

Neural Decoding as Cognitive Inference

Yi Guo^{1,2}, Changhong Jing^{1,3}, Yong Hu¹, Yan Liu³, Michael K. P. Ng⁴,
Shanshan Wang^{1,2}, Shuqiang Wang^{1,2}

¹Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China.

²University of Chinese Academy of Sciences, Beijing, China.

³Department of Computing, The Hong Kong Polytechnic University, Hong Kong.

⁴Department of Mathematics, Hong Kong Baptist University, Hong Kong.

Abstract

The brain maintains stable cognition despite continuously changing neural activity. How to extract stable cognitive states from variable neural observations remains a central problem in neural decoding. Existing neural decoding methods map neural observations to predefined external labels based on the stimulus-response principle, often capturing recording-specific spurious correlations. Inspired by how the brain infers the world, and specifically by Bayesian brain theory, we recast neural decoding as cognitive inference constrained by brain-intrinsic priors, yielding high-level meta-neural semantic representations. In decoding experiments spanning five neural recording modalities and three cognitive domains (motor, perception and internal mentation), our cognitive inference method reorganized the geometry of neural observation representations, yielding meta-neural semantic representations that exhibited consistent geometric relationships across cognitive tasks and enabled the recovery of stable cognitive states from variable neural observations. Our work provides an account of how the brain maintains relatively stable cognition despite continual changes in the external environment. Cognitive stability is sustained through cognitive inference from changing neural activity, without requiring fixed neural activity patterns.

Introduction

The brain maintains relatively consistent cognitive states despite continuous changes in neural activity^{1,2}. Neural observations vary with task context, individual differences and recording sessions, whereas cognitive states remain stable. Similarly, audition and vision depend on different sensory inputs and evoke distinct patterns of neural activity, yet the high-level neural representations derived from this activity show systematic semantic consistency³. A central problem in neural decoding is therefore how to recover stable cognitive

states from changing neural observations⁴⁻⁶. Most existing neural decoding methods remain based on the stimulus-response principle. These methods map neural observations onto predefined labels and often capture recording-specific spurious correlations^{7,8}. If the stability of cognitive states does not depend on fixed patterns of neural activity, rigid stimulus-response mappings are insufficient for decoding stable cognition.

To overcome this limitation, neural decoding needs to reconsider how the brain understands the external world^{9,10}. According to Bayesian brain

theory, the brain does not passively respond to sensory evidence but actively infers the external world from sensory evidence, with brain-intrinsic priors constraining the inference^{11–15}. The brain’s cognitive states are therefore not directly determined by sensory evidence. Even when sensory evidence differs, the brain may infer consistent cognitive states. For neural decoding, this implies that recovering stable cognitive states from changing neural observations requires inference constrained by brain-intrinsic priors.

Building on this view, we recast neural decoding as cognitive inference. Neural observations provide evidence about cognitive states, while brain-intrinsic priors constrain the inference of those states from neural observations. On this basis, we proposed a hierarchical decoding framework inspired by Bayesian brain theory. The framework infers cognitive states from changing neural observations and thereby yields high-level representations that we term meta-neural semantic representations (Fig. 1). Within this framework, brain-intrinsic priors may arise from structural constraints of the brain. Previous studies have shown that cortical geometry systematically constrains neural activity throughout the cerebral cortex and that cortical geometric eigenmodes accurately reconstruct the cortex-wide spatial distribution of this activity^{16–18}. Accordingly, we introduced cortical geometry as a brain-intrinsic prior into cognitive inference. The cortical geometry prior is constructed independently of the cognitive states to be inferred and does not directly encode their specific content. This independence allows us to test whether consistent cognitive states can still be

inferred from variable neural observations despite changes in the underlying patterns of neural activity.

We systematically evaluated the framework using five neural modalities: EEG, MEG, ECoG, fNIRS and fMRI. The experiments covered three functional domains, Motor, Perception and Internal mentation, and assessed whether cognitive inference could recover consistent cognitive states from variable neural observations. (1) Compared with neural observation representations, meta-neural semantic representations were more broadly distributed throughout the cortex and exhibited more consistent geometric relations under different cognitive conditions. (2) Meta-neural semantic representations preserved consistent cognitive state relations among different tasks, and decoding performance remained stable despite neural drift over days to months. (3) Neural observations differed markedly between auditory and visual perception, whereas the inferred meta-neural semantic representations remained aligned with the semantic structure of cognitive content. Cognitive state consistency was preserved despite differences in sensory input. (4) Cognitive inference in the proposed framework recovered internally generated cognitive content from neural observations. The framework extended neural decoding from externally evoked content to internal mentation. Overall, cognitive inference in the proposed framework recovered stable cognitive states from variable neural observations. These findings support the view that cognitive stability does not depend on fixed patterns of neural activity but may instead be

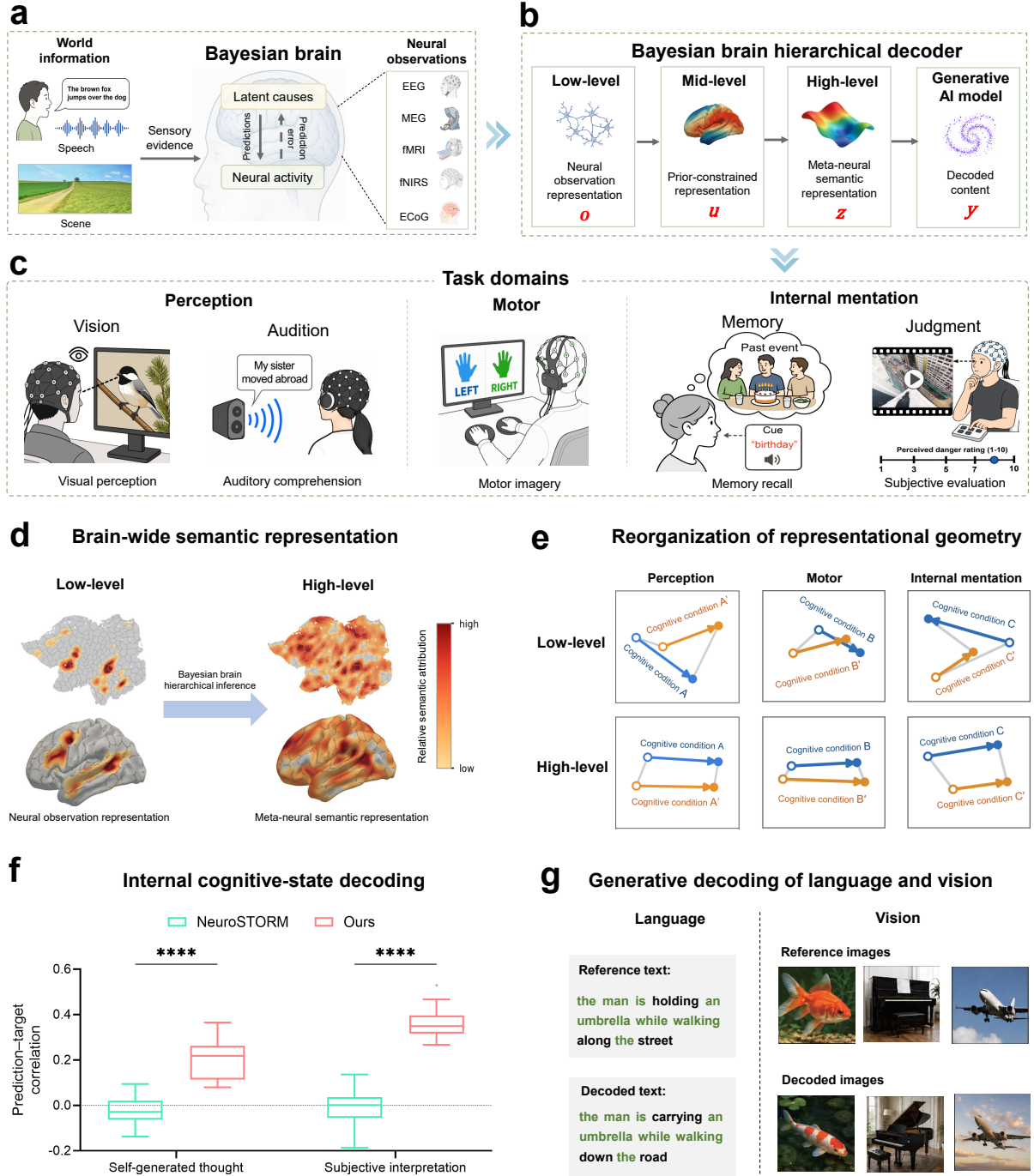


Fig. 1 Overview of the Bayesian brain hierarchical decoding framework. **a**, Bayesian brain account linking speech and scene information to latent causes and neural activity measured with EEG, MEG, fMRI and fNIRS and ECoG. **b**, Hierarchical decoder comprising low-level neural observation o , mid-level prior-constrained u and high-level meta-neural semantic z representations, with z supplied to a generative AI model to produce decoded output y . **c**, Task domains: perception (vision and audition), motor (motor imagery) and internal mentation (memory and judgment). **d**, Cortical semantic-attribution maps showing localized low-level and brain-wide high-level representations. **e**, Representational geometry of low-level and high-level representations across the three task domains. Open and filled circles distinguish cognitive states; blue and orange distinguish cognitive conditions. Arrows connect states within each condition; greater parallelism across conditions indicates greater cognitive stability. **f**, Prediction-target correlations for self-generated thought and subjective interpretation, comparing NeuroSTORM with the proposed framework. Boxes show medians and interquartile ranges (IQRs); whiskers extend to the most extreme values within $1.5 \times \text{IQR}$ of the box limits, with outliers shown individually. **** $P < 0.0001$. **g**, Representative open-ended language and visual decoding outputs. Green words mark shared key semantic content in reference and decoded text.

maintained through continual cognitive inference from variable neural activity.

Results

Brain-intrinsic priors constrain cognitive inference to yield brain-wide meta-neural semantic representations

The Bayesian brain hierarchical decoding framework used cortical geometric eigenmodes to instantiate brain-intrinsic structural priors and inferred meta-neural semantic representations constrained by these priors (Fig. 2a).

Cognitive inference-based encoding used the cognitive state representation constrained by cortical geometric eigenmodes to predict cortical activity, whereas stimulus-response-based encoding mapped stimulus semantics to evoked cortical responses. Cognitive inference-based encoding produced significant predictions over a broader cortical range and achieved higher noise-ceiling-normalized squared prediction correlation than stimulus-response-based encoding (Fig. 2b–c). These results demonstrate the advantage of the framework in encoding broadly distributed cortical responses constrained by brain-intrinsic priors. This validation was conducted on the Natural Scenes Dataset (NSD; $n = 8$ participants).

In NSD, cortical searchlight RSA showed broader cortical correspondence for high-level meta-neural semantic representations than for low-level image features (51.59% versus 28.28%;

Fig. 2d). In SMN4Lang, cortical semantic attribution was more broadly distributed for the high-level meta-neural semantic representation than for the low-level neural observation representation (Fig. 2e; $n = 12$ participants). From a neural-field perspective, attribution of the meta-neural semantic representation concentrated in long-wavelength cortical geometric eigenmodes, peaked at approximately 84 mm and decreased with eigenmode order (Spearman $\rho = -0.820$; Fig. 2f). This scale dependence related the broader semantic range of the representation to large-scale cortical fields.

The whole-cortex semantic-connectivity coefficient was higher for the proposed framework than for CogReader (0.33 versus 0.05), with the same pattern observed within each of the seven Yeo networks (Fig. 2g). The corresponding chord diagrams showed denser relations between network pairs and more long-range connections, as reflected by greater cortical geodesic distances (Fig. 2h). These results indicate that the meta-neural semantic representation was more closely aligned with distributed brain-network structure than the CogReader representation.

Cognitive inference constrained by cortical geometry yielded brain-wide meta-neural semantic representations with large-scale spatial relationships. Cognitive state information was thus organized under cortical structural constraints extending beyond local neural activity. These findings support brain-intrinsic structure as a relatively stable constraint on cognitive inference from limited neural observations.

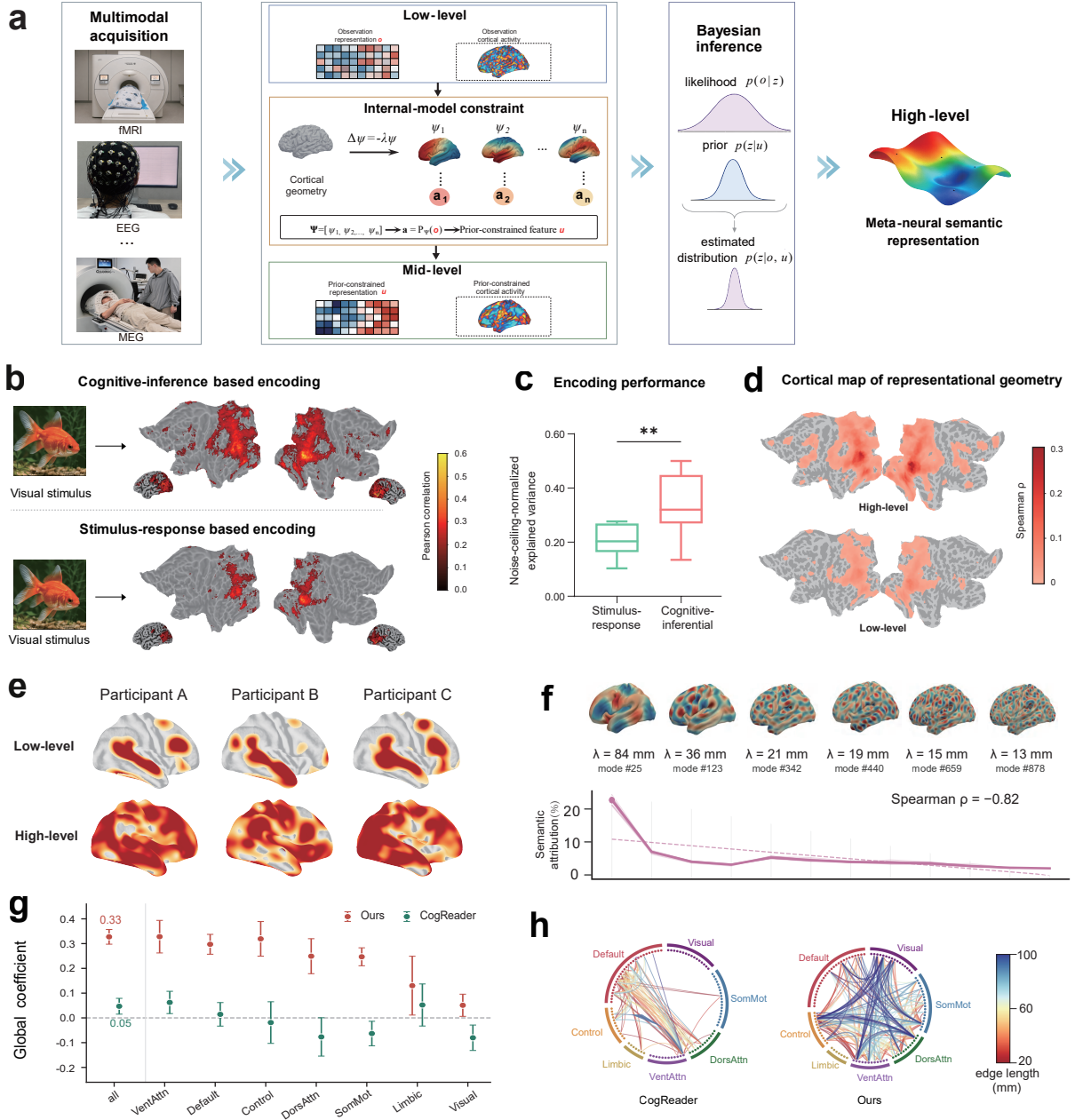


Fig. 2 Brain-intrinsic priors shape brain-wide meta-neural semantic representations. **a**, Cortical geometric eigenmodes instantiate brain-intrinsic priors in the hierarchical decoding framework and constrain inference from neural observations to meta-neural semantic representations. **b**, Vertex-wise correlations between predicted and measured cortical activity for cognitive inference-based encoding and stimulus-response encoding in NSD ($n = 8$); evaluation stimuli were held out during model fitting. **c**, Noise-ceiling-normalized squared prediction correlation for the two encoding models. Boxes show medians and quartiles; whiskers extend to the most extreme values within $1.5 \times \text{IQR}$ of the box limits, with outliers shown individually. **d**, Cortical searchlight RSA maps in NSD show broader cortical correspondence for high-level meta-neural semantic representations than for low-level image features. Color denotes mean Spearman correlation. **e–h**, Results from SMN4Lang ($n = 12$). **e**, Semantic-attribution maps for three representative participants (Participants A–C). Maps for all 12 participants are shown in Extended Data Fig. 1. **f**, Semantic attribution across cortical geometric eigenmodes; the line and shading show the participant mean and s.e.m. The association with eigenmode order was assessed using Spearman’s rank correlation. **g**, Semantic-connectivity coefficients across the whole cortex and seven Yeo networks for CogReader and the proposed framework; values show the participant mean $\pm$ s.e.m. **h**, Thresholded interregional relations for the two methods, grouped by Yeo network. Edge color denotes cortical geodesic length. A two-sided paired t -test was used in **c**. Asterisks denote statistical significance ($*P < 0.05$, $**P < 0.01$, $***P < 0.001$ and $****P < 0.0001$).

Cognitive inference enables generalizable decoding in diverse tasks

We evaluated the generalizability of discriminative decoding based on meta-neural semantic representations across tasks spanning motor, perception and internal mentation. The evaluation considered representational geometry, within-participant decoding accuracy and zero-shot decoding in unseen participants.

Three-dimensional MDS projections showed that state-difference directions were more consistent across cognitive conditions at the high level across motor, perception and internal mentation (Fig. 3a). Quantitative analyses further showed greater cross-condition generalization performance and stronger parallelism of factor directions at the high level across the three domains (Fig. 3b). Complementary analyses showed lower within-state dispersion, greater between-state centroid distance and higher category-structure RSA at the high level (Extended Data Fig. 3). These results reveal a consistent geometric reorganization of the meta-neural semantic representation, in which task-relevant structure was preserved more consistently across cognitive conditions.

Within-participant decoding in BCIC, FACED, SEED-V and Motor Imagery covered motor-imagery, affective-state and movement-state decoding¹⁹⁻²². Among the four evaluated methods, the proposed framework achieved the highest participant-averaged accuracy in all four datasets compared with LaBraM, BrainOmni and CBraMod (Fig. 3c)²³⁻²⁵. The box plots showed consistently

higher median accuracies for the proposed framework, with participant-level distributions shifted above those of the three comparison methods. The corresponding confusion matrices were dominated by diagonal entries in each dataset (Fig. 3d). Thus, the geometric reorganization of the meta-neural semantic representation was accompanied by higher decoding accuracy across multiple tasks.

The proposed framework also achieved the highest mean zero-shot decoding accuracy in each dataset compared with LaBraM, BrainOmni and CBraMod (Fig. 3e). All four methods were fitted using only the training participants and applied to unseen participants without participant-specific adaptation. This result extends the performance of the meta-neural semantic representation from within-participant decoding to zero-shot decoding in unseen participants.

Across motor, perception and internal mentation, cognitive inference yielded more consistent relationships among cognitive states despite differences in tasks and neural observations. This geometric consistency was accompanied by improved within-participant and zero-shot decoding. The results suggest that distinct neural activity patterns can support common relationships among cognitive states without requiring identical neural representations across tasks.

Cognitive inference enables stable decoding across time despite neural drift

Meta-neural semantic representations preserved cognitive state relationships and supported cross-session decoding despite neural drift. Repeated

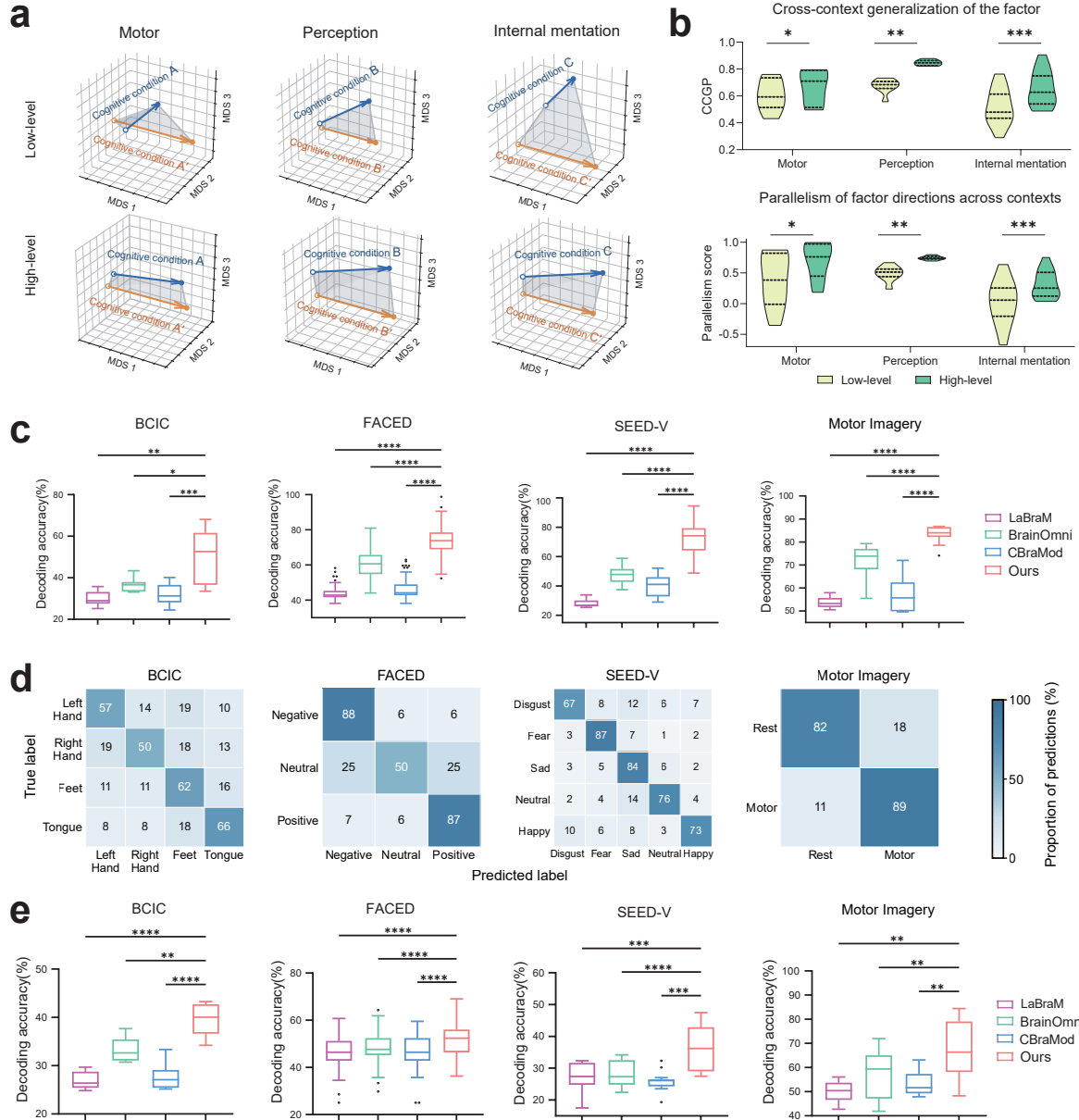


Fig. 3 Meta-neural semantic representations exhibit consistent geometric relationships and improved decoding performance across diverse tasks. **a**, Three-dimensional MDS projections of the low-level neural observation and high-level meta-neural semantic representations in motor, perception and internal mentation, using the Motor Imagery, NSD and Spacemap datasets, respectively. Open and filled circles distinguish cognitive states; blue and orange distinguish cognitive conditions. Arrows connect cognitive states within each condition, and pale lines connect corresponding states across conditions. Greater parallelism across conditions indicates greater cognitive stability. **b**, Cross-condition generalization performance (CCGP) and parallelism score (PS) for the low-level and high-level representations. CCGP measures the generalization of cognitive state decoding across conditions, whereas PS measures the parallelism of the corresponding state-difference directions. Violin plots show the distributions of both metrics. **c**, Within-participant decoding accuracy for LaBraM, BrainOmni, CBraMod and the proposed framework across BCIC, FAcED, SEED-V and Motor Imagery ($n = 9, 123, 16$ and 10 participants, respectively). The datasets comprised four motor-imagery classes, three affective states, five affective states and two movement states, respectively. **d**, Within-participant confusion matrices for the proposed framework; rows denote true classes, columns predicted classes and cells row-normalized trial percentages. **e**, Zero-shot decoding accuracy for participants excluded from model training. In **c** and **e**, boxes show medians and quartiles; whiskers extend to the most extreme values within $1.5 \times \text{IQR}$ of the box limits, with outliers shown individually. Points denote individual participants. One-way repeated-measures ANOVA with Greenhouse–Geisser correction followed by Šidák-adjusted multiple comparisons was used in **c** and **e**. Asterisks denote statistical significance ($*P < 0.05$, $**P < 0.01$, $***P < 0.001$ and $****P < 0.0001$).

recordings spanned 208 days in ECoG Speech and 4 days in AJILE12, both invasive ECoG datasets, and 8 days in non-invasive SHU-MI EEG. These datasets covered speech-command production, natural upper-limb movement and motor imagery.

Normalized amplitude profiles changed markedly between recording sessions in the low-level neural observation representation but remained more consistent in the high-level meta-neural semantic representation (Fig. 4a, c, e). The corresponding cross-session similarity matrices showed higher between-session similarity at the high level in ECoG Speech, AJILE12 and SHU-MI (Fig. 4b, d, f). Mean similarity increased from 0.393 to 0.803 in ECoG Speech, from 0.459 to 0.819 in AJILE12, and from 0.214 to 0.758 in SHU-MI. The high-level representation therefore preserved condition-specific amplitude patterns as low-level neural observations drifted across sessions.

In ECoG Speech, the high-level representation retained clear state separation in both the reference recording and the 42-day follow-up, whereas the low-level representation showed greater within-class dispersion (Fig. 4h). This difference was also reflected in higher silhouette coefficients at the high level (Fig. 4i). Moreover, the meta-neural semantic representation preserved more consistent representational trajectory geometry across sessions, whereas the corresponding low-level trajectories changed more substantially between D0 and D42 (Fig. 4g). Representational displacement relative to the reference recording remained lower at the high level throughout the 208-day period, including follow-ups more than 200 days after day 0

(shaded interval in Fig. 4j). The respective low-level and high-level displacement values were 17.5 and 3.7 in ECoG Speech, 12.7 and 2.5 in AJILE12, and 17.9 and 5.3 in SHU-MI. These values corresponded to reductions by factors of three to five across the three datasets (Fig. 4k). This stability was accompanied by higher mean cross-session decoding accuracy. Across 20 matched random seeds, the framework showed higher mean cross-session accuracy than SPaRCNet at every ECoG Speech follow-up and than both SPaRCNet and BrainOmni across all three datasets (Fig. 4l-m).

Over days to months of neural drift, meta-neural semantic representations preserved relatively stable relationships among cognitive states and supported cross-session decoding. Cognitive states could therefore be recovered consistently from time-varying neural observations, indicating that stable recovery does not require fixed neural activity patterns.

Cognitive states exhibit semantic consistency across perceptual domains

Meta-neural semantic representations supported open-ended decoding of cognitive content across language and vision.

Language decoding was evaluated in Alice, LPPC-fMRI, StudyForrest and LPPC-HK, covering English, Mandarin Chinese, Cantonese, French and German^{26–29}. BGEScore measured the sentence-level semantic similarity between decoded and reference text³⁰. The median BGEScore was 0.588 in Alice, 0.638 in LPPC-fMRI, 0.559 in

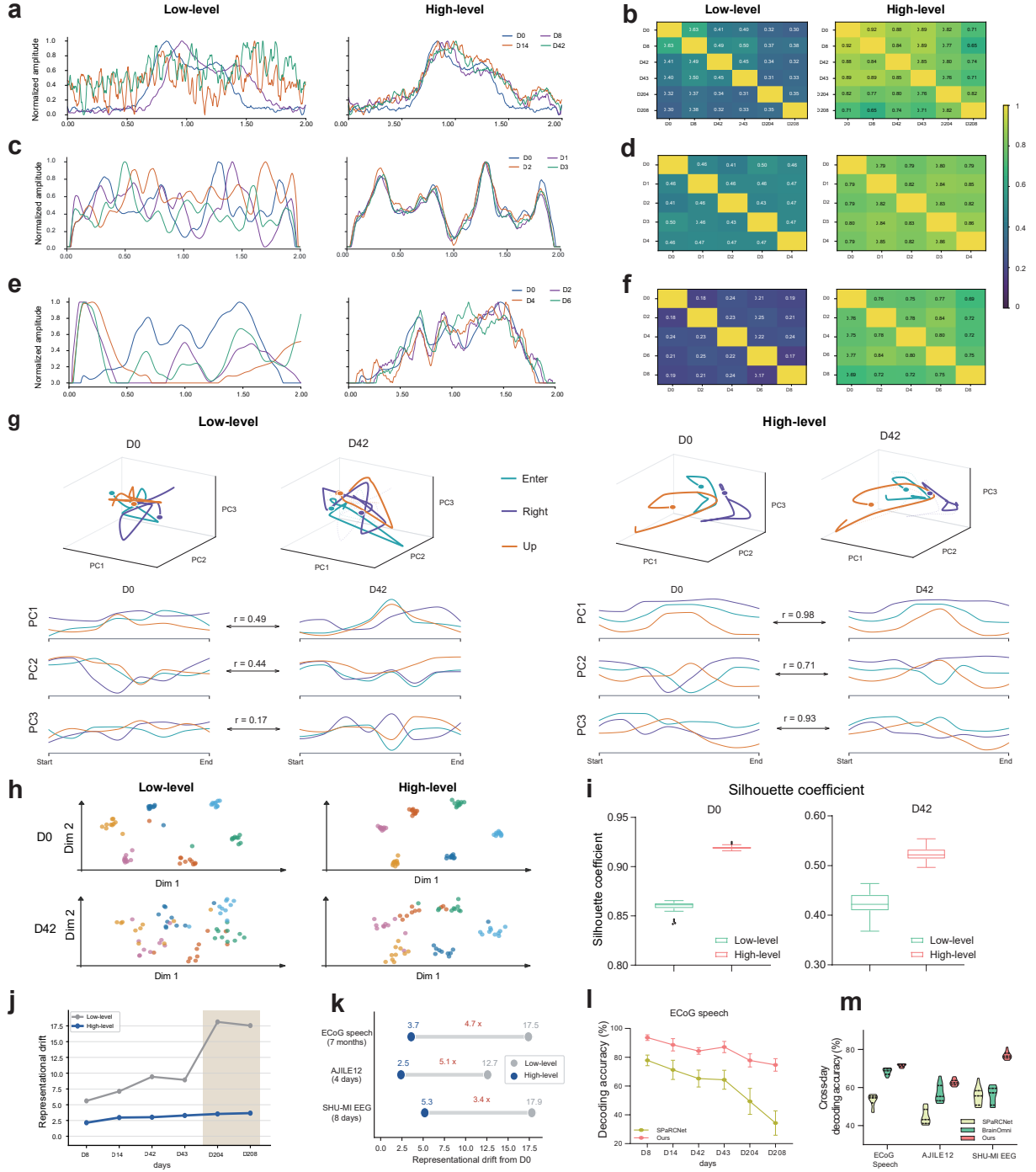


Fig. 4 Meta-neural semantic representations support stable decoding across days to months of neural drift. **a, c, e**, Normalized amplitude profiles of the low-level neural observation and high-level meta-neural semantic representations across sessions in ECoG Speech, AJILE12 and SHU-MI, respectively. **b, d, f**, Corresponding cross-session similarity matrices shown on a common scale. **g**, Low-level and high-level representational trajectories at D0 and D42 in ECoG Speech, projected onto the first three principal components; temporal profiles and Pearson correlations between D0 and D42 are shown for each component. **h**, t-SNE projections at D0 and D42. **i**, Silhouette coefficients at D0 and D42 across 50 matched initializations. Boxes show medians and quartiles; whiskers extend to the most extreme values within $1.5 \times \text{IQR}$ of the box limits, with outliers shown individually. **j**, Representational displacement from D0 across 208 days in ECoG Speech; shading marks follow-ups beyond 200 days. **k**, Low-level and high-level representational displacement across the three datasets. **l**, Mean ECoG Speech cross-session decoding accuracy for SPARCNet and the proposed framework at each follow-up; points and error bars show mean $\pm$ s.d. across 20 matched random seeds. **m**, Mean cross-session decoding accuracy for SPARCNet, BrainOmni and the proposed framework across ECoG Speech (six classes), AJILE12 (two classes) and SHU-MI EEG (two classes). Violin plots show distributions across the same 20 matched random seeds; internal boxes indicate medians and interquartile ranges.

StudyForrest and 0.631 in LPPC-HK, and the proposed framework outperformed CogReader and the random-output control in all four datasets (all $P < 0.0001$; Fig. 5a)³¹. This advantage was sustained throughout entire narratives, as illustrated by the Chinese and French stories in LPPC-fMRI (Fig. 5b–c). Representative decoded passages retained central entities, actions and event relations from the reference narratives (Fig. 5d). These results show that the framework reconstructed text semantics more accurately across multiple languages.

Image decoding from meta-neural semantic representations was evaluated within participants in NSD ($n = 8$ participants)³². Representational similarity analysis (RSA) showed a stronger correspondence between the high-level meta-neural semantic representation and image-description semantics than between the low-level neural observation representation and image-description semantics ($P < 0.01$; Fig. 5e). Compared with Neural Code Conversion (NCC)³³, the proposed framework achieved higher top-1, top-5 and top-10 image-retrieval accuracy ($P < 0.001$ for top-1 and $P < 0.0001$ for top-5 and top-10; Fig. 5f), higher CLIP similarity for reconstruction of viewed images (visual perception) ($P < 0.01$; Fig. 5g) and mental images (visual imagery) ($P < 0.05$; Fig. 5h)³³. Representative retrieval results identified images that matched the target semantics, and representative reconstructions retained the principal objects, scenes and visual content of viewed and imagined targets (Fig. 5i–j).

In language and vision, cognitive inference recovered semantic content consistent with the target cognitive states despite differences in sensory inputs and neural activity patterns. This semantic consistency in both perceptual domains supported open-ended decoding, showing that cognitive content could be recovered through cognitive inference across different forms of sensory input.

Cognitive inference extends neural decoding from externally evoked content to internal mentation

Internal cognitive states can arise endogenously during processes such as imagined speech, memory recall and self-generated thought, and can vary between individuals even in response to the same external stimuli. We therefore asked whether these states could be inferred from neural observations using the Bayesian brain hierarchical decoding framework.

In the Imagined Speech task, we used an optically pumped magnetometer magnetoencephalography (OPM-MEG) dataset collected for this study ($n = 6$ participants). Participants read a text prompt and then imagined its content during a 4-s imagery phase; only neural observations from this phase were used (Fig. 6a). The framework achieved a median semantic correlation of 0.567, which was higher than that obtained with BrainOmni ($P < 0.05$; Fig. 6b). Median three-class accuracy reached 48%, compared with 33% for BrainOmni ($P < 0.05$; Fig. 6c). These results show that meta-neural semantic representations preserved the semantic content of internally generated cognitive states during imagined speech.

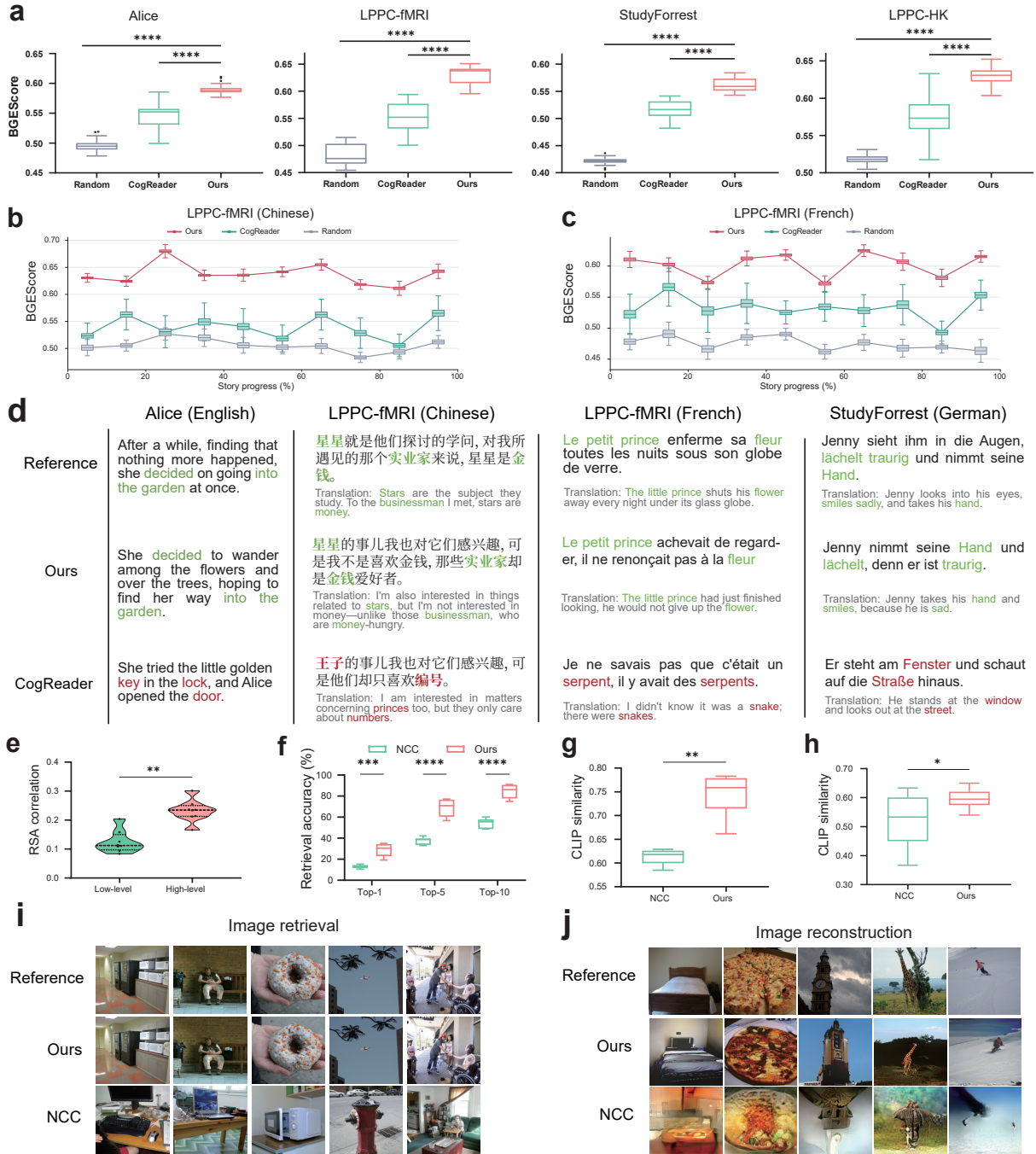


Fig. 5 Cognitive inference enables open-ended decoding of cognitive states across language and visual perceptual domains. **a**, Sentence-level BGE score between decoded and reference text for the proposed framework, CogReader and a random-output control across Alice, LPPC-fMRI, StudyForrest and LPPC-HK ($n = 26, 112, 20$ and 52 participants, respectively). **b–c**, Segment-wise BGE score for Chinese ($n = 35$) and French ($n = 28$) LPPC-fMRI participants, respectively; lines and error bars show mean $\pm$ s.e.m. **d**, Representative decoded passages in English, Chinese, French and German. Green highlighting denotes content semantically consistent with the reference, whereas red highlighting denotes divergent content. **e–h**, Within-participant visual decoding in NSD ($n = 8$). **e**, RSA between image-description semantics and the low-level neural observation or high-level meta-neural semantic representation. **f–h**, Comparison with NCC for top- k image retrieval ($k = 1, 5$ and 10) (**f**), viewed-image reconstruction (**g**) and mental-imagery reconstruction (**h**). For each retrieval trial, the candidate set comprised the target image and 99 randomly selected non-target images from the 1,000-image test set. **i**, Representative image-retrieval results. **j**, Representative image reconstructions. Boxes in **a** and **f–h** show medians and quartiles; whiskers extend to the most extreme values within $1.5 \times \text{IQR}$ of the box limits, with outliers shown individually. The violin plot in **e** shows participant-level distributions with medians and quartiles. Two-tailed paired t -tests were used in **a** against CogReader and in **e–h**; two-tailed Welch’s t -tests were used in **a** against the random-output control. Asterisks denote statistical significance (* $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$ and **** $P < 0.0001$).

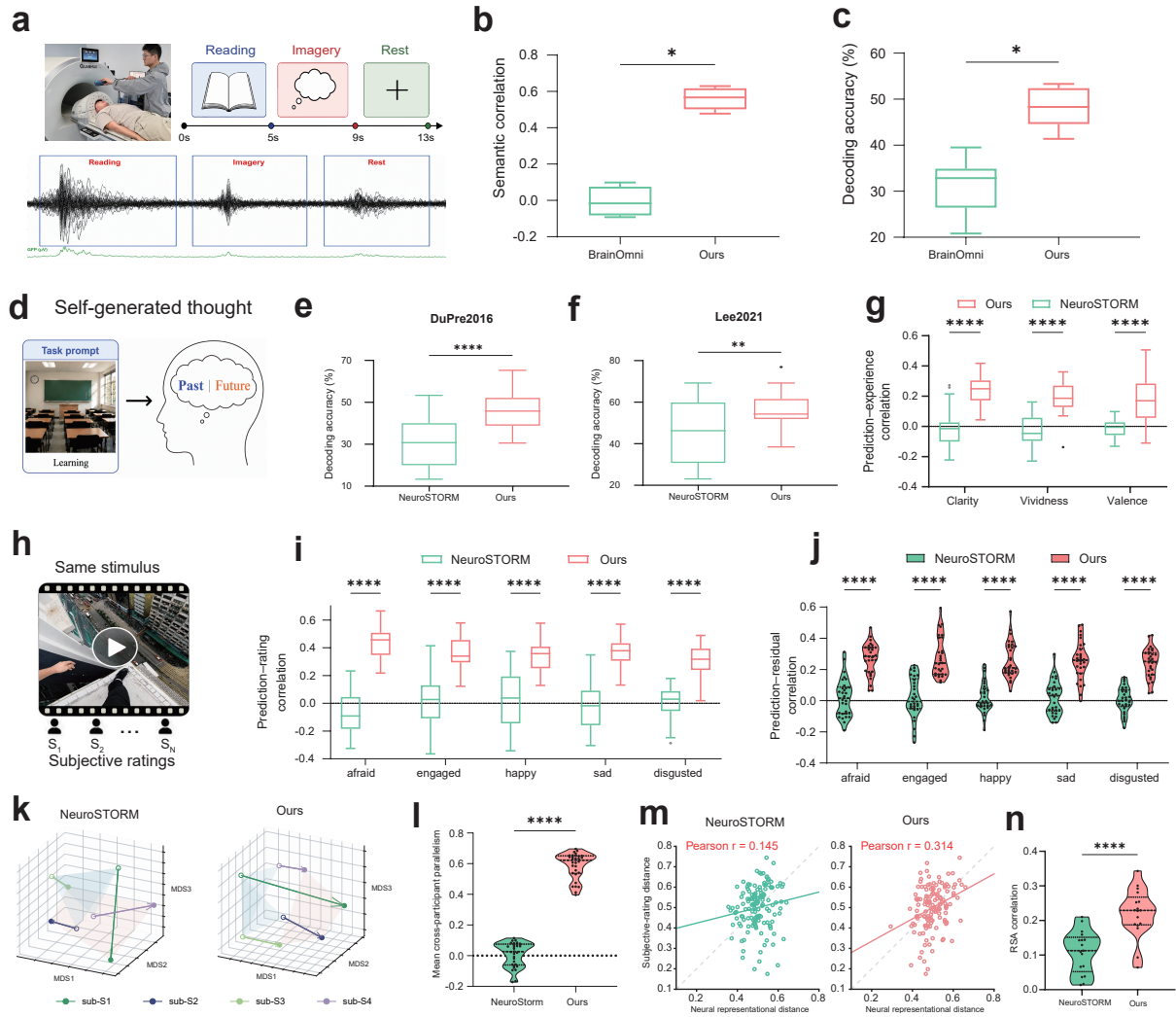


Fig. 6 Meta-neural semantic representations capture cognitive states in internal mentation. **a–c**, Imagined Speech OPM-MEG experiment ($n = 6$). **a**, Reading, imagery and rest phases with corresponding event-related fields. **b**, Semantic correlation during imagery for BrainOmni and the proposed framework. **c**, Three-class semantic-category decoding accuracy for imagined content. **d–g**, Self-generated thought in two public fMRI datasets, comparing NeuroSTORM with the proposed framework. **d**, State-classification tasks. **e**, Three-class decoding accuracy in DuPre2016 ($n = 31$). **f**, Two-class decoding accuracy in Lee2021 ($n = 26$). **g**, Correlations between predicted and reported clarity, vividness and valence. **h–j**, Participant-specific subjective ratings in Spacemap Alignvideo ($n = 30$), comparing NeuroSTORM with the proposed framework. **h**, Shared films and subjective ratings. **i**, Prediction-rating correlations across five emotional dimensions. **j**, Residual prediction-rating correlations after removal of the film-specific component shared across participants. **k**, Representational geometry of happiness for four selected participants, comparing NeuroSTORM with the proposed framework. Arrows represent participants and point towards higher happiness ratings; greater parallelism indicates more consistent representational relationships between happiness states across participants. **l**, Mean cross-participant parallelism of representational changes associated with happiness ratings ($n = 30$). **** $P < 0.0001$. **m**, Pearson correlations between neural representational distance and subjective-rating distance across 136 pairs of 17 participants. **n**, Participant-level RSA correlations between neural representations and subjective ratings in the same participants. Boxes in **b**, **c**, **e–g** and **i** show medians and quartiles; whiskers extend to the most extreme values within $1.5 \times \text{IQR}$ of the box limits, with outliers shown individually. In **l**, violin plots show distributions of participant-level mean parallelism, with individual participants shown as black points. Violin plots in **j** and **n** show participant-level distributions, with internal lines indicating medians and quartiles. Two-tailed paired t -tests were used in **b**, **c**, **e**, **f** and **n**. For **l**, the between-representation difference in mean parallelism was tested using a participant-level leave-one-out jackknife standard error and a two-sided t approximation (29 d.f.), with Holm correction across five emotional dimensions. RSA coefficients in **n** were Fisher z -transformed before testing. Two-way repeated-measures ANOVA followed by Šidák-adjusted multiple comparisons was used in **g**, **i** and **j**. Asterisks denote statistical significance (* $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$ and **** $P < 0.0001$).

Self-generated thought was assessed in the DuPre2016³⁴ and Lee2021 fMRI datasets³⁵. The classification tasks distinguished among self-generated thought states, including autobiographical recollection and imagined future events (Fig. 6d). The framework achieved higher state-classification accuracy than NeuroSTORM in DuPre2016 ($P < 0.0001$; Fig. 6e) and Lee2021 ($P < 0.01$; Fig. 6f)³⁶. Predictions also tracked the clarity of autobiographical recollection and the vividness and valence of imagined future events, with higher prediction–experience correlations than those obtained with NeuroSTORM for all three measures ($P < 0.0001$ for all comparisons; Fig. 6g). Thus, the framework distinguished self-generated thought states and predicted continuous variation in their reported experience.

Subjective feelings elicited by shared films were assessed in Spacetop Alignvideo. The 30 participants with the most complete film ratings viewed the same videos and rated fear, engagement, happiness, sadness and disgust³⁷ (Fig. 6h). Prediction–rating correlations were higher for the proposed framework than for NeuroSTORM in all five dimensions ($P < 0.0001$ for all comparisons; Fig. 6i). To determine whether this performance captured participant-specific experience rather than only film-related responses shared across participants, the shared film-specific component was removed separately from the observed ratings and model predictions. Correlations between the residual components remained higher than those obtained with NeuroSTORM in all five dimensions ($P < 0.0001$ for all comparisons;

Fig. 6j). The residual correlations therefore captured participant-specific subjective variation after accounting for responses to each film that were shared across participants.

Participant-specific decoding was accompanied by consistent representational relationships between subjective states across individuals. For happiness, these relationships were more parallel across individuals in the proposed framework than in NeuroSTORM (mean parallelism, 0.59 versus 0.01; $n = 30$; $P < 0.0001$; Fig. 6k–l). Similar increases were observed for sadness, fear, disgust and engagement (0.37–0.52 versus -0.01 –0.08; all $P < 0.001$). These results indicate that meta-neural semantic representations captured consistent relationships between subjective states across individuals. Across participant pairs, neural representational distance was related to subjective-rating distance for the 17 participants with complete ratings in all five dimensions. The association reached $r = 0.314$ for the proposed framework and $r = 0.145$ for NeuroSTORM (Fig. 6m), indicating that participants with more similar meta-neural semantic representations reported more similar subjective ratings. Within-participant RSA provided a complementary test of the correspondence between representational and rating dissimilarity. Participant-level RSA correlations were higher for the proposed framework than for NeuroSTORM ($P < 0.0001$; Fig. 6n). Figs. 2 and 4 show corresponding results in additional rating tasks. These results demonstrate that the similarity structure of meta-neural semantic representations aligned with that of subjective evaluations both across and within participants.

Meta-neural semantic representations captured both internally generated cognitive content and distinct participant-specific subjective states elicited by the same stimuli. These findings extend neural decoding to cognitive content generated internally and shaped by individual subjective interpretation, beyond what external stimuli alone specify. Within the framework, such cognitive states were inferred from neural observations under brain-intrinsic priors, rather than identified through a direct mapping between stimuli and neural responses.

Discussion

Changes in neural activity patterns do not imply changes in cognitive state². A key question in neural decoding is therefore how to resolve the cognitive states that remain stable in the brain despite continuously changing neural population activity^{4,5}. Existing neural decoding methods, which are largely grounded in the stimulus-response principle, treat predefined external labels as decoding targets and learn statistical correspondences between neural observations and these labels^{7,8}. When neural activity changes continuously, reliance on such statistical correspondences tends to map changes in neural observations onto changes in cognitive state. These methods thus have difficulty resolving stable cognitive states from changing neural observations. Here, we redefine neural decoding as the process of inferring the cognitive state of the brain from neural observations, and we propose a Bayesian brain hierarchical decoding framework as a computational implementation of this inference. The proposed framework recovers relatively stable cognitive states from

changing neural observations, thereby distinguishing changes in neural observations from changes in cognitive states. Accordingly, the framework not only extends the objective of neural decoding but also provides an analytical tool for examining the relationship between changes in neural activity patterns and the stability of cognitive states.

From a neural manifold perspective, cognitive states are embedded within the manifold structure formed by whole-brain neural population activity². Existing neural recording technologies capture only partial observations of this whole-brain activity⁶. Neural observations are therefore projections of whole-brain neural population activity onto a specific observation space, and they reveal the relationships between cognitive states only incompletely. Our Bayesian brain hierarchical decoding framework treats limited neural observations as evidence and constrains cognitive inference with brain-intrinsic priors. This inference yields meta-neural semantic representations that carry the semantic content of cognitive states. Representational geometry analyses further revealed that cognitive inference was accompanied by a systematic geometric reorganization of the meta-neural semantic representation manifolds (Fig. 3a–b; Extended Data Fig. 3). Cognitive inference altered the relative configuration of the manifolds corresponding to different cognitive states in representational space, such that geometric relationships susceptible to variation in finite neural observations became more stable in the high-level representation. Consequently, the high-level representation preserved the relative relationships among cognitive states even as neural observations changed. At the level of

representational geometry, cognitive stability may therefore be understood as the preservation of the relational structure among cognitive states.

Stable cross-session neural decoding does not require neural activity patterns to remain unchanged^{2,38,39}. Neural drift substantially alters the neural observations associated with a given cognitive state across sessions, even when the cognitive state itself remains stable. The free-energy principle provides a theoretical perspective on how changes in neural activity can coexist with cognitive stability. Within this framework, free-energy minimization characterizes how the brain maintains relative stability in a changing environment without requiring fixed patterns of neural activity¹³. From this perspective, neural drift can be understood as cross-session variation in neural activity that is compatible with stable cognitive states, rather than as evidence that the cognitive states themselves have changed. Our results are consistent with this view. Despite neural drift over days to months, cognitive inference with our method yielded meta-neural semantic representations whose representational geometry remained consistent across sessions (Fig. 4). On this basis, the proposed framework maintained stable decoding performance across sessions. Together, these results indicate that stable cross-session neural decoding need not rely on aligning neural observations across sessions, but can instead be achieved by inferring stable cognitive states from changing neural observations.

The cognitive state of the brain is not uniquely determined by external stimuli. Cognitive states

also arise from internal processes such as imagination and spontaneous thought, and the same external stimulus gives rise to different subjective interpretations. Decoding based on the stimulus-response principle takes predefined external labels as targets and thereby prescribes stimulus content as the decoding outcome. Such decoding mainly reproduces stimulus information rather than inferring the cognitive state of the brain. In contrast, results from the internal mentation experiments (Fig. 6) showed that the meta-neural semantic representations inferred by the proposed framework are not restricted to stimulus content but also reflect internally generated cognitive states and subjective interpretations. Our method therefore extends the scope of neural decoding to internal cognitive processes, including memory, imagination, judgment and subjective interpretation, and provides a computational approach for investigating the cognitive states generated by such processes. This extension broadens the range of information accessible to human-machine interaction and provides a computational foundation for cognitive brain-computer interfaces⁴⁰.

Human brain structure exhibits pronounced inter-individual variation⁴¹. Because brain-intrinsic priors are shaped by brain structure, this variation implies that cognitive inference in different individuals is constrained by different priors. From this perspective, individual differences in brain structure become a source of individuality in cognitive inference itself. The individualized cortical geometric priors in the proposed framework capture this individuality, and the decoding improvement they bring supports

this view (Extended Data Fig. 6). At the same time, the cognitive states of different individuals are inferred through the same process, and individual differences enter this process through the constraints imposed by each brain’s prior. Our method thereby achieves personalized decoding under a common inference principle, constraining cognitive inference with each brain’s own structure to recover stable cognitive states from individual neural activity.

In this study, we incorporated constraints imposed by brain structure on neural activity into cognitive inference as brain-intrinsic priors, and instantiated these structural priors using cortical geometric eigenmodes^{16,18}. Future studies could also consider priors defined by neural dynamics and examine how different brain-intrinsic priors shape cognitive stability^{42,43}. For example, temporal priors derived from neural dynamics could allow cognitive inference to directly exploit regularities in the evolution of cognitive states, enabling their formation to be tracked over longer timescales. Overall, our experimental results indicate that, in a continuously changing external environment, the brain need not maintain fixed patterns of neural activity to sustain relatively stable cognitive states, but can instead achieve stability through cognitive inference from changing neural activity.

Methods

Bayesian brain hierarchical decoding framework

Bayesian brain theory motivates hierarchical inference from current neural evidence under constraints imposed by brain-intrinsic priors (Fig. 1). We denote the neural observation representation by o , the prior-constrained representation by u , and the inferred high-level meta-neural semantic representation by z . The inference is formulated as

$$p(z \mid o, u) \propto p(o \mid z) p(z \mid u). \quad (1)$$

Here, $p(o \mid z)$ quantifies the compatibility of a candidate representation z with the current neural observation representation, whereas $p(z \mid u)$ constrains the candidate representation according to the brain-intrinsic prior. Their product defines the inference over z .

Neural observation encoder. Modality-specific pretrained encoders transformed neural recordings from EEG, MEG, ECoG, fNIRS and fMRI into the neural observation representation o . Grouped spatial-coordinate embeddings and factorized attention accommodated differences in recording length and channel configuration. The resulting o represented the evidence provided by the current neural observation.

Brain-intrinsic prior encoder. The prior encoder incorporated brain-intrinsic prior information into the current neural representation to produce the prior-constrained representation u . In the present implementation, the fixed cortical geometric eigenmode basis instantiated the brain-intrinsic

structural prior, and u represented the current neural observation under this structural constraint. Their projection coefficients quantified the expression of the current neural observation within the fixed eigenmode basis, whereas the corresponding eigenvalues indexed spatial scale. These quantities were encoded across spatial scale and time to obtain u , which imposed the brain-intrinsic prior constraint during inference of z . We denote the eigenmode projection operator by P_Ψ , such that $\mathbf{a} = P_\Psi(o)$ gives the eigenmode coefficient vector associated with neural observation representation o .

Inference of meta-neural semantic representations. To infer the meta-neural semantic representation z , we formulated inference under two complementary constraints derived within each recording modality from the same neural observation. The neural observation representation o retained evidence from the current recording, whereas the prior-constrained representation u encoded brain-intrinsic structural constraints. We implemented this Bayesian formulation using a temperature-scaled product of experts⁴⁴:

$$p_T(z \mid o, u) \propto p(o \mid z)^{1/T_o} p(z \mid u)^{1/T_u}. \quad (2)$$

The positive temperatures T_o and T_u controlled the relative contributions of the observation and prior-constrained factors, respectively. At the representation level, the two factors were implemented as isotropic Gaussian experts with a common scale

and means centered at o and u , respectively, yielding the normalized inverse-temperature weights

$$\omega_o = \frac{T_o^{-1}}{T_o^{-1} + T_u^{-1}}, \quad \omega_u = \frac{T_u^{-1}}{T_o^{-1} + T_u^{-1}}, \quad (3)$$

and the inferred representation

$$z = \omega_o o + \omega_u u. \quad (4)$$

The resulting z constituted the meta-neural semantic representation used for subsequent decoding.

Masked-autoencoding pretraining. The neural observation encoder and brain-intrinsic prior encoder were pretrained independently as masked autoencoders, without access to task labels. Fifty percent of the temporal and spatial patches were randomly masked. A reconstruction decoder predicted the masked patches from the corresponding encoder representation, with reconstruction loss evaluated only over the masked index set Ω :

$$\mathcal{L}_{\text{MAE}} = \frac{1}{|\Omega|} \sum_{p \in \Omega} \|\hat{x}_p - x_p\|_2^2. \quad (5)$$

Here, x_p denotes the target patch at masked position p , and $\hat{x}_p$ denotes its reconstruction. Pretraining was performed independently for each neural recording modality using only unlabeled neural recordings from the corresponding training partition.

Cortical geometric eigenmodes

The current implementation instantiates the brain-intrinsic prior through intrinsic cortical geometry,

represented by a fixed eigenmode basis. This geometry constrains the spatial propagation of large-scale neural activity. Its eigenfunctions form an orthogonal basis ordered by spatial scale and fixed across experimental tasks. Modality-specific mappings express this basis in each measurement space. Previous work has shown that these eigenmodes account for the large-scale organization of task activity and resting-state networks^{16,17}.

Cortical eigenmode construction. The Laplace–Beltrami operator Δ was defined on the midthickness triangular cortical surface. Its eigenfunctions satisfy

$$\Delta\phi_k = -\lambda_k\phi_k. \quad (6)$$

The non-negative eigenvalues were ordered from low to high. Smaller eigenvalues correspond to larger spatial scales. The constant zeroth mode was removed. The remaining K modes were assembled as $\Phi = [\phi_1, \dots, \phi_K]$.

Modality-specific eigenmode mapping. Cortical eigenmodes were used as a common geometric basis across neural recording modalities. Because different modalities sample neural activity in distinct observation spaces, the cortical eigenmode basis was mapped to each modality according to its spatial sampling geometry, and the resulting modality-specific basis was used to obtain eigenmode coefficient time series from the observed neural signals.

The fMRI eigenproblem was solved separately on both hemispheres. The analysis used the standard 32k_fs_LR midthickness surface from the Human Connectome Project⁴⁵. Each hemisphere

contains 32,492 vertices. The cortical atlas region of interest mask removed the medial wall. A finite-element solver computed the eigenmodes. The analysis retained 1,000 modes from each hemisphere, giving 2,000 modes in total. Cortical time series were projected by least squares to obtain eigenmode coefficient time courses.

For EEG and MEG, dipole sources were constrained to a common cortical source space, and a three-layer boundary-element model was used to compute the forward solution for each modality. For each modality, the resulting lead-field matrix mapped the cortical basis to sensor space.

The forward operator mixes cortical modes into correlated sensor patterns. Singular value decomposition was used to process the mapped basis, and selected components formed a sensor-space basis for estimating coefficient time series. Cortical spatial-scale information was incorporated into the sensor-space representation.

For ECoG, the cortical eigenmodes were evaluated at the contact locations. Dataset-specific contact coordinates were mapped to the standard surface. Contact j was assigned to its nearest cortical vertex v_j , giving $\tilde{\Phi}_{jk} = \phi_k(v_j)$. Tikhonov-regularized least squares estimated the coefficients from the observed contact signals.

For fNIRS, source-detector midpoints projected the surface eigenmodes into channel space. The components were orthogonalized before separate projection of HbO and HbR.

Participant-specific eigenmodes were constructed analogously using each participant’s cortical surface. Individual surfaces were registered to the standard 32k_fs_LR mesh to maintain vertex

correspondence. The resulting participant-specific basis replaced the standard cortical basis in the modality-specific mapping procedures described above, with all subsequent steps unchanged.

Decoding and encoding with meta-neural semantic representations

Decoding from meta-neural semantic representations. A task-specific linear classification head was applied to the inferred meta-neural semantic representation. For trial i , its categorical estimate was

$$q_i(c \mid z_i) = \text{softmax}(W_{\text{cls}}z_i + b_{\text{cls}})_c. \quad (7)$$

For target class c_i , the classification head was fitted using cross-entropy:

$$\mathcal{L}_D = \frac{1}{N} \sum_{i=1}^N \text{CE}(c_i, q_i(\cdot \mid z_i)). \quad (8)$$

Here, W_{cls} and b_{cls} are the classification weights and bias, respectively, N is the number of training trials and CE denotes cross-entropy. All discriminative accuracy estimates were computed from $q_i(c \mid z_i)$. The pretrained encoders and temperature parameters remained fixed, and only W_{cls} and b_{cls} were fitted on the corresponding training partition.

Open-ended language decoding comprised semantic alignment followed by conditional generation (Fig. 5). BGE-M3 produced token-level text states that were resampled to a fixed-length semantic sequence³⁰. A query-based cross-attention transformer (Q-Former) aligned the meta-neural semantic representation with this sequence⁴⁶. Let $\bar{\kappa}$ denote their mean tokenwise cosine similarity.

The objective combined reconstruction, cosine-distance and symmetric within-batch contrastive losses:

$$\mathcal{L}_A = \alpha_1 \mathcal{L}_2 + \alpha_2 (1 - \bar{\kappa}) + \alpha_3 \mathcal{L}_{\text{con}}. \quad (9)$$

The coefficients α_1 , α_2 and α_3 are loss weights. The term $\mathcal{L}_2$ denotes reconstruction loss, and $\mathcal{L}_{\text{con}}$ denotes the mean of the two cross-modal retrieval cross-entropies computed between mean-pooled neural and text representations within each batch. Parameters of the alignment model and language model were shared across participants. The same story-level 9:1 split was applied to every participant. Semantic alignment was fitted before the language adapters, with both stages using the same training stories. Test stories were excluded from all model-fitting stages and used only for final generation and evaluation. The aligned sequence was projected into the hidden space of the 3.8-billion-parameter Phi-4-mini-instruct model and provided as a soft prompt⁴⁷. PiSSA initialized the LoRA adapters on the attention projections⁴⁸. Causal language-model loss was evaluated only at answer tokens. Beam-search decoding included penalties for sequence repetition and repeated n -grams. Language-specific text normalization was applied to generated sequences before evaluation.

SD Image Variations v2.0 provided normalized CLIP ViT-L/14 image embeddings⁴⁹. The same cross-attention alignment model mapped the meta-neural semantic representation into this conditioning space and was fitted using the reconstruction, cosine-distance and symmetric within-batch

contrastive losses defined above. A frozen latent-diffusion decoder generated images from Gaussian latent noise⁵⁰. Sampling used DDIM and classifier-free guidance⁵¹. Diffusion-model parameters remained fixed. Sampling settings and random seeds were matched within each comparison. NOD trials were used exclusively for visual pretraining. All image-decoding evaluations were conducted within individual NSD participants. Participant-specific models were fitted on each participant’s NSD-core training data. Held-out NSD-core trials were used to evaluate viewed-image reconstruction, and the NSD imagery condition was used to evaluate mental-imagery reconstruction.

Encoding brain activity through meta-neural semantic representations. Let $e_i \in \mathbb{R}^d$ denote the external semantic representation of stimulus i and $b_i \in \mathbb{R}^J$ its cortical GLM beta pattern. A semantic bridge G produced $z_i^{\text{sem}} = G(e_i)$. For training stimuli, the frozen hierarchical model determined the target representation z_i from b_i and its projection constrained by brain-intrinsic priors. Predictions for independent test stimuli were based exclusively on e_i .

The target response was decomposed as $b_i = \Phi a_i + \epsilon_i$, where $a_i = \Phi^+ b_i$ and $\epsilon_i = b_i - \Phi a_i$. Two decoders predicted these components from z_i^{sem} , and the predicted components were combined using positive weights:

$$\hat{b}_i = \gamma_{\text{mode}} \Phi D_{\text{mode}}(z_i^{\text{sem}}) + \gamma_{\text{res}} D_{\text{res}}(z_i^{\text{sem}}). \quad (10)$$

Both γ_{mode} and γ_{res} were constrained to be positive. The decoders D_{mode} and D_{res} predicted a_i

and ϵ_i , respectively. The encoding objective was

$$\begin{aligned} \mathcal{L}_E = & \beta_0 [1 - \cos(z_i^{\text{sem}}, z_i)] \\ & + \beta_1 \|D_{\text{mode}}(z_i^{\text{sem}}) - a_i\|_2^2 \\ & + \beta_2 \|D_{\text{res}}(z_i^{\text{sem}}) - \epsilon_i\|_2^2 \\ & + \beta_3 \left[1 - \text{corr}(\hat{b}_i, b_i)\right]. \end{aligned} \quad (11)$$

Here, $\cos(\cdot, \cdot)$ denotes cosine similarity, $\text{corr}(\cdot, \cdot)$ denotes Pearson correlation and $\beta_0, \dots, \beta_3$ are loss weights. Caption embeddings were the primary semantic input. Category embeddings and image features served as controls. The bridge, response decoders and positive weights were fitted only on training stimuli.

The stimulus-response encoding baseline directly mapped the external semantic representation e_i to the cortical response b_i using an MLP, without hierarchical inference or cortical-geometric prior constraints. It used the same response targets, training and test stimuli, data partition and optimization protocol as the cognitive-inferential encoding model.

Training and implementation details

The neural observation and brain-intrinsic prior encoders were independently pretrained as masked autoencoders with a reconstruction objective (masking ratio, 0.5) using AdamW (learning rate, 3×10^{-4}) for 60 epochs, with batch sizes of 32 and 64, respectively, and cosine learning-rate decay. The checkpoint with the lowest validation reconstruction loss was retained, after which both encoders were frozen. The positive temperatures T_o and T_u were calibrated separately

for each recording modality using the validation partition of the corresponding training data. They controlled the relative strength of the neural observation and prior-constrained factors in the temperature-scaled PoE inference. The temperature pair with the lowest validation loss was retained and fixed for all held-out evaluations. The outputs of the two encoders entered the temperature-scaled product-of-experts (PoE) inference to obtain the meta-neural semantic representation. For language decoding, this representation was aligned with BGE-M3 embeddings through subject-specific layers and a Q-Former using a weighted combination of mean-squared-error, token-wise cosine and contrastive losses for 200 epochs (batch size, 256; AdamW; learning rate, 3×10^{-4} ; cosine learning-rate decay). For text generation, we then jointly optimized a linear projector and PiSSA-LoRA adapters (rank, 16; dropout, 0.05) on frozen Phi-4-mini using cross-entropy with label smoothing of 0.1 for 15 epochs (batch size, 16; AdamW; learning rates, 3×10^{-4} for the projector and 5×10^{-5} for LoRA; cosine learning-rate decay).

For image decoding, the same representation was aligned with CLIP image embeddings through a visual alignment head using a weighted combination of mean-squared-error, cosine and contrastive losses for 200 epochs (batch size, 64; AdamW; learning rate, 2×10^{-4}). The aligned embeddings were decoded into images using a frozen diffusion generator without updating its parameters. All models were trained on two NVIDIA A800 GPUs with 80 GB memory each.

Datasets

The internally acquired Imagined Speech cohort comprised adults aged 25 to 29 years. Each trial consisted of 5 s of reading, 4 s of speech imagery and 4 s of fixation, with only the imagery epoch included in the main analysis. Neural activity was recorded using a 64-channel single-axis OPM-MEG system (Pyramag Epoch 64, Quanamag Healthcare) in a semi-enclosed magnetically shielded cylinder. A T1-weighted anatomical image was acquired for each participant using a 3.0 T uMR 790 scanner (United Imaging Healthcare). The protocol was approved by the Ethics Committee of the Shenzhen Institute of Advanced Technology (approval no. SIAT-IRB-251015-H1027), and all participants provided written informed consent. All public datasets used in this study, together with their cohort sizes, recording modalities and analysis assignments, are summarized in Extended Data Table 1. Analysis-specific sample sizes are reported in the corresponding figure legends.

Data preprocessing

T1-weighted images were processed with FreeSurfer through the ABCD-HCP workflow, which reconstructed native white and pial surfaces and derived the midthickness surface⁵². Each surface was registered and resampled to the standard 32k_fs_LR mesh. The participant-specific cortical region of interest mask was retained. Participant surfaces supported individual eigenmode construction. The group surface was used for datasets without participant anatomical images.

fMRI data were processed with the ABCD-HCP BIDS pipeline⁵², an extension of the Human

Connectome Project minimal workflow. Processing comprised denoising, spatial registration, cortical reconstruction and CIFTI generation. Motion and distortion correction followed the information supplied by each dataset. Datasets without field maps entered the workflow without fieldmap correction. Nuisance regression and motion censoring were performed before cortical projection. The final model input was the cortical CIFTI time series on the standard 32k_fs_LR mesh. Event windows followed the source annotations. A 6 s hemodynamic offset aligned task events to the blood oxygen level dependent response. Image-encoding analyses used GLM beta maps estimated from the preprocessed time series.

EEG datasets released only as preprocessed epochs were analyzed using the preprocessing provided with the original release. For datasets distributed as continuous recordings, signals were notch filtered at the dataset-specific mains frequency, band-pass filtered between 0.1 and 50 Hz, and resampled to 256 Hz. Independent component analysis was applied where continuous recordings and sufficient channel coverage were available to remove ocular, cardiac and stereotyped myogenic components. Task epochs were defined from the event annotations provided with each dataset. Electrode coordinates were retained for forward modeling.

MEG recordings were notch filtered at the dataset-specific mains frequency, band-pass filtered between 0.1 and 50 Hz, and resampled to 256 Hz. Sensor geometry was retained for forward modeling. For OPM-MEG, sensor positions and

orientations were co-registered to the individual anatomical image before source modeling.

fNIRS recordings were processed according to the form in which each dataset was released. Raw intensity data were screened for invalid measurements, converted to optical density, corrected for motion artifacts and band-pass filtered before conversion to hemoglobin concentration changes using the modified Beer-Lambert law. Trial-wise baseline correction was applied where required. Datasets released as processed hemoglobin time series retained the provided preprocessing. Channel correspondence was harmonized across runs before analysis.

ECoG recordings were screened for discontinuities and non-finite samples before filtering. Line noise at the dataset-specific mains frequency and its harmonics was removed, after which signals were band-pass filtered and resampled to 256 Hz. The reference scheme provided with the source release was retained where applicable. Electrode coordinates supplied with each dataset were used for cortical-surface sampling.

Model comparisons

EEG and MEG comparisons used LaBraM, BrainOmni and CBraMod with their released pre-trained weights^{23–25}. A task-specific linear head was attached to each backbone. Within-participant comparisons used identical dataset partitions. Zero-shot cross-participant training excluded every sample from the evaluated participant. Cross-participant evaluation froze each pretrained backbone and fitted only its classification head.

Cross-session ECoG and EEG comparisons also included SPaRCNet⁵³. Every method was trained on the same reference session and evaluated on the same subsequent-session trials. Internal-state fMRI analyses compared the proposed framework with NeuroSTORM³⁶.

Language decoding was compared with CogReader using the same story boundaries³¹. A random control replaced the neural soft prompt with independent standard-normal tokens. The frozen language model and decoding procedure remained unchanged. Image retrieval was compared with NCC on identical NSD candidate sets and test trials within each participant³³. The fNIRS Transformer comparison used the same within-participant fivefold partitions as the proposed framework⁵⁴.

Shared cohorts used identical trial definitions, label vocabularies and test samples across methods. Each released model retained its native backbone and input representation. Data partitions, optimization budgets and evaluation metrics were fixed before test evaluation.

Quantitative analyses

Evaluation protocol. Training and test data were separated at the analysis-specific grouping level. Within-participant classification used run-blocked folds for datasets with multiple runs and trial-blocked folds otherwise. Zero-shot cross-participant evaluation used disjoint participant sets. Language datasets used a 9:1 story-level split shared across participants, with nine parts for training and one for testing. Held-out stories were excluded from model fitting and used only for final

generation and scoring. Visual decoding analyses used separate stimuli for model fitting and evaluation. For cross-session evaluation, models were trained on the reference session and applied to all subsequent sessions without refitting.

Cortical semantic attribution. The semantic-attribution maps in Fig. 2e were computed on independent SMN4Lang test data. Attribution for the neural observation representation (low-level) was defined as the square root of non-negative cross-validated vertex-to-semantic-embedding R^2 after surface smoothing. Attribution for the meta-neural semantic representation (high-level) was obtained from cross-validated eigenmode-to-semantic-embedding R^2 , with squared eigenmode amplitudes used to project these values onto the cortical surface. The two attribution maps were independently rank-normalized for visualization. The displayed percentages indicate the proportion of cortical surface vertices represented in each attribution map.

Surface-searchlight RSA compared cortical coverage of low-level and high-level reference geometries in eight NSD participants, using 907–1,000 Shared1000 images per participant and repetition-averaged fMRI responses at 59,412 fsLR32k cortical vertices. Searchlights included vertices within 10 mm along template midthickness mesh edges, within the same hemisphere and excluding the medial wall. Local brain RDMs contained pairwise correlation distances ($1 - r$, Pearson correlation) between image-response patterns.

Reference RDMs used 768-dimensional CLIP ViT-L/14 embeddings or 1,024-dimensional meta-neural semantic representations predicted by a participant-specific image-to-representation bridge. Its 2,944 inputs combined CLIP ViT-L/14 embeddings, mean-pooled ViT-bigG/14 tokens and 512 training-fitted ConvNeXt principal components. A 2,944–512–1,024 network with GELU and dropout 0.1 predicted fused posterior means inferred from non-test fMRI. Training used mean-squared error, AdamW (learning rate 3×10^{-4} , weight decay 0.01, batch size 512) and a 90:10 fitting–validation split of non-shared images. Validation RDM correlation selected the checkpoint over at most 120 epochs, with checks every five epochs and stopping after 25 epochs without improvement. The configuration was selected on sub-02 validation data and fixed across participants; Shared1000 data were excluded from fitting and validation.

Searchlight RSA correlated the upper triangles of brain and reference RDMs using Spearman correlation. Exact two-sided sign-flip tests across participants (256 assignments) were followed by Benjamini–Hochberg correction jointly over both references and all vertices. Coverage comprised vertices with mean $\rho > 0.01$ and $q < 0.05$, weighted by template vertex area. Paired high-minus-low RSA differences were tested similarly, with correction across vertices as a separate family.

The contribution of each eigenmode band or latent token was quantified by the decrease in decoding performance after occlusion. For eigenmode band g and latent token m , the occlusion

effects were

$$\delta_g = s_{\text{full}} - s_g^{\text{occ}}, \quad \delta_m = s_{\text{full}} - s_m^{\text{occ}}, \quad (12)$$

where s_{full} denotes the score obtained with the complete representation, and s_g^{occ} and s_m^{occ} denote the corresponding scores after occluding eigenmode band g and latent token m , respectively. The score was BGEScore for language decoding and CLIP cosine similarity for visual decoding. Modes were grouped into non-overlapping bands in ascending eigenvalue order. Their representative wavelength was

$$\ell_g = \frac{2\pi}{\sqrt{\bar{\lambda}_g}}. \quad (13)$$

The value $\bar{\lambda}_g$ is the mean eigenvalue in band g . The final spatial-attention layer supplied the normalized token-to-mode weight A_{mk} . This weight was projected through the cortical basis as $\tilde{A}_m(v) = \sum_k A_{mk} \phi_k(v)$. The participant-level cortical attribution map was

$$\mathcal{M}(v) = \sum_m \max(\delta_m, 0) |\tilde{A}_m(v)|. \quad (14)$$

The resulting attribution map was non-negative by construction and was normalized before group aggregation.

Cortical attribution maps were averaged within the Schaefer atlas⁵⁵. Parcels were assigned to the seven Yeo networks⁵⁶. Inspired by activity-flow mapping⁵⁷, we estimated participant-specific functional-connectivity weights from SMN4Lang task-fMRI using pairwise Pearson correlations between parcel time series. Each attribution map

was standardized across parcels, and the diagonal of the resulting connectivity matrix W was set to zero. For target parcel j , connectivity-weighted semantic attribution was

$$\hat{s}_j = \sum_{\ell \neq j} W_{\ell j} s_\ell. \quad (15)$$

Here, s_ℓ denotes semantic attribution at source parcel ℓ . The semantic-connectivity alignment coefficient was

$$\rho = \text{corr}_j(\hat{s}_j, s_j). \quad (16)$$

Here, corr_j denotes Pearson correlation across target parcels j . The same calculation was repeated within each Yeo network. For Fig. 2g, alignment coefficients for the proposed framework and CogReader were compared descriptively. For Fig. 2h, interregional relations were displayed for both methods. Cross-network relations joined parcels assigned to different Yeo networks. Relations were classified as long-distance when their parcel-centroid geodesic distance exceeded the 75th percentile.

Representational geometry analyses.

Analyses crossed two cognitive states with two cognitive conditions: rest versus right-hand imagery with versus without EEG neurofeedback; inanimate versus animate images across small versus large real-world sizes; and low versus high engagement across low versus high sadness, defined by participant-specific rating medians.

Three-dimensional MDS visualizations used centroids of the four state-condition combinations from centered, isotropically scaled representations. Projection onto an orthonormal basis spanning

their state- and condition-difference vectors preserved pairwise Euclidean distances; coordinates were normalized by mean inter-centroid distance. Each representation was embedded separately, with quantitative metrics computed in full feature space.

CCGP used L_2 -regularized logistic regression ($C = 0.01$) to classify states in one cognitive condition after training in the other. Feature standardization was fitted on the training condition. Accuracy was averaged over both transfer directions and 20 resamples, balancing all four state-condition combinations without replacement. PS was the cosine similarity between the two condition-specific state-difference vectors, each obtained by subtracting one state centroid from the other in a consistent order. High-minus-low differences in participant-level CCGP and PS were tested using exact two-sided paired sign flips of the mean difference.

Representational structure and cross-session stability. Frozen representations were standardized before t-SNE. Principal-component reduction was performed before the two-dimensional embedding. For Fig. 4h-i, t-SNE was performed with 50 random initializations. A silhouette coefficient was calculated for each trial and averaged across trials. For Fig. 4i, the same coefficient was calculated after balanced sampling of ten trials per class within each ECoG Speech session. Linear-probe accuracy was used to quantify class separability.

Cosine similarity measured direct alignment between neural and semantic features. Representational similarity analysis used $d_{ij} = 1 - \text{corr}(h_i, h_j)$, where h_i is the representation of item i and

corr denotes Pearson correlation. The RSA coefficient was the Spearman correlation between the upper triangles of two dissimilarity matrices. Group analyses used Fisher-transformed coefficients. Normalized representational-amplitude profiles, cross-session similarity matrices and population-vector displacement quantified cross-session stability. ECoG Speech models were trained on the reference session and applied to every follow-up session. AJILE12 ECoG and SHU-MI used the same reference-session protocol.

Internal cognitive state analyses. Imagined Speech analyses used only the 4 s OPM-MEG imagery epoch. Semantic correlation was quantified using category-level RSA. Trial-level neural representations were averaged within each category to obtain neural centroids. BGE embeddings of the corresponding sentences were averaged within the same categories to obtain semantic centroids. Pairwise distances were calculated separately among neural and semantic centroids, and the Spearman correlation between the corresponding upper-triangular distance values yielded the RSA coefficient. A classifier fitted to imagery epochs yielded content accuracy. DuPre2016 and Lee2021 used their released task labels for self-generated condition decoding. Their trial-level clarity and vividness scores supplied continuous targets. Valence supplied a further continuous target.

Spacetop Alignvideo analyses in Fig. 6h–j included participants with complete ratings across all five dimensions for the common set of videos. A separate regression model was fitted for each dimension. Direct performance was the within-participant correlation between predicted and

observed ratings. Video-specific consensus was estimated from training participants and subtracted from both the test-participant rating and the model prediction. The correlation between these residuals quantified participant-specific interpretation. Participant-level correlations were Fisher transformed for inference.

The RSA analyses in Fig. 6m–n used participants with complete ratings across all five dimensions for the same set of videos. The dimensions were happy, sad, afraid, disgusted and engaged. The item index i identifies a video in this analysis. Rating vectors were standardized using parameters estimated from the training participants and are denoted $\tilde{y}_{si}$. The analyses compared time-pooled representations from NeuroSTORM and the proposed framework.

For the between-participant analysis, each method yielded a representation matrix for each video, with one row per participant. Neural distance between participants s_1 and s_2 for video i was

$$d_i^{(z)}(s_1, s_2) = 1 - \text{corr}(z_{s_1 i}, z_{s_2 i}). \quad (17)$$

Rating distance for the same participant pair was

$$d_i^{(y)}(s_1, s_2) = \|\tilde{y}_{s_1 i} - \tilde{y}_{s_2 i}\|_2. \quad (18)$$

Each video yielded participant-pair distances of each type. Fractional ranks were calculated over participant pairs within each video and averaged over videos:

$$\overline{\mathcal{R}}^{(\xi)}(s_1, s_2) = \frac{1}{N} \sum_{i=1}^N \mathcal{R}_i(d_i^{(\xi)}(s_1, s_2)), \quad (19)$$

$$\xi \in \{z, y\}.$$

Here, N denotes the number of videos in this analysis. The operator $\mathcal{R}_i$ maps the participant-pair distances for video i to fractional ranks between zero and one. The reported coefficient was the Pearson correlation between the upper-triangular entries of the two participant-pair fractional-rank matrices. This analysis quantified the association between neural representational distance and rating distance.

For the within-participant analysis, each participant yielded neural and rating dissimilarity matrices across videos. Their upper-triangular entries yielded one Pearson coefficient r_s . Coefficients were Fisher transformed as $\zeta_s = \text{arctanh}(r_s)$. The transformed values were averaged across participants. The hyperbolic tangent returned the group coefficient to the correlation scale. The participant was the analysis unit. Group differences used a two-tailed paired t -test on Fisher-transformed values.

For both RSA analyses, scaling parameters and model selection were estimated from the training data and held fixed during evaluation. Reported coefficients were then calculated on the corresponding held-out observations.

For Extended Data Fig. 4, Spacetop participants who had ratings for both tasks were included. The tasks were Narratives and Shortvideo. RSA compared meta-neural semantic dissimilarity with rating dissimilarity across the shared stimuli. The participant was the unit of analysis.

Structural controls and robustness. Equal-dimensional Fourier and spherical-harmonic bases served as structural controls. Graph-Laplacian and

cortical-eigenmode bases completed the comparison. Every basis used the same encoder and classifier. Data partitions were also fixed. Mode-number analyses changed only the number of retained basis functions. Group-template and participant-specific bases were evaluated on identical trials.

Robustness was assessed by adding zero-mean Gaussian noise scaled by the training-fold channel standard deviation, masking nested proportions of channels and restricting the training set to nested fractions of trials. At each perturbation level, BrainOmni and the proposed model received the same perturbation and sample subset. A separate cross-participant analysis progressively added training data from the target participant while holding the test set fixed. These analyses are reported in Extended Data Fig. 8.

Performance metrics and statistical analysis. Discriminative decoding was summarized by accuracy. Language generation was evaluated using sentence-level BGE-M3 embedding cosine similarity³⁰, termed BGEScore. Image reconstruction used CLIP image-embedding cosine similarity. Retrieval reported top-1, top-5 and top-10 accuracy in a 100-way task. For each test image, the candidate set comprised the target image and 99 randomly selected non-target images from the 1,000-image test set.

Semantic encoding of brain activity was evaluated on independent test stimuli. Vertex-wise prediction correlation was $r_v = \text{corr}(\{\hat{b}_{i,v}\}_i, \{b_{i,v}\}_i)$. Let n_v denote the repeated-response noise ceiling at vertex v . The noise-ceiling-normalized squared

prediction correlation was

$$r_{\text{norm},v}^2 = \frac{r_v^2}{n_v}. \quad (20)$$

This measure uses squared test-set Pearson correlation, rather than residual-sum-of-squares R^2 . Vertices with $n_v \leq 0$ were excluded. Encoding coverage was the proportion of vertices with positive prediction correlation and a Benjamini–Hochberg adjusted $P < 0.05$. Vertex-wise null distributions used 10^4 permutations of stimulus identity. Semantic-attribution maps and network edges used the same number of permutations. Spatial correspondence, network concentration and coverage differences used 10^4 spherical rotations. Rotations were performed separately in each hemisphere and preserved the medial wall. Empirical probabilities used the standard add-one correction.

All hypothesis tests were two-sided. Comparisons among matched methods or structural bases used one-way repeated-measures ANOVA. Each participant provided one matched row of measurements. Analyses involving method and a second repeated factor used two-way repeated-measures ANOVA. Sphericity was not assumed, and the Greenhouse–Geisser correction was applied where appropriate. Pairwise contrasts used Šidák-adjusted multiple comparisons. Direct comparisons between two matched conditions used two-tailed paired t -tests. Comparisons between language-decoding outputs and the independent Random control used two-tailed Welch’s t -tests.

Line plots show the mean and s.e.m. when specified. The participant was the biological

replicate. Session-level analyses treated the session as the experimental unit only when explicitly stated. Repeated t-SNE initializations and matched random-seed model fits quantified optimization stability. Inferential tests used the participant, session, initialization or random-seed sampling units specified in the corresponding figure legends.

References

- [1] Gonzalez, W. G., Zhang, H., Harutyunyan, A. & Lois, C. Persistence of neuronal representations through time and damage in the hippocampus. *Science* **365**, 821–825 (2019).
- [2] Gallego, J. A., Perich, M. G., Chowdhury, R. H., Solla, S. A. & Miller, L. E. Long-term stability of cortical population dynamics underlying consistent behavior. *Nat. Neurosci.* **23**, 260–270 (2020).
- [3] Tyree, T. J., Metke, M. & Miller, C. T. Cross-modal representation of identity in the primate hippocampus. *Science* **382**, 417–423 (2023).
- [4] Haynes, J.-D. & Rees, G. Decoding mental states from brain activity in humans. *Nat. Rev. Neurosci.* **7**, 523–534 (2006).
- [5] Poldrack, R. A. Inferring mental states from neuroimaging data: From reverse inference to large-scale decoding. *Neuron* **72**, 692–697 (2011).

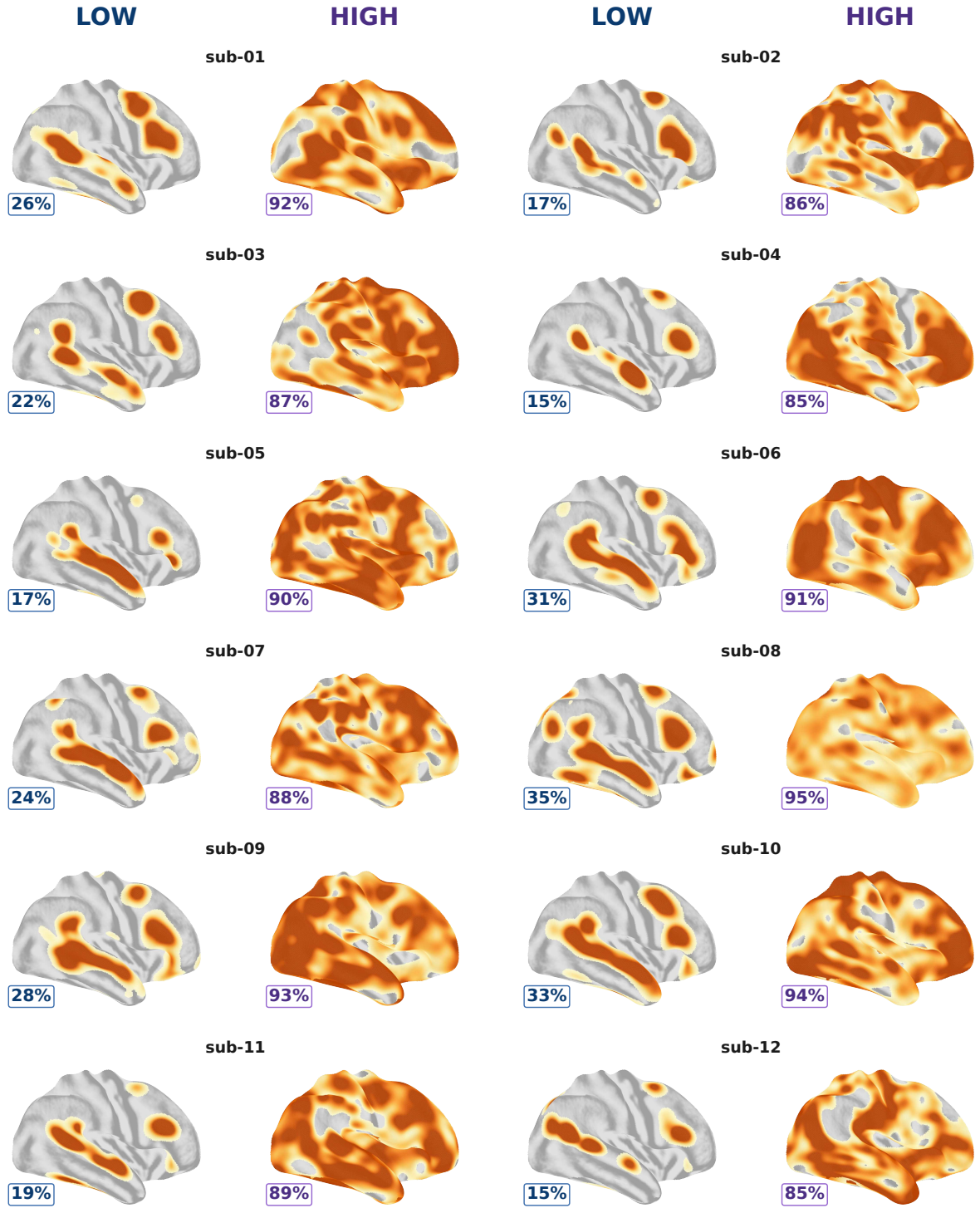
- [6] Kragel, P. A., Koban, L., Barrett, L. F. & Wager, T. D. Representation, pattern information, and brain signatures: from neurons to neuroimaging. *Neuron* **99**, 257–273 (2018).
- [7] Mathis, M. W., Perez Rotondo, A., Chang, E. F., Tolias, A. S. & Mathis, A. Decoding the brain: From neural representations to mechanistic models. *Cell* **187**, 5814–5832 (2024).
- [8] Peelen, M. V. & Downing, P. E. Testing cognitive theories with multivariate pattern analysis of neuroimaging data. *Nat. Hum. Behav.* **7**, 1430–1441 (2023).
- [9] Dijksterhuis, D. E. *et al.* Pronouns reactivate conceptual representations in human hippocampal neurons. *Science* **385**, 1478–1484 (2024).
- [10] Liu, C., Todorova, R., Tang, W., Oliva, A. & Fernandez-Ruiz, A. Associative and predictive hippocampal codes support memory-guided behaviors. *Science* **382**, eadi8237 (2023).
- [11] Knill, D. C. & Pouget, A. The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends Neurosci.* **27**, 712–719 (2004).
- [12] Körding, K. P. & Wolpert, D. M. Bayesian integration in sensorimotor learning. *Nature* **427**, 244–247 (2004).
- [13] Friston, K. The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* **11**, 127–138 (2010).
- [14] Berkes, P., Orbán, G., Lengyel, M. & Fiser, J. Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. *Science* **331**, 83–87 (2011).
- [15] Lange, R. D., Shivkumar, S., Chatteraj, A. & Haefner, R. M. Bayesian encoding and decoding as distinct perspectives on neural coding. *Nat. Neurosci.* **26**, 2063–2072 (2023).
- [16] Pang, J. C. *et al.* Geometric constraints on human brain function. *Nature* **618**, 566–574 (2023).
- [17] Normand, F. *et al.* Geometric constraints on the architecture of mammalian cortical connectomes. *Cell* **189**, 5283–5303.e17 (2026).
- [18] Atasoy, S., Donnelly, I. & Pearson, J. Human brain networks function in connectome-specific harmonic waves. *Nat. Commun.* **7**, 10340 (2016).
- [19] Brunner, C., Leeb, R., Müller-Putz, G. R., Schlögl, A. & Pfurtscheller, G. BCI competition 2008—Graz data set A (2008). https://www.bbci.de/competition/iv/desc_2a.pdf.
- [20] Chen, J. *et al.* A large finer-grained affective computing EEG dataset. *Sci. Data* **10**, 740 (2023).
- [21] Liu, W., Qiu, J.-L., Zheng, W.-L. & Lu, B.-L. Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition. *IEEE Trans. Cogn. Dev. Syst.* **14**, 715–729 (2022).

- [22] Lioi, G. *et al.* Simultaneous EEG-fMRI during a neurofeedback task, a brain imaging dataset for multimodal data integration. *Sci. Data* **7**, 173 (2020).
- [23] Jiang, W.-B., Zhao, L.-M. & Lu, B.-L. Large brain model for learning generic representations with tremendous EEG data in BCI. *The Twelfth International Conference on Learning Representations* 35429–35450 (2024). URL <https://openreview.net/forum?id=QzTpTRVtrP>.
- [24] Xiao, Q. *et al.* BrainOmni: A brain foundation model for unified EEG and MEG signals. *Advances in Neural Information Processing Systems* **38**, 46081–46114 (2025).
- [25] Wang, J. *et al.* CBraMod: A criss-cross brain foundation model for EEG decoding. *The Thirteenth International Conference on Learning Representations* 62056–62092 (2025). URL <https://openreview.net/forum?id=NPNUHgHF2w>.
- [26] Bhattasali, S., Brennan, J., Luh, W.-M., Franzuebbers, B. & Hale, J. The Alice Datasets: fMRI & EEG observations of natural language comprehension. *Proceedings of the Twelfth Language Resources and Evaluation Conference* 120–125 (2020). URL <https://aclanthology.org/2020.lrec-1.15/>.
- [27] Li, J. *et al.* Le Petit Prince multilingual naturalistic fMRI corpus. *Sci. Data* **9**, 530 (2022).
- [28] Hanke, M. *et al.* A high-resolution 7-Tesla fMRI dataset from complex natural stimulation with an audio movie. *Sci. Data* **1**, 140003 (2014).
- [29] Momenian, M. *et al.* Le Petit Prince Hong Kong (LPPHK): Naturalistic fMRI and EEG data from older Cantonese speakers. *Sci. Data* **11**, 992 (2024).
- [30] Chen, J. *et al.* M3-Embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Findings of the Association for Computational Linguistics: ACL 2024* 2318–2335 (2024). URL <https://aclanthology.org/2024.findings-acl.137/>.
- [31] Lu, W. *et al.* Brain-inspired fMRI-to-text decoding via incremental and wrap-up language modeling. *Advances in Neural Information Processing Systems* **38**, 166540–166563 (2025).
- [32] Allen, E. J. *et al.* A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nat. Neurosci.* **25**, 116–126 (2022).
- [33] Wang, H. *et al.* Inter-individual and inter-site neural code conversion without shared stimuli. *Nat. Comput. Sci.* **5**, 534–546 (2025).
- [34] DuPre, E., Luh, W.-M. & Spreng, R. N. Multi-echo fMRI replication sample of autobiographical memory, prospection and theory of mind reasoning tasks. *Sci. Data* **3**, 160116 (2016).

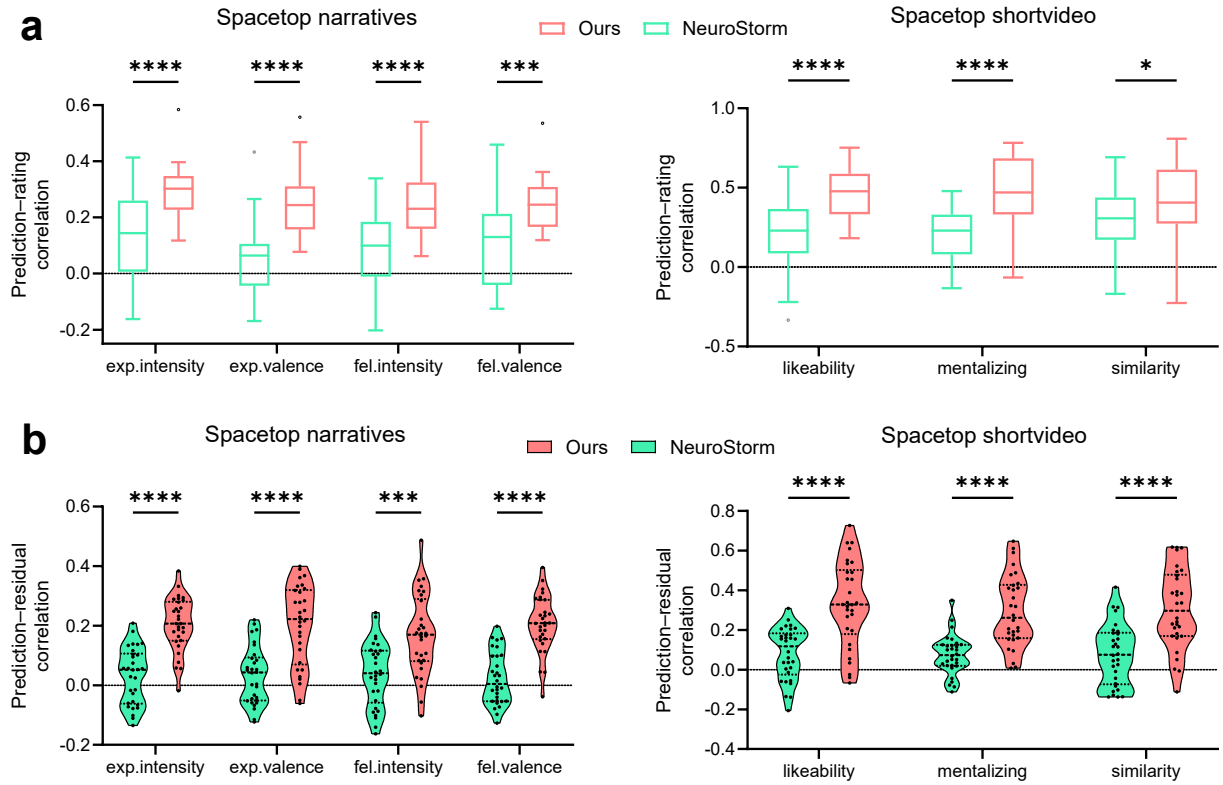
- [35] Lee, S., Parthasarathi, T. & Kable, J. W. The ventral and dorsal default mode networks are dissociably modulated by the vividness and valence of imagined events. *J. Neurosci.* **41**, 5243–5250 (2021).
- [36] Wang, C. *et al.* Towards a general-purpose foundation model for functional MRI analysis. *Nat. Biomed. Eng.* (2026). <https://doi.org/10.1038/s41551-026-01666-y>.
- [37] Jung, H. *et al.* Spacetop: A multimodal fMRI dataset unifying naturalistic processes with a rich array of experimental tasks. *Sci. Data* **12**, 1465 (2025).
- [38] Degenhart, A. D. *et al.* Stabilization of a brain–computer interface via the alignment of low-dimensional spaces of neural activity. *Nat. Biomed. Eng.* **4**, 672–685 (2020).
- [39] Sussillo, D., Stavisky, S. D., Kao, J. C., Ryu, S. I. & Shenoy, K. V. Making brain–machine interfaces robust to future neural variability. *Nat. Commun.* **7**, 13749 (2016).
- [40] Saez, I. The emerging field of cognitive brain–computer interfaces. *Trends Cogn. Sci.* (2026). <https://doi.org/10.1016/j.tics.2026.06.012>.
- [41] Bethlehem, R. A. *et al.* Brain charts for the human lifespan. *Nature* **604**, 525–533 (2022).
- [42] Breakspear, M. Dynamic models of large-scale brain activity. *Nat. Neurosci.* **20**, 340–352 (2017).
- [43] Demirtaş, M. *et al.* Hierarchical heterogeneity across human cortex shapes large-scale neural dynamics. *Neuron* **101**, 1181–1194.e13 (2019).
- [44] Hinton, G. E. Training products of experts by minimizing contrastive divergence. *Neural Computation* **14**, 1771–1800 (2002).
- [45] Glasser, M. F. *et al.* The minimal preprocessing pipelines for the human connectome project. *NeuroImage* **80**, 105–124 (2013).
- [46] Li, J., Li, D., Savarese, S. & Hoi, S. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *Proceedings of the 40th International Conference on Machine Learning* **202**, 19730–19742 (2023). URL <https://proceedings.mlr.press/v202/li23q.html>.
- [47] Abouelenin, A. *et al.* Phi-4-Mini technical report: Compact yet powerful multi-modal language models via mixture-of-LoRAs. arXiv:2503.01743 [cs.CL] (2025).
- [48] Meng, F., Wang, Z. & Zhang, M. PiSSA: Principal singular values and singular vectors adaptation of large language models. *Advances in Neural Information Processing Systems* **37**, 121038–121072 (2024).
- [49] Radford, A. *et al.* Learning transferable visual models from natural language supervision. *Proceedings of the 38th International Conference on Machine Learning* **139**, 8748–8763 (2021). URL <https://proceedings.mlr.press/v139/radford21a.html>.
- [50] Rombach, R., Blattmann, A., Lorenz, D., Esser, P. & Ommer, B. High-resolution image

- synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition* 10674–10685 (2022).
- [51] Song, J., Meng, C. & Ermon, S. Denoising diffusion implicit models. *The Ninth International Conference on Learning Representations* (2021). URL <https://openreview.net/forum?id=St1gjarCHLP>.
- [52] Feczko, E. *et al.* Adolescent brain cognitive development (ABCD) community MRI collection and utilities. bioRxiv 2021.07.09.451638 [Preprint] (2021). <https://doi.org/10.1101/2021.07.09.451638>.
- [53] Jing, J. *et al.* Development of expert-level classification of seizures and rhythmic and periodic patterns during EEG interpretation. *Neurology* **100**, e1750–e1762 (2023).
- [54] Wang, Z., Zhang, J., Zhang, X., Chen, P. & Wang, B. Transformer model for functional near-infrared spectroscopy classification. *IEEE J. Biomed. Health Inform.* **26**, 2559–2569 (2022).
- [55] Schaefer, A. *et al.* Local–global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. *Cereb. Cortex* **28**, 3095–3114 (2018).
- [56] Yeo, B. T. T. *et al.* The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *J. Neurophysiol.* **106**, 1125–1165 (2011).
- [57] Cole, M. W., Ito, T., Bassett, D. S. & Schultz, D. H. Activity flow over resting-state networks shapes cognitive task activations. *Nat. Neurosci.* **19**, 1718–1726 (2016).
- [58] Simistira Liwicki, F. *et al.* Bimodal electroencephalography–functional magnetic resonance imaging dataset for inner-speech recognition. *Sci. Data* **10**, 378 (2023).
- [59] Dubois, J. *et al.* Reliability of brain metrics derived from a time-domain functional near-infrared spectroscopy system. *Sci. Rep.* **14**, 17500 (2024).
- [60] Duwadi, S. *et al.* Decoding spatial attention in the cocktail party problem using wearable whole-head high-density fNIRS. bioRxiv 2026.05.06.722322 [Preprint] (2026). <https://doi.org/10.64898/2026.05.06.722322>.
- [61] Ma, J. *et al.* A large EEG dataset for studying cross-session variability in motor imagery brain–computer interface. *Sci. Data* **9**, 531 (2022).
- [62] Golmohammadi, M., Harati Nejad Torbati, A. H., Lopez de Diego, S., Obeid, I. & Picone, J. Automatic analysis of EEGs using big data and hybrid deep learning architectures. *Front. Hum. Neurosci.* **13**, 76 (2019).
- [63] Lin, F.-H. *et al.* somatomotor. OpenNeuro [Data set] (2025). URL <https://doi.org/10.18112/openneuro.ds006035.v1.0.0>.
- [64] Wang, S., Zhang, X., Zhang, J. & Zong, C. A synchronized multimodal neuroimaging

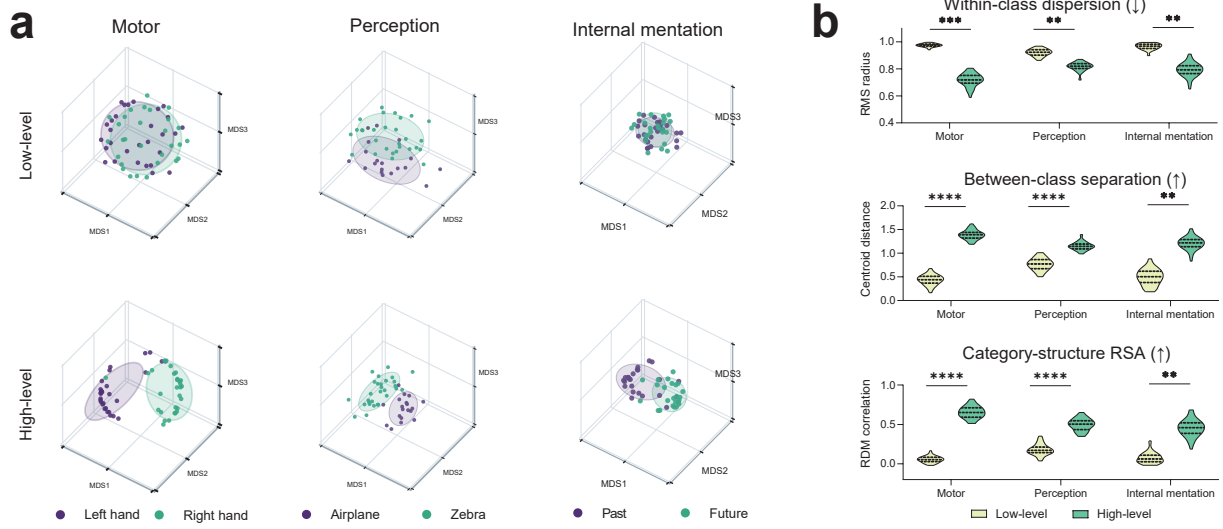
- dataset for studying brain language processing. *Sci. Data* **9**, 590 (2022).
- [65] Nastase, S. A. *et al.* The “Narratives” fMRI dataset for evaluating models of naturalistic language comprehension. *Sci. Data* **8**, 250 (2021).
- [66] Gong, Z. *et al.* A large-scale fMRI dataset for the visual processing of naturalistic scenes. *Sci. Data* **10**, 559 (2023).
- [67] Peterson, S. M. *et al.* AJILE12: Long-term naturalistic human intracranial neural recordings and pose. *Sci. Data* **9**, 184 (2022).
- [68] Angrick, M. *et al.* Online speech synthesis using a chronically implanted brain–computer interface in an individual with ALS. *Sci. Rep.* **14**, 9617 (2024).



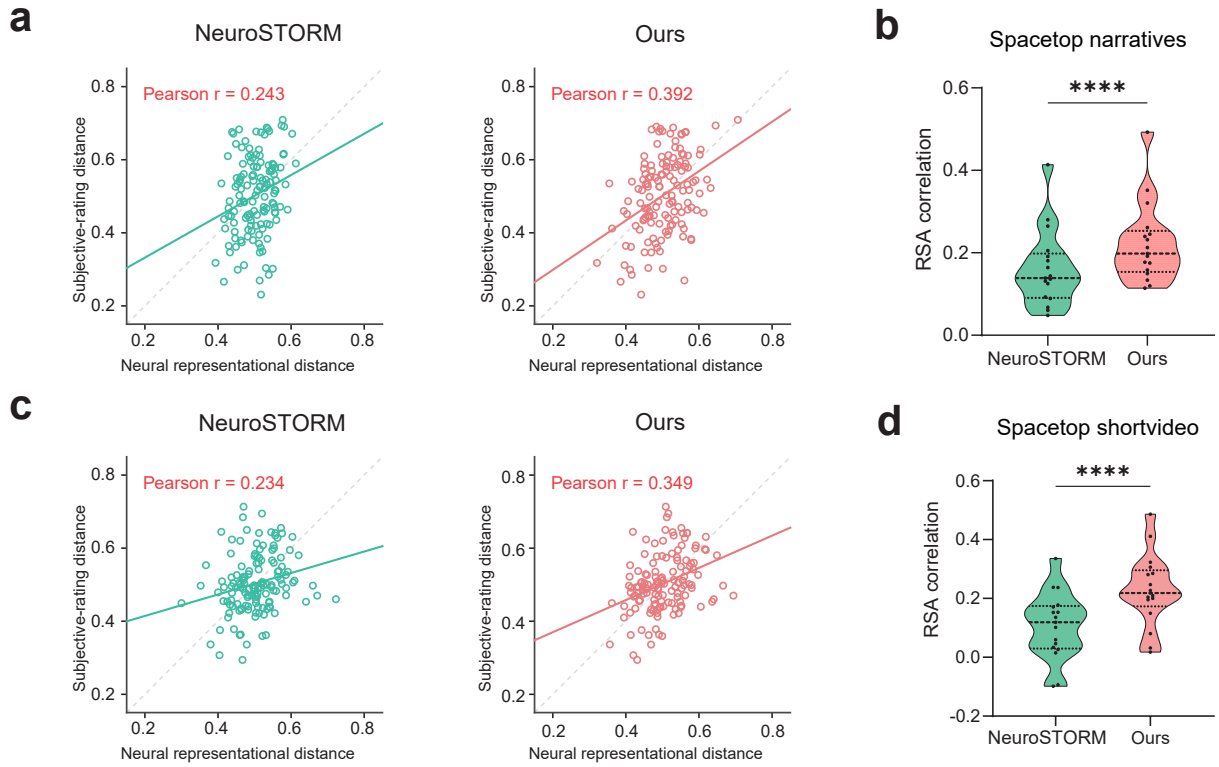
Extended Data Fig. 1 Semantic-attribution maps across participants. Maps of the low-level neural observation representation and high-level meta-neural semantic representation are shown for 12 SMN4Lang participants. Percentages indicate cortical coverage.



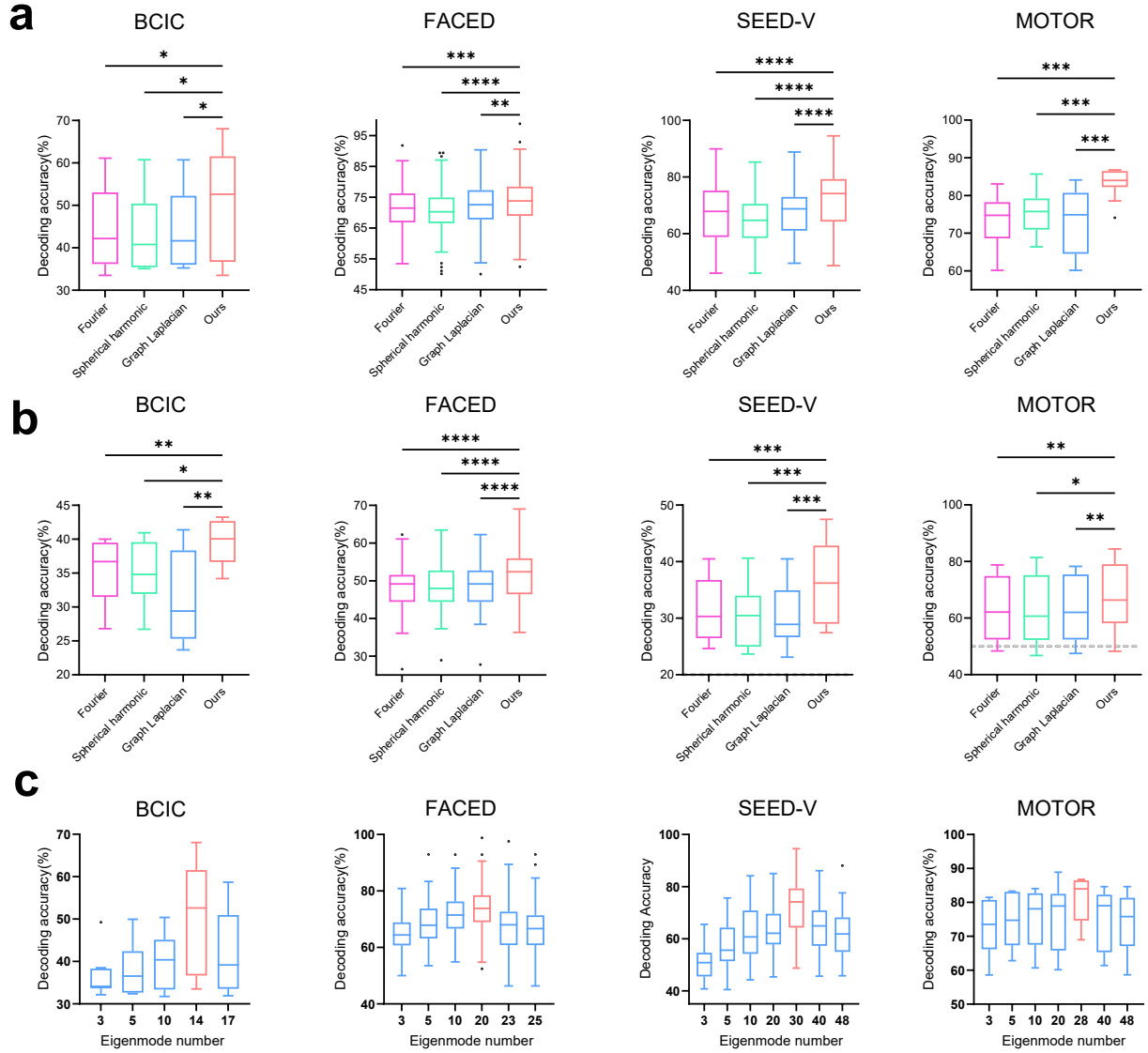
Extended Data Fig. 2 Subjective-state prediction across Spacetop tasks. **a**, Prediction-rating correlations for expectation intensity, expectation valence, feeling intensity and feeling valence in Spacetop Narratives, and for likeability, mentalizing and similarity in Spacetop Shortvideo. Each target included $n = 30$ participants. **b**, Prediction of participant-specific rating residuals after removal of the shared stimulus-locked component. Green denotes NeuroSTORM and red denotes the proposed method. Box plots show the median (center line), 25th and 75th percentiles (box limits), and $1.5 \times \text{IQR}$ whiskers; values outside the whiskers are shown individually. Violin plots show participant-level distributions. Correlation coefficients were Fisher z -transformed before statistical testing; raw Pearson r values are shown. Two-way repeated-measures ANOVA followed by Šidák-adjusted multiple comparisons was used in each task, with comparisons between methods at each rating dimension. Asterisks denote statistical significance ($*P < 0.05$, $**P < 0.01$, $***P < 0.001$ and $****P < 0.0001$).



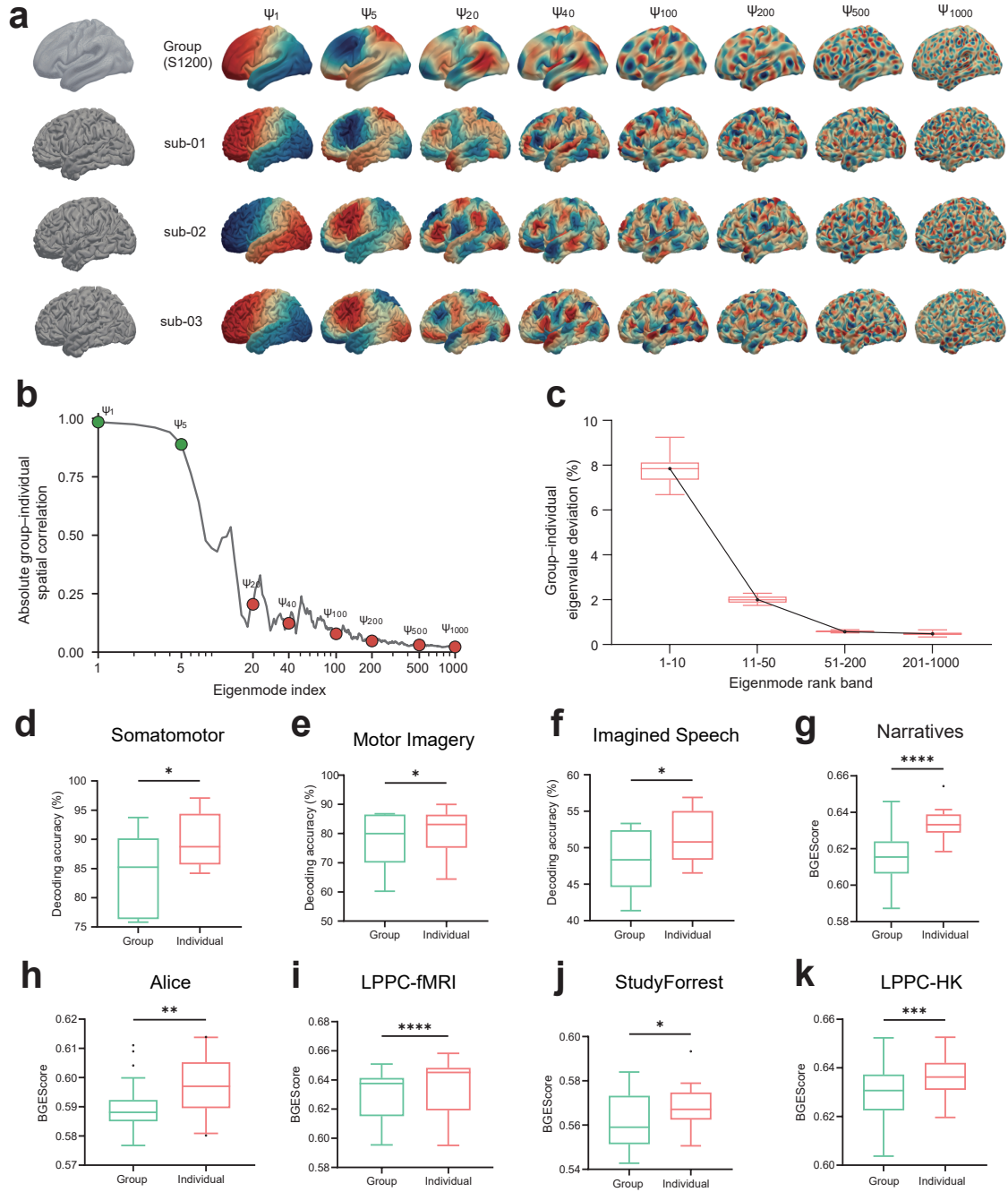
Extended Data Fig. 3 Representational geometry of low-level and high-level representations across cognitive categories. **a**, Multidimensional-scaling visualizations for motor, perception and internal-mentation categories in the low-level neural observation and high-level meta-neural semantic representations. **b**, Within-class dispersion, between-class separation and category-structure representational similarity analysis (RSA) for the two representations. Violin plots show distributions, with internal lines indicating medians and quartiles. Asterisks denote statistical significance (* $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$ and **** $P < 0.0001$).



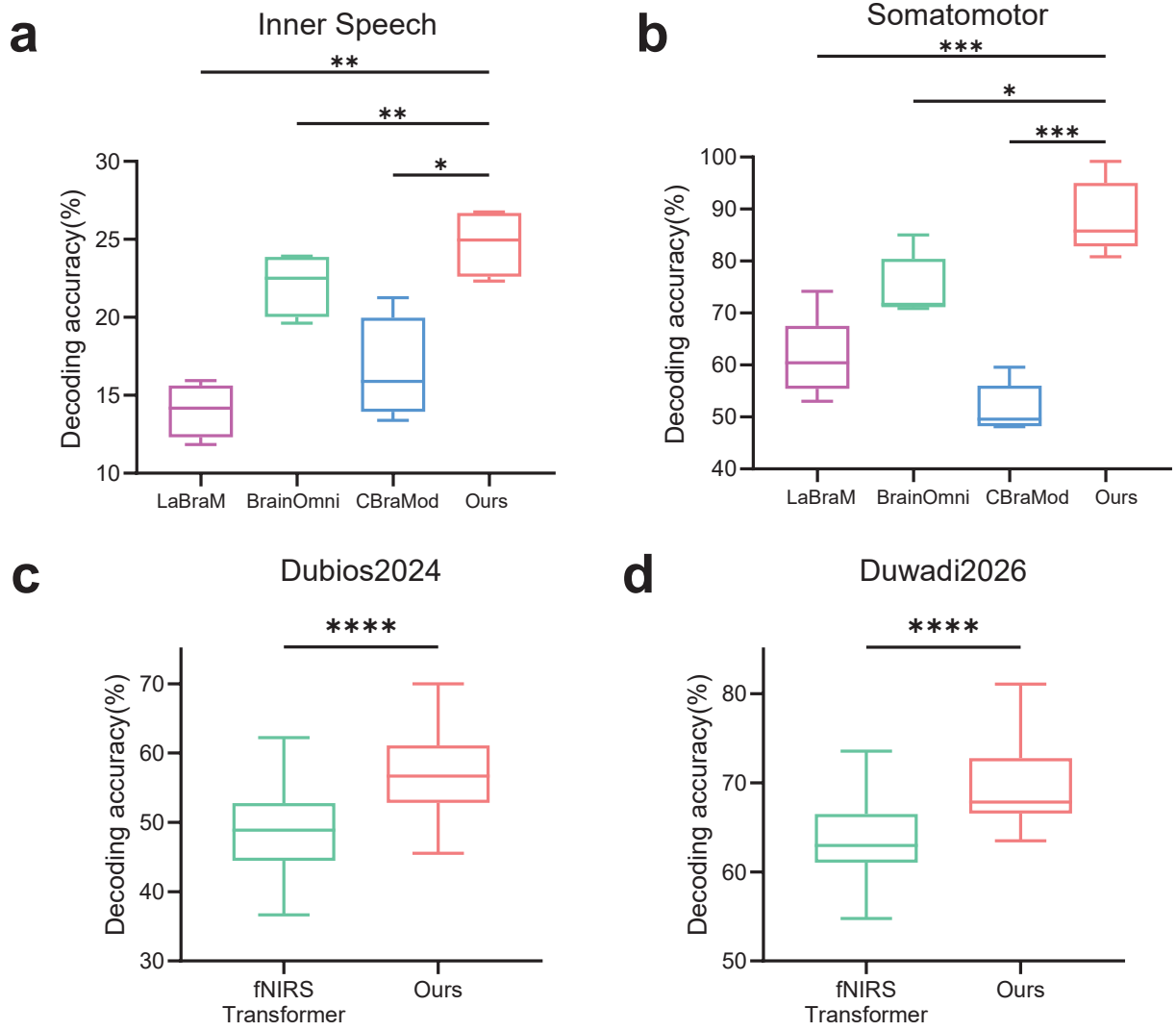
Extended Data Fig. 4 Neural representations relate to subjective ratings in Spacetop. **a**, Association between neural representational distance and rating distance across 136 participant pairs formed from $n = 17$ participants in Narratives. **b**, Participant-level RSA correlations between neural representations and Narratives ratings. **c**, Corresponding participant-pair analysis for Shortvideo ratings. **d**, Participant-level RSA correlations for Shortvideo ratings. Green denotes NeuroSTORM and red denotes the proposed framework. Points in **a** and **c** denote participant pairs, and lines show linear fits with Pearson correlation coefficients. Violin plots in **b** and **d** show participant-level distributions; internal lines indicate the median and quartiles. Two-tailed paired t -tests on Fisher z -transformed participant-level RSA coefficients were used in **b** and **d**. Asterisks denote statistical significance ($*P < 0.05$, $**P < 0.01$, $***P < 0.001$ and $****P < 0.0001$).



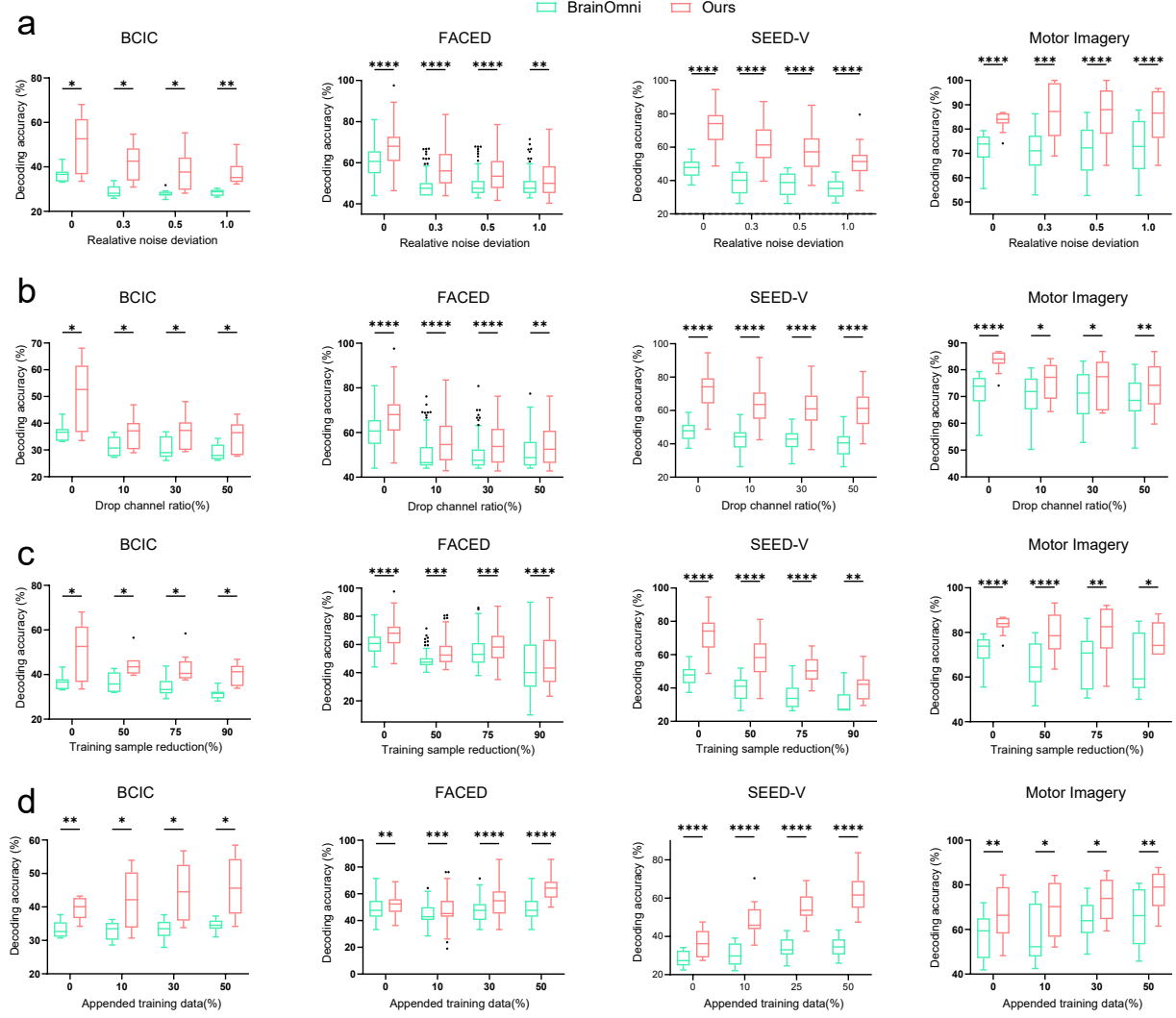
Extended Data Fig. 5 Ablation analyses of cortical geometric eigenmodes. **a**, Within-participant decoding accuracy using Fourier, spherical-harmonic, graph-Laplacian and cortical geometric eigenmode bases. **b**, Cross-participant decoding accuracy for the same bases. BCIC included $n = 9$ participants, FAGED $n = 123$, SEED-V $n = 16$ and Motor Imagery $n = 10$. **c**, Decoding accuracy as a function of the number of retained modes. Red boxes mark the dimensionalities used in the main analyses: 14, 20, 30 and 28 modes. Box plots show the median (center line), 25th and 75th percentiles (box limits), and $1.5 \times \text{IQR}$ whiskers; values outside the whiskers are shown individually. One-way repeated-measures ANOVA with Greenhouse–Geisser correction followed by Šidák-adjusted multiple comparisons was used in **a** and **b**. Asterisks denote statistical significance ($*P < 0.05$, $**P < 0.01$, $***P < 0.001$ and $****P < 0.0001$).



Extended Data Fig. 6 Individual cortical geometry and decoded features. **a**, Cortical surfaces and geometric eigenmodes 1, 5, 20, 40, 100, 200, 500 and 1,000 from the HCP S1200 group reference and three representative participants. **b–c**, Group-individual correspondence across cortical eigenmode ranks. **b**, Absolute same-rank spatial correlations $|r|$ between HCP S1200 group eigenmodes and individual eigenmodes for one participant from the Alice dataset. **c**, Group-individual deviations in cortical-area-normalized eigenvalues across four eigenmode-rank bands in 26 Alice participants. For each participant, deviations were averaged within each band, and box plots summarize the participant distributions. Higher-rank eigenmodes showed weaker spatial correspondence but increasingly consistent eigenvalue scales. **d–f**, Decoding accuracy using group-level and participant-specific cortical geometric eigenmodes in SomatoMotor (two-class; $n = 5$), Motor Imagery (four-class; $n = 10$) and the Imagined Speech OPM-MEG dataset collected for this study (three-class; $n = 6$), respectively. **g–k**, BGEScore using group-level and participant-specific cortical geometric eigenmodes in Narratives, Alice, LPPC-fMRI, StudyForrest and LPPC-HK, with $n = 37, 26, 112, 20$ and 52 participants, respectively. Box plots in **c–k** show the median (center line), the 25th and 75th percentiles (box limits), and $1.5 \times \text{IQR}$ whiskers; values outside the whiskers are shown individually. Two-tailed paired t -tests were used in **d–k**. Asterisks denote statistical significance (* $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$ and **** $P < 0.0001$).



Extended Data Fig. 7 Additional discriminative decoding results. **a**, Inner-speech decoding accuracy across eight classes in Bimodal Inner Speech ($n = 4$ participants) for LaBraM, BrainOmni, CBraMod and the proposed framework⁵⁸. **b**, Somatomotor decoding accuracy across two classes in SomatoMotor ($n = 5$) for the same methods.⁵⁹ **c**, Decoding accuracy across three classes in Dubois2024 ($n = 49$)⁵⁹ for the fNIRS Transformer⁵⁴ and the proposed framework. **d**, Two-class decoding accuracy for cocktail-party fNIRS; 30 of the 38 source participants had valid recordings for analysis⁶⁰. Box plots show the median (center line), 25th and 75th percentiles (box limits), and $1.5 \times \text{IQR}$ whiskers; values outside the whiskers are shown individually. Brackets indicate pairwise comparisons. One-way repeated-measures ANOVA with Greenhouse–Geisser correction followed by Šidák-adjusted multiple comparisons was used in **a** and **b**. Two-sided paired t -tests were used in **c** and **d**. Asterisks denote statistical significance ($*P < 0.05$, $**P < 0.01$, $***P < 0.001$ and $****P < 0.0001$).



Extended Data Fig. 8 Robustness of decoding performance. **a**, Decoding accuracy after adding increasing levels of noise relative to the training-fold channel standard deviation. **b**, Accuracy after dropping increasing proportions of channels. **c**, Accuracy as progressively smaller fractions of within-participant training data were used. **d**, Accuracy in the cross-participant setting as increasing amounts of training data from the target participant were introduced. Each row compares BrainOmni with the proposed framework; columns show BCIC, FACED, SEED-V and Motor Imagery, with $n = 9, 123, 16$ and 10 participants, respectively. Box plots show the median (center line), 25th and 75th percentiles (box limits), and $1.5 \times \text{IQR}$ whiskers; values outside the whiskers are shown individually. Brackets indicate pairwise comparisons. Two-way repeated-measures ANOVA followed by Šidák-adjusted multiple comparisons was used for each dataset and analysis, with methods compared at each level. Asterisks denote statistical significance ($*P < 0.05$, $**P < 0.01$, $***P < 0.001$ and $****P < 0.0001$).

Extended Data Table 1 Datasets used in this study. N denotes the reported cohort size (participants).

No.	Dataset	Modality	N	Task	Ref.
1	FACED	EEG	123	Affective video viewing	20
2	SEED-V	EEG	16	Affective video viewing	21
3	BCIC-IV-2a	EEG	9	Motor imagery	19
4	SHU-MI	EEG	25	Motor imagery	61
5	TUEV	EEG	518	Clinical EEG event classification	62
6	SomatoMotor	MEG + EEG	5	Somatosensory stimulation	63
7	Duwadi2026	fNIRS	38	Auditory spatial attention	60
8	Dubois2024	fNIRS	49	Auditory and cognitive tasks	59
9	Motor Imagery	fMRI + EEG	10	Motor imagery with neurofeedback	22
10	Bimodal Inner Speech	fMRI + EEG	4	Inner speech production	58
11	SMN4Lang	fMRI + MEG	12	Narrative listening	64
12	Narratives	fMRI	345	Narrative listening	65
13	Alice	fMRI	26	Narrative listening	26
14	LPPC-fMRI	fMRI	112	Narrative listening	27
15	LPPC-HK	fMRI + EEG	52	Narrative listening	29
16	StudyForrest	fMRI	20	Naturalistic movie viewing	28
17	NOD	fMRI	30	Natural image viewing	66
18	NSD	fMRI	8	Natural image viewing	32
19	Spacetop	fMRI	101	Affective and cognitive tasks	37
20	DuPre2016	fMRI	31	Autobiographical cognition	34
21	Lee2021	fMRI	26	Future event imagination	35
22	AJILE12	ECoG	12	Naturalistic upper-limb movement	67
23	ECoG Speech	ECoG	1	Speech production	68