the early work of Hodgkin and Huxley[3] who described how ion currents generate action potentials, to Wilson and Cowan[4] who described how excitatory-inhibitory neuronal interactions shape population-level dynamics.

More recent work has extended this approach beyond single cells and neuronal populations to characterize whole-brain dynamics[5–10]. From a dynamical systems perspective, large-scale brain activity unfolds over time as a trajectory within a high-dimensional state space, where each dimension typically represents the activity of a brain region. This state space is shaped by an attractor landscape, or hills and valleys that guide the trajectory of neural activity (Fig. 1). Research has shown that trajectories visit a small number of recurring brain states, each defined by a unique pattern of regional activity and interactions[11–15]. These brain states function as attractors, or deep valleys in the landscape, where neural activity is most likely to converge[16–19]. In turn, neural dynamics are characterized by transitions between or perturbations around these attractors.

While attractor landscapes govern neural dynamics, it is important to understand how they relate to our internal states, including physiological (e.g., hunger, stress, arousal) and cognitive states (e.g., attention, emotion, motivation)[20]. Prior studies have provided initial evidence of this connection, showing that certain brain states are more or less likely to occur when a person is focused versus unfocused[21], performing better or worse on a task[22,23], or engaged versus

[1]Center for Theoretical and Computational Neuroscience, Washington University in St. Louis, St. Louis, MO, USA. [2]Division of Biology and Biomedical Sciences, Washington University in St. Louis, St. Louis, MO, USA. [3]Department of Psychological and Brain Sciences, Dartmouth College, Hanover, NH, USA. [4]Department of Psychological and Brain Sciences, Washington University in St. Louis, St. Louis, MO, USA. [5]Department of Psychology, University of Chicago, Chicago, IL, USA. [6]Neuroscience Institute, University of Chicago, Chicago, IL, USA. [7]Institute for Mind and Biology, University of Chicago, Chicago, IL, USA. [8]Department of Electrical and Systems Engineering, Washington University in St. Louis, St. Louis, MO, USA. e-mail: omasong17@gmail.com

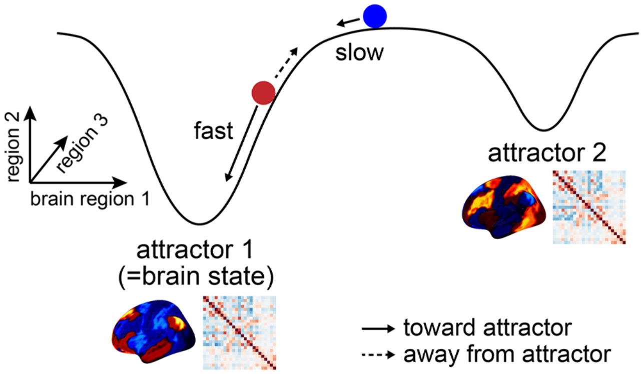

Fig. 1 | Schematics of the geometry of neural dynamics on the attractor landscape. A state space is defined where each dimension represents the activity of a brain region spanning the cortex. The hills and valleys represent the attractor landscape with valleys indicating the attractors. Each attractor corresponds to a recurring brain state that is identified from large-scale patterns of regional activity and interaction. The circles represent the neural activity at a specific moment. The trajectory of neural activity (indicated with black arrow) is largely determined by the landscape but can also be affected by external perturbations. For example, the red circle is more likely to fall toward the attractor based on the intrinsic landscape but may move away from the attractor when perturbed by external forces, such as stimuli, task demands, or behavior. The speed and direction of the movement on this landscape defines the geometry of neural dynamics. Example brain state figures are adapted from Song et al.[24].

disengaged in a movie[24]. However, the computational mechanism of this connection remains unknown[25]. One way to characterize such mechanisms is through the geometry of neural dynamics—the structure and flow of trajectories that describe how brain activity evolves over time along the attractor landscape (Fig. 1). This geometry can be quantified by the speed and direction with which brain activity moves toward or away from attractors. We hypothesized that the geometry of neural dynamics changes over time in relation to internal state fluctuations. Supporting this idea, Munn et al.[26] showed evidence of flattened attractor landscape during phasic bursts in noradrenergic locus coeruleus activity (related to arousal)[27] and steepened attractor landscape during phasic bursts in cholinergic basal forebrain activity (related to vigilance or attentional focus)[28]. This suggests a possibility that one's internal state at a given moment may correspond to whether the brain occupies a flatter or steeper region of the attractor landscape, which in turn shapes the trajectory of neural dynamics.

Here, we propose that the geometry of neural dynamics on the attractor landscape characterizes moment-to-moment and context-to-context variations in internal states. In this study, we specifically test this in relation to measures of sustained attention. Dynamical systems models were fit to whole-brain fMRI data collected during rest, tasks, and movie-watching. Our model separates neural activity into two components: one that is intrinsic, driven internally by regional activity and interactions, and the other extrinsic, driven externally from the stimulus. By simulating neural trajectories from the model, we identify attractors toward which the neural activity converges. At each moment, we estimate the speed and direction of these simulated trajectories, using them to infer the steepness of the local attractor landscape. Importantly, we relate the change in geometries to behavioral measures of attention which were collected with the fMRI data. These measures include participants' continuous ratings of engagement while watching comedy sitcoms and button response times during controlled sustained attention tasks.

Prior work has typically fit a set of static model parameters from neural activity dynamics, meaning variations across time have often been reduced to a single model that is agnostic to temporal change. Here, we reproduce evolving neural geometry from a model of fixed parameters and relate these dynamics to fluctuations in internal states. With this approach, we test a hypothesis that the geometry of large-scale cortical dynamics along the attractor landscape systematically reflects changes in attentional states during task and movie-watching contexts.

A dynamical systems model of large-scale cortical activity

We applied a large-scale parametric dynamical systems model, developed and validated by Singh et al.[9] and Chen et al.[19], to fit the time series of BOLD activity measured in human cortex with fMRI. The model defines the rate of change in neural activity $\dot{x}_t$, as a function of the neural activity x_t ($x_t \in \mathbb{R}^n$, $n =$ number of neural units) and time-aligned experimental variables u_t ($u_t \in \mathbb{R}^p$, $p =$ number of experimental variables), such as task, stimulus, or behavior. In our model, X corresponds to the BOLD activity time series of 200 parcels covering the cortex[29]. U corresponds to audiovisual and semantic features extracted from the stimuli that participants watched and heard inside the scanner, which were reduced to 100 principal component dimensions. Mathematically, the model can be described as $\dot{x}_t = F(x_t) + G(u_t)$, where F and G are deterministic functions that represent, respectively, the impact of intrinsic activity and input-driven perturbations on the evolution of neural activity.

The followings are the specifics of our model, where $F(x_t) = W\psi_\alpha(x_t) - D \odot x_t$ and $G(u_t) = \beta u_t$ with W, D, α, and β being model parameters (Fig. 2).

$$\dot{x}_t \approx \frac{x(t+\Delta t)-x(t)}{\Delta t} = W\psi_\alpha(x_t) - D \odot x_t + \beta u_t \tag{1}$$

$$\psi_\alpha(x_t) = \sqrt{\alpha^2 + (bx_t + 0.5)^2} - \sqrt{\alpha^2 + (bx_t - 0.5)^2} \tag{2}$$

Given that Δt equals to 1 TR in our data, the equation can be simplified as follows.

$$\hat{x}_{t+1} = x_t + W\psi_\alpha(x_t) - D \odot x_t + \beta u_t \tag{3}$$

The goal of the model is to predict neural activity pattern of the consecutive time step ($\hat{x}_{t+1}$), by minimizing the prediction error (i.e., difference between the predicted $\hat{x}_{t+1}$ and observed x_{t+1}) using algorithmic optimization (here, stochastic gradient descent). The prediction of the next time step is based on the neural activity x_t plus the signals received by connections from other parcels ($W\psi_\alpha(x_t)$) minus the self-decay ($D \odot x_t$) plus the neural activity driven by the external inputs (βu_t). In turn, fitting this model corresponds to decomposing intrinsic and extrinsic (i.e., input-driven) neural dynamics.

The weight matrix ($W \in \mathbb{R}^{n \times n}$) represents directional interaction between neural units, namely the effective connectivity. Nonlinearity is introduced by a parametrized sigmoidal transfer function (ψ_a), which maps neural activity to a bounded output at a range from −1 to 1. The slope of the transfer function differs for every parcel, parameterized by α ($\alpha \in \mathbb{R}^n$). b is fixed at 20/3 following prior studies[9,19]. A self-decay ($D \in \mathbb{R}^n$) captures a return to baseline in absence of external inputs or interactions, with $\odot$ denoting element-wise product. Having high decay rate corresponds to having low temporal autocorrelation, indicating less persistence in neural activity. β represents linear transformation from the stimulus embedding space to neural activity pattern space. Note that this is an individualized model where a set of parameters are estimated for each fMRI run of each participant.

We analyzed two openly available fMRI datasets, the SONG dataset[24] ($N = 27$) and the HCP dataset[30–32] ($N = 119$). In both datasets, each participant underwent multiple sessions of rest, task, and movie-watching conditions. We extracted time series of stimulus features for

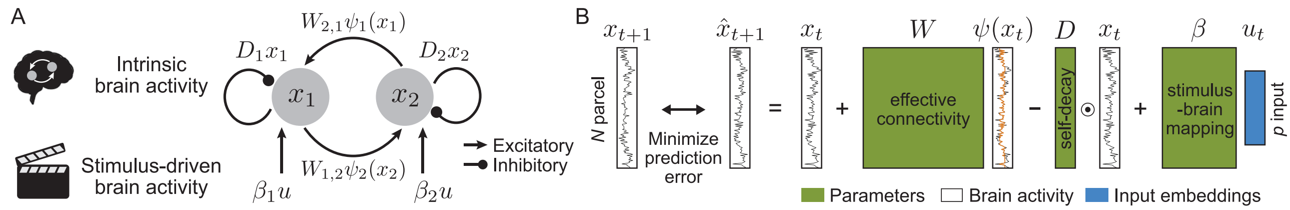

Fig. 2 | Model of large-scale cortical dynamics. A Model schematics. x represents the activity time series of a cortical parcel and u represents input from the stimulus. Model parameters include directional interactions between neural units (W), self-decay that determines autocorrelation (D), and the stimulus-to-brain relationship (β). Although only two units are visualized for simplicity, the model was fit on the time series of 200 cortical parcels. **B** Model optimization. The model was trained to minimize the difference between the observed and predicted neural activity patterns of consecutive time steps. Green denotes parameters that are estimated during training, black lines denote observed neural activity pattern at a time step (x_t), with orange indicating sigmoidal bound from −1 to 1 given the nonlinear transfer function, and blue denotes stimulus embeddings at the corresponding time step (u_t).

the movie-watching runs, including low-level visual and visuo-semantic features of video frames, low-level audio of the sounds, and audio-semantic features of the speech and dialogue in the movies. These features were projected onto 100 principal components that explained the largest variance across runs. Only low-level visual and visuo-semantic features were extracted from the SONG dataset task runs because the stimuli were purely visual. Given that no stimulus was provided for resting-state runs in either dataset, we fit a simplified model that only considers intrinsic but not extrinsic neural dynamics: $\dot{x}_t = W\psi_\alpha(x_t) - D \odot x_t$ (removing the input term βu_t from the equation).

Model parameters recapitulate functional brain connectivity and stimulus encoding

We first tested whether our model, optimized to predict neural activity of successive time steps, captured parameters sensitive to individual and cognitive state differences. These properties are critical, as their emergence indicates the model's biological plausibility. We analyzed the movie-watching runs of the SONG and HCP datasets because they were fit on the full model that includes the input term (see Supplementary Fig. S1 for the range of model parameter estimates).

Model performance was estimated based on how well the model predicted neural activity of the next time steps, specifically $r^2(\hat{x}_{2:T}, x_{2:T})$. The total explained variance was, on average, $r^2 = 0.452 \pm 0.043$ across a total of 80 runs in the SONG dataset and $r^2 = 0.561 \pm 0.053$ across 476 runs in the HCP dataset (Fig. 3A). This indicates that our model explained approximately half of the variance in neural activity. We then decomposed the explained variance into three components. The interareal interaction explained $r^2 = 0.041 \pm 0.009$ (SONG), 0.055 ± 0.015 (HCP), local recurrence explained $r^2 = 0.284 \pm 0.051$, 0.237 ± 0.060, and external input-driven activity explained $r^2 = 0.017 \pm 0.003$, 0.015 ± 0.004 of the variances.

We validated model performance across runs. Model parameters estimated from one run of a participant were better at predicting that same participant's brain activity in other runs than predicting other participants' brain activity (SONG: within: $r^2 = 0.418 \pm 0.037$, across: $r^2 = 0.402 \pm 0.011$; paired t-test $t(79) = 4.049$, $p = 0.0001$; HCP: within: $r^2 = 0.526 \pm 0.052$, across: $r^2 = 0.508 \pm 0.015$; $t(475) = 9.375$, $p = 2.8\text{e-}19$; Supplementary Fig. S2). This indicates that the model captured individual-specific patterns of brain dynamics and explained more variance within individuals than across individuals.

We further validated the model by comparing model parameters, W and β, to their analogue descriptive statistics that are commonly used in the field. W was compared to an undirected functional connectivity (FC), estimated as the parcel-by-parcel Fisher's transformed Pearson's correlation coefficients. β was compared to regression coefficients estimated from a linear encoding model, that predicts neural activity from stimulus time series: $x_t = (\text{encoding coefficient}) \times u_t + (\text{residual})$ (encoding coefficient $\in \mathbb{R}^{n\times p}$, residual $\in \mathbb{R}^{n}$). We found that the estimated W was highly comparable to FC for both the SONG (cosine similarity = 0.929 ± 0.009; $z = 1178.36$, $p < 0.0001$ compared to shuffled chance distribution) and HCP datasets (cosine similarity = 0.913 ± 0.014; $z = 2830.05$, $p < 0.0001$) (Fig. 3B, D). Likewise, β was highly comparable to encoding coefficients for both the SONG (cosine similarity = 0.324 ± 0.100; $z = 448.58$, $p < 0.0001$) and HCP datasets (cosine similarity = 0.243 ± 0.036; $z = 1022.18$, $p < 0.0001$) (Fig. 3C).

Do these parameters capture differences between individuals as well as differences in cognitive states? Models were considered sensitive to individual differences if the parameters estimated from runs of the same person were more similar compared to parameters of different individuals (Fig. 3E). Both W and β were sensitive to individual differences, and W more strongly reflected individual differences than β (Wilcoxon rank-sum test between same vs. different individual pairs; W: $z = 14.957$, $p = 1.4\text{e-}50$ for SONG, $z = 44.667$, $p = 0.0$ for HCP; β: $z = 2.714$, $p = 0.007$ for SONG, $z = 13.828$, $p = 1.7\text{e-}43$ for HCP) (Fig. 3F). This aligns with prior findings that functional connectivity is stable within an individual and distinctive across individuals, thus driven more by trait than state differences[33,34].

Models were considered sensitive to cognitive state differences if parameters estimated from runs of the same movie stimulus were more similar compared to runs of different stimuli (Fig. 3E). Both W and β were sensitive to cognitive state differences, and β more strongly reflected cognitive state differences than W (W: $z = 21.414$, $p = 9.9\text{e-}102$ for SONG, $z = 75.509$, $p = 0.0$ for HCP; β: $z = 45.189$, $p = 0.0$ for SONG, $z = 242.537$, $p = 0.0$ for HCP) (Fig. 3F). The corresponding descriptive statistics—the FC and encoding coefficients—closely followed this trend. These results indicate that the parameters estimated from our dynamical systems model are comparable to validated descriptive statistics and hold biological plausibility.

The model reveals stable attractors organized along the cortical hierarchy

In dynamical systems, the differential equation determines and predicts how the system's trajectory would evolve over time from a given initial state, assuming that no internal or external perturbation exists beyond what is parameterized. Depending on the nature of the system, the trajectory may eventually converge to a set of stable fixed point attractors, settle onto stable periodic orbits known as limit cycles, bounce between saddle points when attractors are absent, or exhibit chaotic behavior[19].

We hypothesized that our data-driven models would reveal a finite set of attractors, or stable patterns of neural activity toward which neural trajectories are likely to converge. To find attractors, we ran forward simulations of the model starting from multiple initial states, each corresponding to the observed 200-parcel neural activity pattern at every time step of the fMRI time series (Table 1). The observed neural activity at each time step was simulated 5000 time-steps forward in time, which we considered was sufficient for convergence. This method simulates the intrinsic drift, or the likely path that neural activity would follow in the absence of external

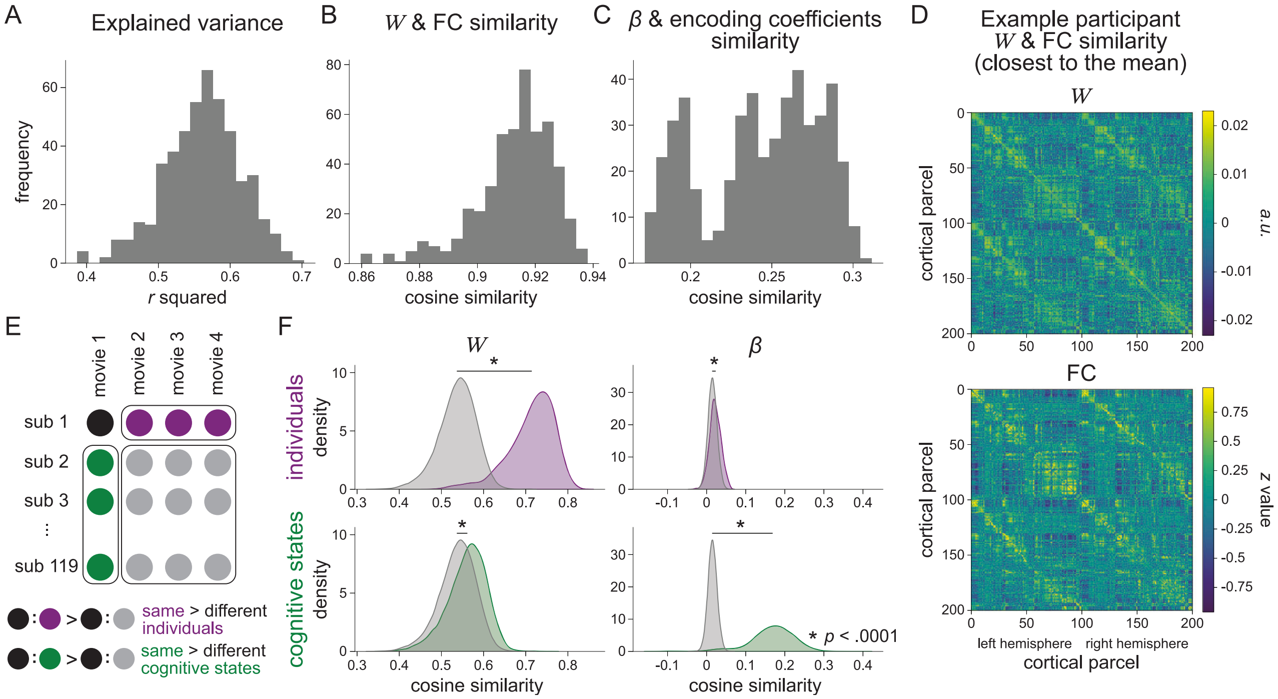

Fig. 3 | Model validation with the HCP dataset. A Model performance was primarily assessed based on the explained variance of how well the model predicted neural activity of the next time step given the current time step (the explicit training objective). The histogram includes r^2 estimates from all movie-watching runs of 119 participants included in the HCP dataset. **B**, **C** Model parameters were compared to descriptive statistics, specifically W to functional connectivity (FC) estimates and β to coefficients estimated from the stimulus-to-brain encoding models. These aspects of the model were not explicitly optimized. **D** Example participant's W estimate compared to their FC matrix. A representative participant's data was selected for visualization (whose cosine similarity between W and FC was closest to the mean in (**B**)). **E** Individual differences were assessed by comparing parameter similarities between different movie-watching runs of the same participant (black-purple) to those of different participants (black-grey). Cognitive state differences were assessed by comparing parameter similarities of different participants' same movie-watching runs (black-green) to those of different movie-watching runs (black-grey). **F** Density functions comparing similarities in parameter estimates between the same vs. different individuals (*top*) and the same vs. different cognitive states elicited by the same vs. different movies (*bottom*). Lines on top of the density functions connect the means of the two distributions, with asterisks indicating $p < 0.0001$. Supplementary Fig. S3 shows the same model validation results, analyzed with the SONG dataset.

Table 1 | Forward simulation

```
$X$ represents the fMRI activity time series matrix of size (200 parcels × time).
$U$ represents the input time series matrix of size (100 stimulus embedding PCs × time).
Fit a dynamical systems model to estimate parameters ($W, D, \alpha, \beta$) specific to the fMRI run.
For each time step, with $x_t$ as the initial condition and $u_t$ as the input at that corresponding moment, run a forward simulation (5000 iterations).
  for $t$ from 1 to the number of time steps:
    for 5000 iterations:
      if 1st iteration:
        $x_{t+1} = x_t + W\psi_\alpha(x_t) - D \odot x_t + \beta u_t$
      else:
        $x_{t+1} = x_t + W\psi_\alpha(x_t) - D \odot x_t$
      $x_t = x_{t+1}$
```

For every time step of the neural data, we simulated the model continuously to predict the likely path of neural dynamics, assuming it strictly follows the equation of the model. For each time step, βu_t was simulated only one time step in the future because stimulus-driven activity for future simulations cannot be predicted. In contrast, the intrinsic drift could be predicted through simulations, by using the predicted x_{t+1} as the next x_t and iterating this 5000 times. If the forward simulation reached a state where no further change occurred within the last ten steps of the iterations ($||\dot{x}||_\infty < 1e\text{-}6$), we considered it to have converged to a fixed point. Across all time steps, fixed points separated by less than 0.1 in Euclidean distance were grouped as the same attractor. The attractors identified this way represent states that the neural activity tend toward, assuming no further perturbation beyond what's driven by the input u_t.

perturbations beyond those parameterized by the model. If the forward simulation reached a state where no further change occurred over the last ten iteration steps, we considered it to have converged to a fixed point, representing the bottom of a valley in the inferred attractor landscape (Fig. 1). Fixed points estimated across all time steps were grouped as the same attractor if their pairwise Euclidean distance was less than 0.1 (Fig. 4A).

We focused this analysis on the HCP dataset, given that it contained a larger number of total runs (1666 runs) compared to the SONG dataset (188 runs). We included rest, task, and movie-watching runs in the analysis. When conducting forward simulations on all runs respectively, a majority of runs converged onto a set of point attractors: many converged onto 2 (41.06%) or 4 attractors (51.08%), and some converged onto 6 (7.02%) or 8 attractors (0.66%) (Fig. 4A). The stability of the attractors in all of these runs was confirmed by computing the Jacobian matrix of the model and verifying that all eigenvalues occupy the unit circle, indicating that small perturbations around an attractor would return the system to that attractor. Only 3

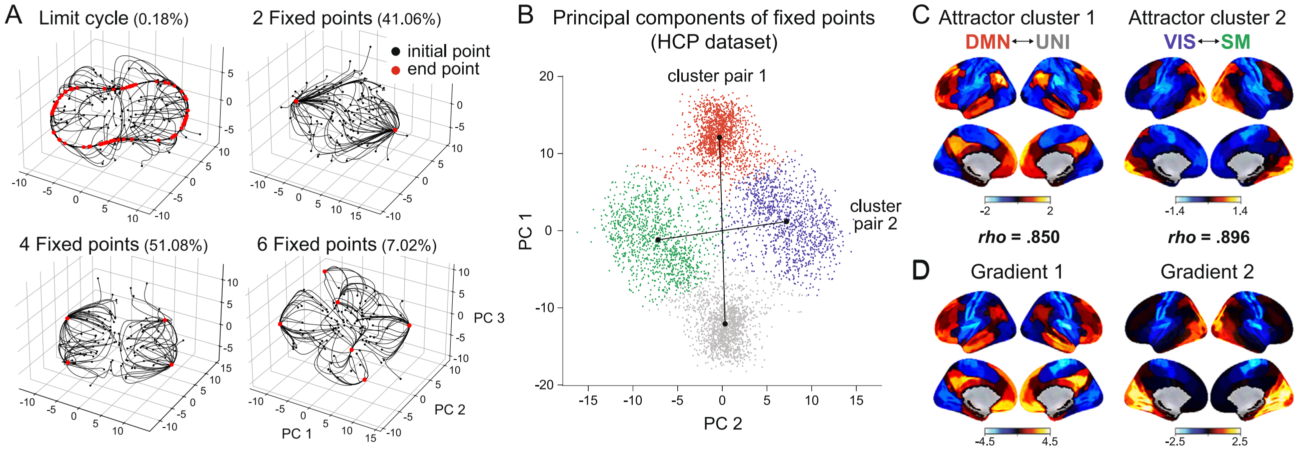

Fig. 4 | Attractor landscapes of large-scale cortical activity. A Four types of forward simulation results visualized from example fMRI runs. Black dots indicate 100 randomly sampled initial states (i.e., 100 time steps sampled from the observed neural activity). Each line indicates a trajectory taken over the course of a forward simulation (5000 iterations). Red dots indicate the end states of the simulations, corresponding to the attractors. Principal component analysis was conducted per run to visualize the results in a 3D principal component (PC) space. **B** Attractors identified from all runs of the HCP dataset, color-coded based on the outcome of k-means clustering into 4 clusters and projected onto the shared PC space for visualization. Dots correspond to attractors estimated from the 1666 HCP runs. Lines connect the pairs of cluster centroids that have anticorrelated patterns. **C** Neural patterns of the two identified attractor clusters. **D** Neural patterns of the top two gradients identified from Margulies et al.[36]. ρ values in between (**C** and **D**) indicate rank correlation values of the top and bottom neural patterns. Colors indicate loadings on the first two axes of the PC and gradient analyses. DMN default mode network, UNI unimodal network, VIS visual network, SM somatosensory-motor network.

out of 1666 runs (0.18%) did not converge to fixed point attractors but exhibited oscillatory limit cycles. This indicates that when large-scale cortical activity evolves according to the dynamics specified by the model equations, it is most likely to fall into a set of attractors. Note that the model algorithm is designed to identify an even number of attractors, each representing opposing patterns of neural activity (Supplementary Text S1).

We predicted that these attractors would tile the core gradients of cortical organization that span between the transmodal default mode network (DMN) areas and the unimodal sensory and motor areas[35,36]. Furthermore, we expected that these attractors would correspond to canonical brain states—distinct and recurring patterns of neural activity—that have been replicated in multiple studies[24,37]. To test these hypotheses, we combined the activity patterns of all fixed point attractors estimated in every run. We applied a k-means clustering to find 4 attractor clusters, because a majority of runs exhibited either four or less attractors (Fig. 4B).

The mean activity patterns of these attractor clusters (Fig. 4C) were compared to the known cortical gradients estimated by Margulies et al.[36] (Fig. 4D). This previous work applied a nonlinear dimensionality reduction algorithm on hundreds of participants' resting-state functional connectivity data to find top gradients that explained the largest variances. The primary gradient distinguished unimodal from transmodal areas, and the secondary gradient distinguished sensory from motor areas, which were argued to be an "intrinsic coordinate system" of the human brain[38] and replicated by multiple research groups[39].

Our two attractor clusters were highly comparable to the top cortical gradients (ρ values = 0.850 and 0.896, p values < 0.0001). Specifically, the attractor clusters served as axes that separated (i) the DMN from the unimodal (UNI) sensory and motor areas and (ii) the visual network (VIS) from the sensorimotor network (SM). We did not find meaningful clusters of attractors nor resemblance to gradients when the attractors were estimated from the randomly circular-shifted fMRI data (Supplementary Fig. S4). We also found no significant participant-level differences in attractor positions, indicating that the attractor locations were largely consistent across participants (Supplementary Text S2). These results provide a dynamical systems explanation for the recurrence of a small number of brain states over time: attractors correspond to brain states, the positions of these attractors are constrained by the brain's functional network organization, and neural dynamics unfold along this attractor landscape.

Geometry of neural dynamics differs across fluctuating attentional states during task and movie-watching

We characterized the attractor landscapes of large-scale neural dynamics, or the likely path that neural activity will take when following the model's equations (Fig. 4A). Would the geometry of these paths vary depending on a person's attentional state at a given moment? In other words, would one's attentional state—from a state of being focused on a task versus zoning out to being engaged in a movie versus being bored—relate to where the brain state is positioned within the hills and valleys?

Revisiting the forward simulation in Table 1, we identified two vectors at each time step based on the observed neural activity x_t: one defining the intrinsic drift, or the flow governed by regional connections and self-decays (called the "intrinsic vector" defined by $W\psi_\alpha(x_t) - D \odot x_t$, shortened as $\vec{WD}$), and the other defining the flow nudged by the external inputs at the corresponding moment (called the "extrinsic vector" defined by βu_t, shortened as $\vec{B}$). To quantify this, for every time step, we extracted the intrinsic and extrinsic vectors and calculated their angles and magnitudes (Fig. 5A). The angle indicates the direction of the vector with respect to the position of the attractor it eventually converged, with high angle indicating the vector directing away from the attractor. The magnitude represents the degree of change from the initial state, with high magnitude indicating a fast-moving vector. Because these measures were estimated at each time step, we were able to extract their time series over the course of the fMRI run. In essence, they provide a geometric characterization of neural dynamics on the attractor landscape.

We asked whether neural trajectories toward attractors—characterized by the angle and magnitude of the intrinsic and extrinsic activity—systematically varied based on the attentional state of a participant. We focused our analysis on the SONG dataset, which contained both naturalistic movie-watching (i.e., two runs of comedy

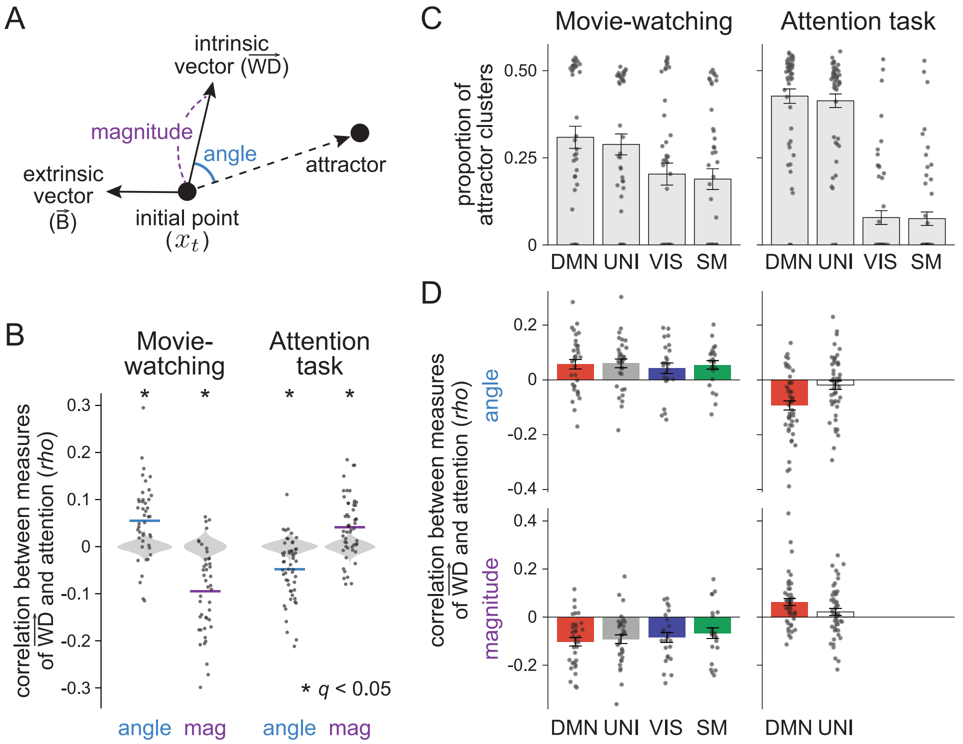

Fig. 5 | Neural dynamics toward attractors and their relationship with attention. A Geometric measures were estimated from neural activity at every time step at the first instance of forward simulation. In $x_{t+1} = x_t + W\psi_\alpha(x_t) - D \odot x_t + \beta u_t$, we decomposed a vector representing the intrinsic drift ($W\psi_\alpha(x_t) - D \odot x_t$) and a vector driven by external inputs (βu_t). The magnitude of the vector was calculated using the Euclidean norm. The direction of the vector was calculated using the cosine angle with respect to the position of the attractor. Because these four measures were calculated at every time step, we were able to extract their respective time series for each fMRI run. **B** Correlations between attention measures and the angle (blue) and magnitude (purple) of the intrinsic vectors across movie-watching and attention task runs, compared to the respective chance distribution (two runs in each condition; $N = 27$). Light grey areas indicate permuted chance distributions, black dots indicate correlation values between attention measures and the estimated angle or magnitude, and colored lines indicate the mean of these correlation values. Asterisks denote statistical significance shown in Table 2. Mag: magnitude. **C** Proportions of identified attractor clusters. The identified attractor at each time step was categorized to either one of the two ends of the primary gradient (default mode network [DMN] or unimodal network [UNI]) or the two ends of the secondary gradient (visual network [VIS] or somatosensory-motor network [SM]). The proportion of the attractor cluster was calculated at each run (black dot), which was averaged across runs to be summarized as a bar graph. Error bar indicates the standard error of the mean. **D** Correlations between individuals' attention measures and the angle and magnitude of the intrinsic vector, categorized based on the positions of the attractors. Black dots indicate estimates from every run of every participant. Colored bars indicate the mean of correlation values that are significantly different from the permuted chance distribution (corrected for false discovery rate, $q < 0.05$), whereas empty bars indicate non-significance. Because VIS and SM attractors were less likely to occur during tasks, they were excluded from the analyses. Supplementary Fig. S6 shows separate results for the two runs in each condition.

sitcom episode watching) and controlled attention tasks (i.e., two runs of gradual-onset continuous performance task; gradCPT)[24]. Importantly, these runs included time-varying behavioral measures of attention. During sitcom episodes, participants continuously rated how engaging they found the episode by adjusting the scale bar[40]. This measure was collected after the fMRI scan, as participants re-watched the same episode. (Narrative engagement has been characterized as a state of heightened emotional arousal and attentional focus[40–43]. Because narrative engagement accompanies changes in both arousal and attentional states, denoting them as "attentional state" is a simplification made in this article.) During gradCPT, participants pressed a button at every second whenever target images appeared inside the scanner. The inverse of response time variability served as a proxy of sustained attention, with moments of stable response times indicating high attention and variable response times indicating low attention[44]. We correlated the behavioral time series of each participant with the four geometric measures' time series, which was compared to a respective chance distribution where each geometric measure was correlated with circular-shifted behavioral time courses (Table 2 and Fig. 5B).

Both the angle and magnitude of intrinsic activity ($\overrightarrow{WD}$) were significantly correlated with attention dynamics (Table 2 and Fig. 5B). The relationships were comparable between repeated runs of the movie-watching and attention tasks, highlighting the reliability of our results. Interestingly, an opposite relationship to attention dynamics was found between the two contexts. The angle of the intrinsic vector was large when participants reported high engagement toward episodes, whereas the angle was small when participants performed stably in gradCPT. The magnitude of the intrinsic vector was small when participants reported high engagement toward episodes, whereas the magnitude was large when participants performed stably in gradCPT. This means that neural dynamics toward attractors—largely determined by the landscape—not only varied across attentional state dynamics but in a manner different across contexts.

On the other hand, the angle and magnitude of the extrinsic activity ($\vec{B}$) did not meaningfully relate to attention (Table 2). There were weak trends of correlations, which were seemingly derived from emergent correlations amongst the four geometric measures (Supplementary Fig. S5). This highlights that attention relates to changes in intrinsic neural dynamics, but not stimulus-driven neural dynamics.

In a study by Song et al.[24] in which this original data was collected, different brain states were associated with high attention in movie-watching and attention task contexts. Participants reported high engagement to sitcom episodes during the "base" state, a state where

Table 2 | Correlations between attention measures and the angle (representing direction) and magnitude (representing speed) of the neural trajectory across different conditions, compared to the respective chance distribution

	Movie		Task	
	Sitcom ep1	Sitcom ep2	GradCPT face	GradCPT scene
Angle ($\vec{WD}$, attractor)	$z = 3.570$, $q = 0.002$*	$z = 4.047$, $q < 0.0001$*	$z = -3.451$, $q = 0.002$*	$z = -3.277$, $q = 0.002$*
Magnitude ($\vec{WD}$)	$z = -4.779$, $q < 0.0001$*	$z = -5.776$, $q < 0.0001$*	$z = 2.147$, $q = 0.048$*	$z = 3.096$, $q = 0.005$*
Angle ($\vec{B}$, attractor)	$z = -2.275$, $q = 0.040$*	$z = 0.773$, $q = 0.440$	$z = 1.817$, $q = 0.093$	$z = 1.815$, $q = 0.093$
Magnitude ($\vec{B}$)	$z = -4.778$, $q < 0.0001$*	$z = -1.377$, $q = 0.207$	$z = 1.137$, $q = 0.293$	$z = 0.825$, $q = 0.432$

Analyses were conducted on four runs of the SONG dataset: two sitcom episodes watching and gradual-onset continuous performance task (gradCPT) with either face or scene images. False discovery rate (FDR) correction was applied to control for multiple comparisons across 16 significance tests. Asterisks indicate FDR-corrected $p < 0.05$.

no functional network exhibited dominant activity and was positioned at the center of the latent manifold. On the contrary, participants exhibited stable task performance to gradCPT during the DMN state, a state defined by high activity in the DMN (consistent with findings by refs. 45–47). Motivated by these results, we categorized attractors into either of the four ends of gradients 1 (DMN and UNI) and 2 (VIS and SM), based on the neural pattern similarity between attractors and the gradients. We assessed the relationships between the geometric measures and attention dynamics, separately for these four attractors. The goal was to see if the relationship, shown in Fig. 5B and Table 2, differs depending on where the neural trajectories converge.

We primarily found that the likelihood of attractors differed across the two contexts (Fig. 5C). The neural activity was nearly equally likely to fall into one of the four attractors during movie-watching, whereas the ends of gradient 1, the DMN and UNI, were much more likely to serve as attractors during sustained attention tasks, compared to the ends of gradient 2. This finding showing different likelihood of attractors across contexts indicates that the attractor landscapes differed across contexts.

During movie-watching, we found that the main relationship between the geometric measures and attention dynamics remained consistent, irrespective of which attractors the neural activity fell into. When participants reported high engagement to movies, the intrinsic drift directed away from the attractors with decreased magnitude (Fig. 5D). This implies that the brain was in a flattened or shallow region of the landscape when people were attentive, such that the brain activity was more likely to lie at the center of the manifold and gravitated less toward the attractors (Fig. 6A). On the contrary, during attention tasks, the main effect we found in Table 2 was specific to when the neural activity converged onto the DMN attractor (Fig. 5D). When attentive, neural activity approached the DMN attractor faster and more directly. This implies that the brain was on a steeper region of the attractor landscape near the DMN attractor when attentive (Fig. 6B).

Together, the results suggest that the neural dynamics toward attractors differed across high and low attentional states, in a context-dependent manner. This indicates that the attractor landscape of large-scale cortical dynamics systematically varies depending on attentional states and task demands.

Discussion

In this study, we fit a dynamical systems model to large-scale fMRI data to investigate how the geometry of neural dynamics differs across changes in attentional states in different contexts. By simulating brain dynamics over time, we identified stable attractors that aligned with cortical gradients separating transmodal from sensorimotor regions as well as sensory from motor regions. Neural trajectories toward these attractors systematically varied with moment-to-moment attentional state fluctuations, in a manner different across situational contexts. These results indicate that the geometry of neural dynamics along the attractor landscape reflects changing attentional states and different task demands across situations.

The dynamical systems model used in this study is a simplified neural mass model that is tailored to simulate large-scale regional activities and their interactions, rather than local neuronal activities within a region. We found that the model parameters effectively reproduced descriptive statistics such as functional connectivity and stimulus encoding and were sensitive to trait- and state-level differences. We compared the model parameters with these metrics because, although they are imperfect measures, they are the most widely used in the field and therefore provide an interpretable benchmark. Beyond reproducing descriptive statistics, the model's strength comes from decomposing components of neural activity that are driven by interactions between brain regions, autocorrelation within each region, and external inputs—which together explained nearly half of the variance in the observed neural activity. These parameter estimates became the basis for estimating attractors and trajectories along the landscape. Moreover, the model was not tailored specifically to fit fMRI data, meaning the model can be used with other data modalities. These together suggest that our model provides a mathematical description of neural dynamics that are generative, biologically plausible, and generalizable to other research domains.

Forward simulation of the model revealed a set of attractors to which neural activity was more likely to converge. While the model identified an attractor landscape that governs cortical dynamics, this does not imply that cortical activity must converge to those attractors. Instead, the landscape provides the underlying structure over which activity evolves, with trajectories continuously influenced by stimuli, task demands, and behavior. The attractors identified from this analysis recapitulated canonical brain states that were identified in previous studies, which tiled the known gradients of cortical hierarchy. Specifically, the attractors were marked by high activities in the DMN, VIS, and SM—an emergent property of the model rather than a feature imposed by the model design. This conceptually replicates many studies in human systems neuroscience that have revealed the existence of a low-dimensional manifold of macroscale neural activity[36,48–50] that is conserved across evolution[51,52], stable across contexts[53,54], and confined by structural architecture and genetic makeup of the brain[55–58]. This indicates that the positions of the attractors are largely fixed, confined by the brain's canonical functional architecture.

Although the positions of the attractors are largely determined, it does not mean that the attractor landscape is fixed. Rather, the attractor landscape as well as the traversal along its hills and valleys can be flexible, which motivated us to study neural dynamics along the attractor landscape in relation to attention dynamics. We found that not only did the attractor landscapes differ across situational contexts, but even within a context, regions occupied within the landscape varied depending on a person's attentional state. When participants were engaged in sitcom episodes, the intrinsic drift directed away from the attractors with decreased magnitude. This was a depiction of neural activity being less prone to fall into attractors as it situated on a shallow landscape at moments of engagement. In contrast, when participants were paying attention to an effortful psychological task, the intrinsic drift directed specifically toward the DMN attractor with

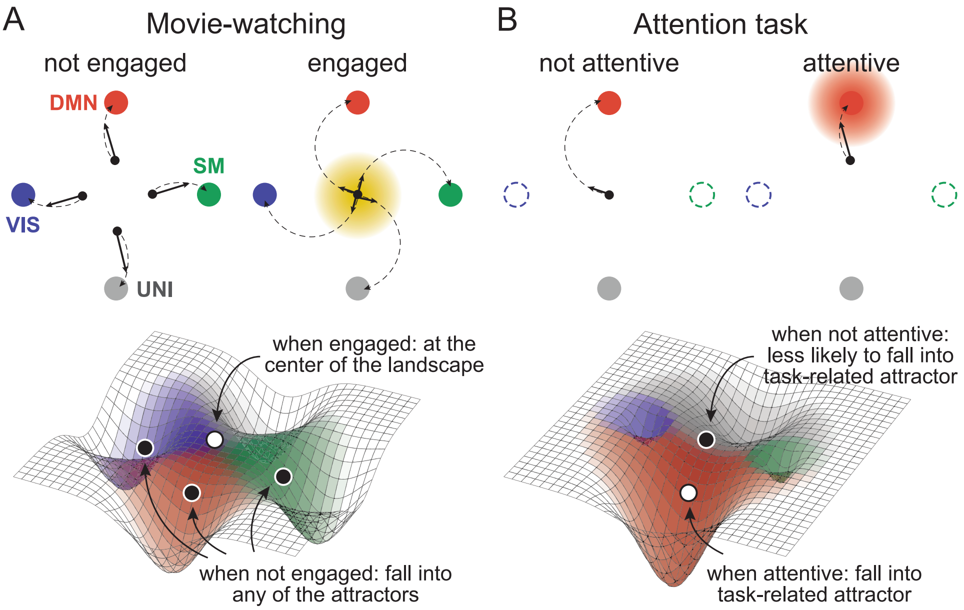
A
Movie-watching
not engaged
engaged
DMN
SM
VIS
UNI
B
Attention task
not attentive
attentive
when engaged: at the center of the landscape
when not engaged: fall into any of the attractors
when not attentive: less likely to fall into task-related attractor
when attentive: fall into task-related attractor

Fig. 6 | Neural geometry across attentional states and contexts. Schematic illustration of the results in Fig. 5D for **A** movie-watching and **B** attention task conditions. (*Top*) The angle and magnitude of the intrinsic vector during engaged vs. not engaged and attentive vs. not attentive states in movie-watching and attention task contexts. The black node illustrates the initial state and the four colored nodes illustrate attractors, with the vertical axis indicating gradient 1 and the horizontal axis indicating gradient 2 as in Fig. 4B. The illustration is based on the angle and magnitude of the vectors from the initial state to the attractors, not the positions of the initial states with respect to the attractors nor the distance amongst attractors. Black dashed lines illustrate intrinsic drifts of neural activity toward attractors. Black solid lines illustrate a one-time-step vector of the drift. Shaded areas indicate positions of brain states that were reported to be associated with high attention in Song et al.[24]. The VIS and SM attractors are dashed in attention task context because they are less likely to occur in this context (Fig. 5C). (*Bottom*) The 3D schematic illustrations of the attractor landscapes in the two contexts. White circles indicate brain states in attentive states and black circles indicate brain states in inattentive states. DMN default mode network, UNI unimodal network, VIS visual network, SM somatosensory-motor network.

increased magnitude. This indicates that the neural activity more easily fell into the DMN attractor because it was in a steeper landscape when attentive. This is in line with studies that reported high activity in DMN at moments of optimal performance during this task[45–47]. These results highlight the flexible geometry of neural dynamics on the large-scale attractor landscape—it systematically changes across attentional state fluctuations.

The findings that brain activity tended to lie on a shallow attractor landscape when engaged in sitcoms and a steep local attractor when attentive to tasks resemble results reported by Munn et al.[26], which characterized cortical dynamics within an energy landscape framework. Energy landscape formulations conceptualize fluctuations in brain activity in terms of the energy required to move between brain states and have been the basis for influential hypotheses regarding functions such as predictive coding, learning, and inference[59]. Munn et al.[26], specifically, found evidence of flattened cortical landscape (i.e., decreased energy) upon activation of the noradrenergic arousal system, which projects broadly across the cortex, and deepened cortical landscape (i.e., increased energy) upon activation of the cholinergic vigilance system, which projects to relatively local functional networks. This suggests a hypothesis that different neuromodulatory circuits may underlie internal states of being immersed in engaging narratives versus being attentive to effortful and controlled tasks. In line with previous findings that engagement correlates with perceived emotional arousal during narratives[40–43], our results hint that the attractor landscape of being engaged may resemble a state of heightened arousal, more so than heightened vigilance. Future work can address this hypothesis by administering pharmacological agents that modulate noradrenergic and cholinergic activity during similar task and movie-watching conditions.

In sum, the attractor landscape is flexible across situational contexts, and cortical dynamics along this landscape reflect changes in attentional states. By modeling neural dynamics, we offer a geometric framework that can explain how internal states arise from large-scale brain activity.

Methods

Model description

The dynamical systems model used in this study is adopted from the neural mass model called the mesoscale individualized neurodynamic (MINDy) model[9,19]. The model is designed to fit the neural activity time series—whichever units or scales the neural activities are sampled from—and the time-aligned experimental variables, such as task, stimulus, or behavior. For our use, the neural activity corresponded to the BOLD activity of 200 cortical parcels collected from human fMRI. The experimental variable corresponded to audiovisual and semantic feature embeddings of the movies that participants watched inside the scanner and visual feature embeddings of the images that were presented as task stimuli. However, the choice of neural units and experimental variables can vary depending on the study and research question. No tailoring specific to the fMRI data (e.g., deconvolution of the hemodynamic response function) was made.

To reiterate, the model is defined as the following equation, with the neural activity at time t represented as x_t ($x_t \in \mathbb{R}^n$, $n = 200$ parcels) and stimuli at time t represented as u_t ($u_t \in \mathbb{R}^p$, $p = 100$ PCs of the feature embeddings), with Δt set to 1 TR.

$$\hat{x}_{t+1} = x_t + W\psi_\alpha(x_t) - D \odot x_t + \beta u_t \tag{4}$$

$$\psi_\alpha(x_t) = \sqrt{\alpha^2 + (bx_t + 0.5)^2} - \sqrt{\alpha^2 + (bx_t - 0.5)^2} \tag{5}$$

The weight matrix ($W \in \mathbb{R}^{n\times n}$) represents directional interaction between neural units, which conceptually corresponds to effective connectivity. $W\psi_\alpha(x_t)$ represents the weighted sum of the neural unit's nonlinearly transformed activity at time t multiplied by its directed connections from every other neural unit. To constrain the estimation of W as the sum of sparse random component and a low-rank structured component[60], we decomposed the parameter into $W = W_S + W_L$ where W_S ($W_S \in \mathbb{R}^{n\times n}$) represents a sparse matrix after L_1 regularization and $W_L = W_1 W_2^T$ ($W_{1,2} \in \mathbb{R}^{n\times r}$) is given by low-rank approximation ($r = n/3$) with sparsity also given to W_1 and W_2 with L_1 regularization. This formalization allows the model to capture global low-dimensional motifs of connectivity superimposed on a sparse network, which gives rise to the known low-dimensional and modular structure of large-scale functional organization.

Nonlinearity follows a parametrized sigmoidal transfer function (ψ_a), which maps neural activity to a bounded output at a range of −1 to 1. The slope of the transfer function is determined by the estimated α. b is fixed at 20/3 in our model.

$\beta \in \mathbb{R}^{n\times p}$ represents a linear coupling between neural activity and incoming stimuli that are time-aligned with one another. The time was aligned by convolving the stimulus time series with the canonical hemodynamic response function. β conceptually corresponds to a linear mapping from the stimulus space to the neural space.

A decay term ($D \in \mathbb{R}^n$) represents convergence to baseline activity at the absence of external inputs or interactions, with $\odot$ denoting element-wise product. D is an algorithmically important parameter, because D is initially set to a value significantly higher than 1, which flips the right-hand side of the equation toward a large negative factor of x_t. This accentuates the difference in neural activity between units, thus allowing effective estimation of the directed connectivity W. From our empirical tests, if D is set to a biologically plausible value of <1, the estimated W does not recapitulate the descriptive functional connectivity measure.

Model fitting

We fit the neural activity time series acquired at each run, in batches of size 300 TRs. Specifically, we predicted the neural activity of consecutive time steps $\hat{x}_{2:T}$ based on the observed neural activity $x_{1:T-1}$ (where $T = 300$). Model parameters were optimized across 2500 iterations using stochastic gradient descent, specifically the Nesterov-accelerated Adaptive Moment Estimation optimizer as chosen in previous studies[9,19]. The loss $\mathscr{L}$ was calculated as follows. Λ represents the regularization term, which enforces sparsity in connections. Regularization terms were fixed to $\lambda_1 = 0.075$, $\lambda_2 = 0.2$, $\lambda_3 = 0.05$.

$$\mathscr{L} = \frac{1}{2}||\hat{x}_{2:T} - x_{2:T}||_2^2 + \Lambda \tag{6}$$

$$\Lambda = \lambda_1||W_S||_1 + \lambda_2 \mathrm{Tr}(|W_S|) + \lambda_3(||W_1||_1 + ||W_2||_1) \tag{7}$$

Because D was initialized at a value higher than 1, the prediction accuracy (i.e., rank correlation between the predicted and observed neural activity patterns) started at a value near −1 at the start of the training. The prediction accuracy increased gradually across iterations. In contrast, W quickly became comparable to a functional connectivity matrix in the initial phase of training, but across more iterations, it gradually approached a diagonally dominant matrix, which is conceptually similar to the first-order autoregressive model. To prioritize biologically meaningful parameter optimization rather than brain activity prediction, we stopped the training at 2500 iterations (in batches of 300 consecutive TRs) at which point W was similar to functional connectivity, but the model's accuracy still remained negative. To boost prediction accuracy, we fit an ordinary least squares linear regression model to estimate pW, pD, and pB, which are scalar values ($pW, pD, pB \in \mathbb{R}$). Parameters W, D, and β were scaled by these value estimates, respectively.

$$\dot{x}_t = pW^*W\psi_\alpha(x_t) - pD^*D \odot x_t + pB^*\beta u_t \tag{8}$$

Table 3 | List of hyperparameters and parameter initializations

Learning rate	2.5e-5
Training iterations	2500
Batch size	300
W_S, W_1, W_2	$\mathscr{N}(0, 0.01^2)$
D	$\mathscr{N}(5, 0.5^2)$
α	$\mathscr{N}(5, 0.05^2)$
b	20/3
β	$\mathscr{N}(0, 0.01^2)$
λ_1	0.075
λ_2	0.2
λ_3	0.05

Model fitting and ensuing analyses were conducted in Python (v3.12.2) and Pytorch (v2.4.1).

Hyperparameters

Hyperparameter selection largely followed the original implementation of the MINDy model[9], but with some simplifications made (Table 3).

FMRI datasets

Two openly available fMRI datasets were analyzed: the SONG dataset (participants recruited in South Korea; $N = 27$) and the HCP dataset (participants recruited in the USA; $N = 119$). Both datasets were approved by the institutional review boards and de-identified for sharing. No information on participants' age, sex, or gender was collected in the present study. Preprocessing steps followed Song et al.[24]. We applied a 200-parcel cortical atlas by Schaefer et al.[29] where the BOLD activities of voxels corresponding to each parcel were averaged to represent the parcel activity[9,19]. The SONG dataset includes two runs of resting-state, two runs of controlled sustained attention task called the gradual-onset continuous performance task (gradCPT), two runs of comedy sitcom-watching, and one run of educational documentary-watching (3 T scans with TR = 1 s). The HCP dataset includes four runs of resting-state in 3 T (TR = 0.72 s) and four in 7 T (TR = 1 s), two runs of working memory tasks in 3 T (6 different types of cognitive task runs were excluded from analyses because the total TRs were less than 350 TRs), and four runs of movie-watching in 7 T. These runs varied in total duration, ranging from 405 to 1486 TRs.

Input features

Low-level visual features were characterized by hue, saturation, and pixel intensity, which were estimated per frame and averaged across frames within an event (rgb2hsv function in MATLAB R2024a). In the gradCPT run with face images, hue and saturation were excluded from the analyses because the images were presented in grayscale. Low-level audio features were represented with amplitude and pitch of left and right stereos, which were estimated per frame and averaged across frames within an event (audioread and pitch functions in MATLAB). Visuo-semantic features were represented by 512-dimensional embeddings of OpenAI's pretrained Contrastive Language-Image Pretraining model[61] (huggingface; clip-vit-base-patch32). For audio-semantic features, we first applied OpenAI's WhisperX model to the audio file to transcribe speech and dialogues, along with their time stamps[62]. Transcripts were sampled at a TR resolution and were transformed using 512-dimensional embeddings of Google's Universal Sentence Encoder[63] (tensorflow v2.19.0; https://tfhub.dev/google/universal-sentence-encoder-multilingual/3). A mean value of the respective dimension was assigned to moments when speech or dialogue did not exist.

Embedding time series were convolved with a canonical hemodynamic response function and normalized across time per dimension. Principal component analyses were conducted on the concatenated embedding time series of the 3 movie-watching runs in SONG and 4 runs in HCP, respectively. We selected the top 100 principal component time series, which explained 74.97% and 77.69% of variance, respectively. Principal component analysis was conducted on the gradCPT face run in isolation, given that the same images were presented in the same sequence for all participants (100.00% of explained variance for 100 principal components). For gradCPT scene runs, because presented images and sequences differed for all participants, participant-specific embeddings were concatenated for a principal component analysis (85.34% of explained variance). Resulting principal component time series were again normalized across time per dimension to be used as stimulus input u in the model.

Validation of model parameters

Parameters were estimated through training the model on the data collected from an individual fMRI run (2500 training iterations). After parameters were fixed, we predicted the next-time-step neural activity from the observed neural activity time series to calculate explained variance (r^2). Note that the data used for training and testing were the same. When comparing parameter W with an undirected functional connectivity matrix (parcel-by-parcel Fisher's transformed Pearson's correlation coefficients), we took the average of the upper and lower triangles of W (region $i \rightarrow j$ and region $j \rightarrow i$ in a directed graph) and took the cosine similarity with the edge strengths of the functional connectivity matrix. Due to the low temporal resolution of the fMRI data, our W estimates are largely symmetrical. Encoding coefficients were estimated using an ordinary least squares regression with a residual term. Cosine similarity between all values in β and values in encoding coefficients were computed. For both metrics, values in descriptive statistics were randomly shuffled 10,000 times, which served as the respective chance distribution. Z statistics and two-tailed p values were calculated with respect to chance distributions. Cosine similarities were calculated between parameter estimates of run pairs. The cosine similarity values were grouped into whether they correspond to the same vs. different individuals or the same vs. different movie stimuli, which were compared using Wilcoxon rank sum tests.

Attractor clusters and gradients

Attractors were estimated by forward simulating the observed neural activity pattern at each time step based on the estimated model parameters (Table 1). For each time step, we performed 5000 simulations. If there was no longer a change in the last 10 iterations of the 5000th forward simulations ($||\dot{x}||_\infty < 1e-6$), we considered the system to have converged to a fixed point. If the Euclidean distance between the pair of estimated fixed points obtained from different time steps was less than 0.1, those were grouped as the same attractor. Using this approach, we found that most fMRI runs converged to 2 or 4 attractors. Principal component analysis was applied to each individual fMRI run only for visualization purpose in a 3D space (Fig. 4A).

We then aggregated attractors (each defined as a neural activity pattern across 200 parcels) across all fMRI runs and applied k-means clustering to identify 4 attractor clusters. The choice of $k=4$ was motivated by the observation that most runs exhibited either 2 or 4 attractors, with higher numbers occurring less frequently. The mean activity pattern of each cluster represented the attractor cluster (Fig. 4B, C). Two pairs of clusters exhibited a pattern correlation of −1. Principal component analysis was also applied to the aggregated attractor patterns, only for visualization purpose (Fig. 4B).

Cortical voxel gradients estimated by Margulies et al.[36] were obtained from NeuroVault (https://identifiers.org/neurovault.collection:1598). Gradient values were averaged within each parcel to generate gradient estimates of the 200 parcels (Fig. 4D). The neural activity patterns of the attractor clusters were then compared to the parcel-level gradient values using Spearman's rank correlation.

Relating speed and direction of neural dynamics toward attractors with measures of attention

With the neural activity pattern at each time step as an initial state (x_t), we considered one-step forward simulation of $\hat{x}_{t+1} = x_t + W\psi_\alpha(x_t) - D \odot x_t + \beta u_t$ and separated vectors representing internally-driven change ($W\psi_\alpha(x_t) - D \odot x_t$, shortened to $\vec{WD}$) and externally-driven change (βu_t, shortened to $\vec{B}$). The angle between the respective vector and the eventual end state was calculated with the following equation,

$$\theta = \arccos\left(\frac{v1 \cdot v2}{||v1|| \cdot ||v2||}\right) \tag{9}$$

where $v1 \cdot v2$ represents the dot product of the two vectors, and $||v1||$ and $||v2||$ are the Euclidean norms of the vectors. Before applying the inverse cosine, the cosine similarity was clipped to the range of [−1, 1]. The magnitude of the vector was calculated as the Euclidean norm. Because these measures were estimated at each time step, repeating this across all time steps within a run resulted in their time series.

To probe attentional state changes during movie-watching, Song et al.[24] asked participants to continuously rate their engagement, on a scale of 1 to 9, as they re-watched the movies after the fMRI scan. Each participant's engagement rating was normalized across time and convolved with a hemodynamic response function. To probe attentional state changes during gradCPT, we analyzed participants' button response times as in the previous study[24]. After linear interpolation of no button response trials and regressing out the linear trend, we calculated response time variability by taking the deviance from the mean response time at every TR. Because studies showed that moments of low response time variability correspond to high sustained attention or optimal task performance, the inverse response time variability time course was used as a proxy for attention dynamics[44]. Again, the inverse response time variability was normalized across time and convolved with a hemodynamic response function. See Song et al.[24] for details of behavioral experiments and data analyses.

Angle and magnitude time series were correlated with attention measures sampled at a TR resolution using Spearman's rank correlation, for each individual run. The mean of all participants' Fisher's r-to-z transformed ρ values were compared to permuted null distributions to calculate z statistics, where behavioral time series were circular-shifted across time with a random multiplication of either 1 or −1 (10,000 iterations). FDR correction was applied across 16 significance tests (Table 2).

The relationship between geometric measures and attention was analyzed separately depending on the position of the attractor at each time step, or where the eventual end state lies at the final round of forward simulation. Because we found that attractor clusters lie at the ends of the known primary and secondary gradients (Fig. 4), we categorized the attractor of each time step to one of the four ends of the two gradients. The attractor was labeled as either the DMN, UNI, VIS, or SM, based on the highest correlation coefficient. Correlation between the geometric measure and attention was calculated from a subset of time series corresponding to the respective attractor cluster category. In a majority of participants' gradCPT runs, VIS and SM attractors did not appear. Because fewer than 10 fMRI participants' gradCPT face or gradCPT scene runs included VIS and SM attractors, statistical analysis was not conducted for these cases.

Significance was again tested by comparing the mean of Fisher's r-to-z transformed ρ values to null distributions created from circular-shifted behavioral time series. Twelve comparisons were corrected for in Fig. 5D, and 24 comparisons were corrected for in Supplementary Fig. S6B using FDR correction.

Reporting summary

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

Data availability

Processed fMRI and behavioral data are openly available in ref. 64: https://github.com/hyssong/dynamicalsystems. Raw data of the SONG dataset can be accessed in OpenNeuro[24]: https://openneuro.org/datasets/ds004592/versions/1.0.1. Processed SONG data used in this study can be found in GitHub: https://github.com/hyssong/dynamicalsystems/tree/main/data/song2023elife. HCP dataset was accessed through the ConnectomeDB[30–32]: https://www.humanconnectome.org/study/hcp-young-adult. A list of participants analyzed in the study can be found in Github: https://github.com/hyssong/neuraldynamics/blob/main/behavior/hcpparticipants.csv. Source data for Fig. 5B–D are provided with this paper.

Code availability

Model and analysis codes are openly available in ref. 64: https://github.com/hyssong/dynamicalsystems.

References

1. Breakspear, M. Dynamic models of large-scale brain activity. *Nat. Neurosci.* **20**, 340–352 (2017).
2. Vyas, S., Golub, M. D., Sussillo, D. & Shenoy, K. V. Computation through neural population dynamics. *Annu. Rev. Neurosci.* **43**, 249–275 (2020).
3. Hodgkin, A. L. & Huxley, A. F. A quantitative description of membrane current and its application to conduction and excitation in nerve. *J. Physiol.* **117**, 500–544 (1952).
4. Wilson, H. R. & Cowan, J. D. Excitatory and inhibitory interactions in localized populations of model neurons. *Biophys. J.* **12**, 1–24 (1972).
5. Honey, C. J., Kötter, R., Breakspear, M. & Sporns, O. Network structure of cerebral cortex shapes functional connectivity on multiple time scales. *Proc. Natl. Acad. Sci. USA* **104**, 10240–10245 (2007).
6. Deco, G. & Jirsa, V. K. Ongoing cortical activity at rest: criticality, multistability, and ghost attractors. *J. Neurosci.* **32**, 3366–3375 (2012).
7. Demirtaş, M. et al. Hierarchical heterogeneity across human cortex shapes large-scale neural dynamics. *Neuron* **101**, 1181–1194 (2019).
8. Wang, P. et al. Inversion of a large-scale circuit model reveals a cortical hierarchy in the dynamic resting human brain. *Sci. Adv.* **5**, eaat7854 (2019).
9. Singh, M. F., Braver, T. S., Cole, M. W. & Ching, S. Estimation and validation of individualized dynamic brain models with resting state fMRI. *NeuroImage* **221**, 117046 (2020).
10. Singh, M. F., Braver, T. S., Cole, M. & Ching, S. Precision data-driven modeling of cortical dynamics reveals person-specific mechanisms underpinning brain electrophysiology. *Proc. Natl. Acad. Sci. USA* **122**, e2409577121 (2025).
11. Baker, A. P. et al. Fast transient networks in spontaneous human brain activity. *eLife* **3**, e01867 (2014).
12. Chen, S., Langley, J., Chen, X. & Hu, X. Spatiotemporal modeling of brain dynamics using resting-state functional magnetic resonance imaging with Gaussian hidden Markov model. *Brain Connect.* **6**, 326–334 (2016).
13. Vidaurre, D., Smith, S. M. & Woolrich, M. W. Brain network dynamics are hierarchically organized in time. *Proc. Natl. Acad. Sci. USA* **114**, 12827–12832 (2017).
14. Liu, X., Zhang, N., Chang, C. & Duyn, J. H. Co-activation patterns in resting-state fMRI signals. *NeuroImage* **180**, 485–494 (2018).
15. Yousefi, B. & Keilholz, S. Propagating patterns of intrinsic activity along macroscale gradients coordinate functional connections across the whole brain. *NeuroImage* **231**, 117827 (2021).
16. Kelso, J. A. S. Multistability and metastability: understanding dynamic coordination in the brain. *Philos. Trans. R. Soc. B: Biol. Sci.* **367**, 906–918 (2012).
17. Cocchi, L., Gollo, L. L., Zalesky, A. & Breakspear, M. Criticality in the brain: a synthesis of neurobiology, models and cognition. *Prog. Neurobiol.* **158**, 132–152 (2017).
18. Roberts, J. A. et al. Metastable brain waves. *Nat. Commun.* **10**, 1056 (2019).
19. Chen, R., Singh, M., Braver, T. S. & Ching, S. Dynamical models reveal anatomically reliable attractor landscapes embedded in resting-state brain networks. *Imaging Neurosci.* **3**, imag_a_00442 (2025).
20. Greene, A. S., Horien, C., Barson, D., Scheinost, D. & Constable, R. T. Why is everyone talking about brain state?. *Trends Neurosci.* **46**, 508–524 (2023).
21. Yamashita, A., Rothlein, D., Kucyi, A., Valera, E. M. & Esterman, M. Brain state-based detection of attentional fluctuations and their modulation. *NeuroImage* **236**, 118072 (2021).
22. Taghia, J. et al. Uncovering hidden brain state dynamics that regulate performance and decision-making during cognition. *Nat. Commun.* **9**, 2505 (2018).
23. Cornblath, E. J. et al. Temporal sequences of brain activity at rest are constrained by white matter structure and modulated by cognitive demands. *Commun. Biol.* **3**, 261 (2020).
24. Song, H., Shim, W. M. & Rosenberg, M. D. Large-scale neural dynamics in a shared low-dimensional state space reflect cognitive and attentional dynamics. *eLife* **12**, e85487 (2023).
25. John, Y. J. et al. It's about time: linking dynamical systems with human neuroimaging to understand the brain. *Netw. Neurosci.* **6**, 960–979 (2022).
26. Munn, B. R., Müller, E. J., Wainstein, G. & Shine, J. M. The ascending arousal system shapes neural dynamics to mediate awareness of cognitive states. *Nat. Commun.* **12**, 6016 (2021).
27. Sara, S. J. The locus coeruleus and noradrenergic modulation of cognition. *Nat. Rev. Neurosci.* **10**, 211–223 (2009).
28. Hasselmo, M. E. & Sarter, M. Modes and models of forebrain cholinergic neuromodulation of cognition. *Neuropsychopharmacol* **36**, 52–73 (2011).
29. Schaefer, A. et al. Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. *Cereb. Cortex* **28**, 3095–3114 (2018).
30. Barch, D. M. et al. Function in the human connectome: task-fMRI and individual differences in behavior. *NeuroImage* **80**, 169–189 (2013).
31. Van Essen, D. C. et al. The WU-minn human connectome project: an overview. *NeuroImage* **80**, 62–79 (2013).
32. Finn, E. S. & Bandettini, P. A. Movie-watching outperforms rest for functional connectivity-based prediction of behavior. *NeuroImage* **235**, 117963 (2021).
33. Cole, M. W., Bassett, D. S., Power, J. D., Braver, T. S. & Petersen, S. E. Intrinsic and task-evoked network architectures of the human brain. *Neuron* **83**, 238–251 (2014).
34. Gratton, C. et al. Functional brain networks are dominated by stable group and individual factors, not cognitive or daily variation. *Neuron* **98**, 439–452.e5 (2018).

35. Mesulam, M. M. From sensation to cognition. *Brain* **121**, 1013–1052 (1998).
36. Margulies, D. S. et al. Situating the default-mode network along a principal gradient of macroscale cortical organization. *Proc. Natl. Acad. Sci. USA* **113**, 12574–12579 (2016).
37. Bolt, T. et al. A parsimonious description of global functional brain organization in three spatiotemporal patterns. *Nat. Neurosci.* **25**, 1093–1103 (2022).
38. Huntenburg, J. M., Bazin, P.-L. & Margulies, D. S. Large-scale gradients in human cortical organization. *Trends Cogn. Sci.* **22**, 21–31 (2018).
39. Bernhardt, B. C., Smallwood, J., Keilholz, S. & Margulies, D. S. Gradients in brain organization. *NeuroImage* **251**, 118987 (2022).
40. Song, H., Finn, E. S. & Rosenberg, M. D. Neural signatures of attentional engagement during narratives and its consequences for event memory. *Proc. Natl. Acad. Sci. USA* **118**, e2021905118 (2021).
41. Busselle, R. & Bilandzic, H. Measuring narrative engagement. *Media Psychol.* **12**, 321–347 (2009).
42. Bilandzic, H., Sukalla, F., Schnell, C., Hastall, M. R. & Busselle, R. W. The narrative engageability scale: a multidimensional trait measure for the propensity to become engaged in a story. *Int. J. Commun.* **13**, 32 (2019).
43. Ke, J., Song, H., Bai, Z., Rosenberg, M. D. & Leong, Y. C. Dynamic brain connectivity predicts emotional arousal during naturalistic movie-watching. *PLoS Comput. Biol.* **21**, e1012994 (2025).
44. Rosenberg, M., Noonan, S., DeGutis, J. & Esterman, M. Sustaining visual attention in the face of distraction: a novel gradual-onset continuous performance task. *Atten. Percept. Psychophys.* **75**, 426–439 (2013).
45. Esterman, M., Rosenberg, M. D. & Noonan, S. K. Intrinsic fluctuations in sustained attention and distractor processing. *J. Neurosci.* **34**, 1724–1730 (2014).
46. Fortenbaugh, F. C., Rothlein, D., McGlinchey, R., DeGutis, J. & Esterman, M. Tracking behavioral and neural fluctuations during sustained attention: a robust replication and extension. *NeuroImage* **171**, 148–164 (2018).
47. Kucyi, A. et al. Electrophysiological dynamics of antagonistic brain networks reflect attentional fluctuations. *Nat. Commun.* **11**, 325 (2020).
48. Hong, S.-J. et al. Toward a connectivity gradient-based framework for reproducible biomarker discovery. *NeuroImage* **223**, 117322 (2020).
49. Shafiei, G. et al. Topographic gradients of intrinsic dynamics across neocortex. *eLife* **9**, e62116 (2020).
50. Dong, H.-M., Margulies, D. S., Zuo, X.-N. & Holmes, A. J. Shifting gradients of macroscale cortical organization mark the transition from childhood to adolescence. *Proc. Natl. Acad. Sci. USA* **118**, e2024448118 (2021).
51. Oligschläger, S. et al. Gradients of connectivity distance in the cerebral cortex of the macaque monkey. *Brain Struct. Funct.* **224**, 925–935 (2019).
52. Xu, T. et al. Cross-species functional alignment reveals evolutionary hierarchy within the connectome. *NeuroImage* **223**, 117346 (2020).
53. Cross, N. et al. Cortical gradients of functional connectivity are robust to state-dependent changes following sleep deprivation. *NeuroImage* **226**, 117547 (2021).
54. Samara, A., Eilbott, J., Margulies, D. S., Xu, T. & Vanderwal, T. Cortical gradients during naturalistic processing are hierarchical and modality-specific. *NeuroImage* **271**, 120023 (2023).
55. Burt, J. B. et al. Hierarchy of transcriptomic specialization across human cortex captured by structural neuroimaging topography. *Nat. Neurosci.* **21**, 1251–1259 (2018).
56. Paquola, C. et al. Shifts in myeloarchitecture characterise adolescent development of cortical gradients. *eLife* **8**, e50482 (2019).
57. Vázquez-Rodríguez, B. et al. Gradients of structure–function tethering across neocortex. *Proc. Natl. Acad. Sci. USA* **116**, 21219–21227 (2019).
58. Pang, J. C. et al. Geometric constraints on human brain function. *Nature* **618**, 566–574 (2023).
59. Friston, K. The free-energy principle: a unified brain theory?. *Nat. Rev. Neurosci.* **11**, 127–138 (2010).
60. Mastrogiuseppe, F. & Ostojic, S. Linking connectivity, dynamics, and computations in low-rank recurrent neural networks. *Neuron* **99**, 609–623.e29 (2018).
61. Radford, A. et al. Learning transferable visual models from natural language supervision. *Proceedings of the 38th International Conference on Machine Learning*, (PMLR 139, 2021).
62. Bain, M., Huh, J., Han, T. & Zisserman, A. WhisperX: time-accurate speech transcription of long-form audio. *INTERSPEECH* (2023).
63. Cer, D. et al. Universal sentence encoder. Preprint at https://arxiv.org/abs/1803.11175 (2018).
64. Song, H. et al. Geometry of neural dynamics along the cortical attractor landscape reflects changes in attention. *Zenodo* https://doi.org/10.5281/zenodo.17978061 (2025).

Acknowledgements

We thank JeongJun Park with conceptualization and helpful discussions and comments on the manuscript. The research was supported by the McDonnell Center for Systems Neuroscience and the McDonnell Center for Cellular and Molecular Neurobiology at Washington University in St. Louis (H.S.) and NIH R01NS130693 (S.C.).

Author contributions

H.S.: Conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, software, visualization, writing–original draft, writing–review & editing. R.C.: Investigation, methodology, software, writing–review & editing. T.L.B.: Data curation, investigation, methodology, writing–review & editing. T.S.B.: Investigation, methodology, writing–review & editing. M.D.R.: Conceptualization, data curation, investigation, writing–review & editing. J.M.Z.: Investigation, supervision, writing–review & editing. S.C.: Conceptualization, funding acquisition, investigation, methodology, supervision, writing–original draft, writing–review & editing.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-026-69041-8.

Correspondence and requests for materials should be addressed to Hayoung Song.

Peer review information *Nature Communications* thanks Joon-Young Moon, Brandon Munn, and the other, 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