Cross-task, explainable, and real-time decoding of human emotion states by integrating grey and white matter intracranial neural activity

Yuxiao Yang[1,2,3,4,5,&,*], Wenjun Chen[3,4,&], Yuyang Chen[3,4], Ling Ding[3,4], Chi Zhang[3,4], Hongjie Jiang[1,6,7], Zhoule Zhu[1,6,7], Xinxia Guo[1,6,7], Shuang Wang[1,8], Gang Pan[3,5], Ning Wei[9,2], Shaohua Hu[9,2,*], Junming Zhu[1,6,7,*], and Yueming Wang[2,3,*]

[1]Department of Neurosurgery, Second Affiliated Hospital, Zhejiang University, Hangzhou 311100, China
[2]Nanhu Brain-computer Interface Institute, Hangzhou 311100, China
[3]College of Computer Science and Technology, Zhejiang University, Hangzhou 311100, China
[4]MOE Frontier Science Center for Brain Science and Brain-machine Integration, Zhejiang University, Hangzhou 311100, China
[5]State Key Laboratory of Brain-machine Intelligence, Zhejiang University, Hangzhou 311100, China
[6]Clinical Research Center for Neurological Diseases of Zhejiang Province, Hangzhou, 311100, China
[7]Clinical Brain-Computer Interface Zhejiang Provincial Engineering Research Center, Hangzhou, 311100, China
[8]Epilepsy Center, Department of Neurology, Second Affiliated Hospital, Zhejiang University, Hangzhou 311100, China
[9]Department of Psychiatry, First Affiliated Hospital, Zhejiang University, Hangzhou 311100, China
[&]These authors contributed equally
*Corresponding authors, yuxiao.yang@zju.edu.cn; dorhushaohua@zju.edu.cn; dr.zhujunming@zju.edu.cn; ymingwang@zju.edu.cn

Abstract: Decoding human emotion states from neural activity has significant applications in human-computer interfaces, psychiatric diagnostics, and neuromodulation therapies. Intracranial electroencephalogram (iEEG) provides a promising modality for decoding by balancing temporal, spatial, and noise resolutions. However, real-world application of decoding requires high performance that integrates neural activity from both grey and white matter, stable generalization across different contexts, sufficient neural encoding explainability, and robust real-time implementation, all of which remain elusive. Here, we simultaneously recorded iEEG and abundant self-rated valence and arousal scores—measuring the two primitive dimension of emotion—across two emotion-eliciting tasks in eighteen epilepsy subjects. We developed self-supervised deep learning models that achieved high-performance decoding by integrating grey and white matter signals, generalizing across tasks. The models provided strong explainability by revealing shared and preferred mesolimbic-thalamo-cortical subnetworks encoding valence and arousal, as well as the structural connectivity basis underlying grey and white matter integration. Finally, the models were implemented online and realized robust real-time decoding in four new subjects. Our results have implications for advancing emotion decoding neurotechnology and understanding emotion encoding mechanisms.

Main

Emotion represents a fundamental yet complex biological state underlying human survival, social interactions, and mental health[1,2]. To date, there is no consensus on a universal model for emotion[1,2]. Discrete categorical models are adopted by the basic emotion theory[3,4] while continuous dimensional models are preferred by the cognitive[5,6] and constructed[7,8] emotion theory, where valence (a pleasure–displeasure continuum) and arousal (an activation–inactivation continuum) are the most widely recognized core dimension[1,2]. A recent cross-species emotion primitive theory[9,10] views emotions as evolutionarily conserved central brain states arising from external and internal inputs, being encoded by widespread neural circuits, and causally affecting various output behavioural variables. The central brain state is described by a collection of dimensional features termed emotion primitives, including valence and arousal, that distinguish emotion from other behavioural states or reflexes. The theory describes clear causal links among emotion primitives, neural activity, and behaviour, offering a suitable computational framework for decoding (i.e., statistically inferring) human emotion states from observed neural signals[11–14], which is crucial for advancing human-computer interfaces[15], embodied AI[16], psychiatric diagnostics[17], and neuromodulation therapies[18].

Experimental design, signal modality, and data amount are basic considerations for decoding[11–14,19]. Task-based experiments employing standardized stimuli enable more consistent elicitation and more frequent measurement (usually via self-reports) of emotion than spontaneous reports or clinical interviews under open environments[20]. Numerous task-based neuroimaging studies have investigated emotion encoding circuits[21], but their low temporal resolution is not suitable for continuous decoding over time[22]. Scalp electroencephalography (EEG) enabled decoding over time but its low spatial and noise resolutions restrict most applications within classifying limited emotion categories[23–25]; only few recent studies explored regression of valence[26,27]. By contrast, intracranial EEG (iEEG) strikes a balance among temporal, spatial, and noise resolutions and has led to more detailed delineation of emotion encoding neural dynamics[28–36], more accurate emotion classification[37–39], and fine-grained regression of mood or depression symptom scores[40–60]. While related, mood or symptom is not identical with emotion[7,8] and suffers from sparse data collection (<40 reports per subject, see Discussion), hindering robust decoding evaluation and interpretation. Thus, decoding emotion from task iEEG can provide multiple technological advantages but asks for high performance that integrates useful neural activity sources, stable generalization across different tasks, sufficient neural encoding explainability, and robust real-time implementation, all of which remain elusive.

First, high-performance decoding requires a computational model integrating useful neural activity sources. Existing iEEG studies exclusively use grey matter signals for decoding[35–60] because white matter signals are much weaker, traditionally considered as noise, and disregarded[61]. However, white matter tracts provide the structural basis for inter-regional communication in emotion processing[62,63]. Thus, we hypothesize that integrating grey and white matter iEEG signals can boost decoding. But such integration will need the design of specialized decoding models, especially given the higher neural feature dimensionality resulted from adding white matter signals. Existing iEEG decoding uses linear[41–60] or classical machine learning models[37–40], which are likely insufficient to deal with the high dimensionality and nonlinearity of neural dynamics, asking for new decoding models designed from a more advanced deep learning perspective[19].

Second, generalization across different emotion-eliciting tasks is a critical requirement for decoding. The emotion primitive theory proposes that different stimuli activate a common central brain state that causally leads to downstream responses. This is a core feature of emotion because it allows the individual to quickly sense, process, and respond to a wide range of stimuli across various contexts[9,10]. The theory thus predicts that the neural representations and decoding model associated with emotion primitives, as identified in a single task, should be transferable to apply for other tasks. However, existing decoding studies have exclusively conducted single-task analyses[11–14,19,23–25], which did not reflect the central role of emotion across diverse contexts. For practical applications, cross-task transferable decoding is necessary for rapid plug-and-play implementation of a trained decoding model across various scenarios[25,64].

Third, real-world utility of decoding needs the endorsement of a sufficient level of neural encoding explainability provided by the decoding model. Spatially, EEG decoding models have only investigated cortex lobe encoding[23–25] due to their low resolution; iEEG decoding models have offered more precise, region-level explainability (especially at subcortical regions), but have only pointed to the general importance of mesolimbic regions for encoding composite score representations of mood/symptom[40–60]. How different subcortical and cortical regions preferentially encode different emotion primitives remains underexplored. At the brain network level, how decoding models correlate with white matter connectivity pattern has not been explored either. Moreover, the spectral and temporal characteristics of emotion-related neural dynamics have been studied but with no consensus[11,12,19], requiring evidence from more subjects and data.

Fourth, real-time validation is the ultimate test of any decoding model if it is to be used in real-world applications such as affective robotics[65], neurofeedback training[66], and adaptive neuromodulation[67]. For example, in adaptive deep brain stimulation (aDBS) for affective disorders, emotion-related states need to be decoded from neural activity in real time to provide feedback for delivering optimized stimulation[68–71]. Unfortunately, existing decoding studies largely remain limited to offline analyses[35–59], relying on a non-causal cross-validation procedure for decoding evaluation, meaning training data can happen after test data in time; thus, it is difficult to infer if these models are robust enough for real-time applications requiring causal decoding. Also, technical difficulties such as system communication instability, environmental disturbances, and real-time computation constraints have hindered the development of online emotion decoding systems.

Here, we close the above gaps by demonstrating an integrated framework of high-performance, cross-task, explainable, and real-time decoding of emotion states using iEEG in subjects with epilepsy. We first collected wide-coverage iEEG and abundant emotion self-reports (>170 reports per subject) from eighteen subjects participating in two emotion-eliciting tasks (Fig. 1a-c). We then developed self-supervised deep learning models that achieved fine-grained decoding of continuous valence and arousal values, whose performance was boosted by integrating grey and white matter activity and outperformed state-of-the-art iEEG emotion decoding models (Fig. 1d). We next showed interchangeable and transferable decoding across the two tasks (Fig. 1e). We further used the decoding models to delineate shared and preferred mesolimbic-thalamo-cortical subnetworks encoding valence and arousal, and explain the structural connectivity basis underlying grey and white matter integration (Fig. 1f). We finally implemented the models online and demonstrated robust real-time decoding in four new subjects (Fig. 1g). Overall, our results have implications for advancing emotion decoding neurotechnology and understanding emotion encoding mechanisms.

Results

Overview of experimental design, neural recording, and decoding model. We recruited 18 subjects with epilepsy undergoing iEEG monitoring for seizure onset zone localization (Fig. 1a and Supp. Table 1). Each subject completed 4.72±0.03 (mean±s.e.m.) rounds of experiments distributed across 2.27±0.24 days. Each round had three interleaved sessions of two types of emotion-eliciting tasks—static image-viewing task (image-task) and dynamic video-viewing task (video-task) (Fig. 1b, see Methods, Supp. Fig.1, and Supp. Table 2). Each session consisted of 20-23 trials for image-task or 7-10 trials for video-task, where the subject self-rated valence and arousal states using 2 separate visual-analog-scales (VAS) spanning 0-100 for each trial. Each trial lasts for 12-18 seconds for the image-task and 1-4 minutes for the video-task. Subjects not being devoted to completing the tasks were removed from further analyses (see Methods), resulting in 18 and 14 subjects included in the offline video-task and image-task decoding analyses, respectively (Supp. Fig. 2 and Supp. Table 3). Overall, after trial inclusion validation, for each subject, 124±7.25 (valence) and 132.36±6.85 (arousal) trials of image-task and 46.94±1.63 (valence) and 47.17±1.52 (arousal) trials of video-task were included in subsequent decoding (see Methods), where the self-rated scores spanned reasonable ranges (Supp. Fig. 3, Supp. Tables 4 and 5). The subjects' rated scores showed a typical 'boomerang' shape in the valence-arousal space, consistent with prior studies using similar tasks[72] (Supp. Fig. 3). Another 4 subjects with epilepsy were separately recruited in the online decoding experiment (see Methods).

iEEG signals were recorded concurrently with the tasks (Fig. 1c, see Methods). Across subjects, iEEG contacts were implanted in mesolimbic regions (such as amygdala, hippocampus, cingulate, and insula), subcortical regions (mainly thalamus), widespread cortical regions, and white matter tracts (such as inferior longitudinal fasciculus, thalamic radiation superior, and uncinate fasciculus), totaling 41 grey matter regions based on the Desikan-Killiany-Tourville (DKT)[73] and FreeSurfer ASEG[74] atlases, 26 white matter tracts based on the Human-Connectome-Project (HCP)[75] atlas, 1408 grey matter contacts, and 1406 white matter contacts (Supp. Fig.4, Supp. Tables 6 and 7). Each subject's implantation location was purely determined by the clinical need and was different, but covered most of the above locations, with 19.61±0.81 grey matter regions, 17.50±0.40 white matter tracts, 78.39±3.23 grey matter contacts, and 78.11±5.09 white matter contacts per subject (Supp. Fig. 5 and Supp. Table 8).

Our dataset of widespread iEEG signals and abundant emotion measurements allowed in-depth decoding and explainability analyses, as expanded in the next few sections. The foundation for these analyses is the construction of decoding models that map neural activity to valence and arousal ratings. The paradox between the high-dimensional spatial-spectral-temporal neural features and the relatively sparse ratings (although they were much denser than existing studies) makes direct supervised learning difficult, but the large amount of unlabeled iEEG data enables sufficient unsupervised learning. We thus designed a self-supervised deep learning model for dimensionality reduction and decoding (Fig. 2a, see Methods). First, we extracted iEEG spectral powers at δ (1-4 Hz), θ (4-8 Hz), α (8-13 Hz), β (13-30 Hz), low γ (30-70 Hz), and high γ (70-200 Hz) bands every 1 second for each channel to form a neural feature vector time-series per subject (feature dimension of 708±35.89; total time step of 3710.71±217.37 and 3965.07±209.55 in image-task and 5748.28±240.63 and 5781.00±235.58 in video-task for valence and arousal, respectively). Second, we used a supervised F-statistic-driven method to identify the top discriminative features. Third, we used an

unsupervised long short-term memory (LSTM) autoencoder to capture nonlinear neural dynamics through low-dimensional representations that enable one-step-ahead prediction of future features. We chose LSTM over other models, such as the Transformer, because it delivered comparable prediction performance with a simpler architecture (Supp. Fig. 7). Fourth, we built supervised multi-layer perceptron (MLP) modules to use the low-dimensional representations to predict (via regression) fine-grained continuous valence/arousal ratings or categorize (via classification) coarse-grained discretized valence/arousal states. Implementation details are provided in Methods and Supp. Figs. 6 and 7.

We trained, validated, and tested the above decoding model separately for each subject using rigorous 10-fold leave-trials-out cross-validation (Fig. 2b; including cross-validating the model's hyperparameters, see Methods), leading to a personalized decoding framework. The primary decoding performance measures were the cross-validated correlation coefficient (cvCC) in predicting the continuous valence/arousal ratings and the secondary measure was the cross-validated classification accuracy (cvACC) of high, median, and low valence/arousal states. Other typical regression and classification measures were also computed for completeness. Statistical significance of decoding was assessed by applying the identical framework to randomly permuted valence/arousal ratings, with permutation *P*-values quantifying the probability of random decoding performance (chance-level) exceeding the true decoding performance. All *P*-values were corrected with false discovery rate (FDR) control wherever needed.

High-performance single-task decoding of emotion states by integrating grey and white matter signals. Our personalized models enabled high-performance decoding of valence and arousal states in distinct emotion-eliciting tasks. In terms of the video-task, the personalized model significantly decoded fine-grained continuous valence ratings in each subject, with high cvCC values ranging from 0.55 to 0.88 (n = 34 to 56 per subject, where n represents the number of ratings; FDR-corrected permutation $P < 0.008$ for each subject; Fig. 2c; Supp. Fig. 8), as confirmed by the strong consistency of temporal decoding traces with the ground-truth rating pattern (Fig. 2c). At the population level, the pooled cvCC was 0.70±0.02, significantly larger than chance-level (N=18, where N represents the number of subjects, n=845 across subjects, $P = 7.63\times10^{-9}$; Fig. 2c; Supp. Fig. 9a). For arousal, the individual cvCC ranged from 0.34 to 0.76 (FDR-corrected $P < 0.03$ for each subject; Fig. 2d; Supp. Fig. 8) and the population-pooled cvCC was 0.52±0.03 ($P = 7.63\times10^{-9}$; Fig. 2d; Supp. Fig. 9a). Moreover, a simple variant of our model enabled significant classification of high, median, and low valence/arousal states with individual-level cvACC ranging from 0.65 to 0.83 and 0.59 to 0.81, respectively (FDR-corrected $P < 0.0008$ for each subject; Supp. Fig. 10), and pooled cvACC reaching 0.75±0.01 and 0.70±0.01, respectively ($P < 0.01$ for both cases; Supp. Fig. 10). A coarser classification of two-class high and low valence/arousal states led to population-pooled cvACC of 0.81±0.01 and 0.74±0.01, respectively ($P < 0.01$ for both cases; Supp. Fig. 10). Other typical regression and classification performance measures for each subject were included in Supp. Table 9.

Similar results held for image-task decoding. First, our model significantly decoded valence and arousal, with cvCC ranging from 0.41 to 0.70 and 0.33 to 0.68, respectively, for each subject (n = 61 to 174 and 69 to 175 per subject, respectively; FDR-corrected $P < 0.0007$ and $P < 0.0046$, respectively; Fig. 2e and 2f; Supp. Fig. 11; Supp. Table 10) and a population-pooled cvCC of 0.57±0.02 and 0.51±0.03, respectively (N=14, n=1736 and 1853 across subjects, respectively; $P = 1.22\times10^{-4}$ for both; Fig. 2e and 2f, Supp. Fig. 9b). Second, the individual cvACC of three-class valence and arousal classifications ranged from 0.62 to 0.79 and 0.53 to 0.8, respectively (FDR-corrected $P < 0.006$ for each subject; Supp. Fig. 10). The population-pooled cvACC achieved 0.67±0.01 and 0.66±0.02, respectively ($P < 0.01$ for both cases; Supp. Fig. 10; Supp. Table 10).

The high decoding performance was achieved by integrating grey and white matter signals, outperforming each single type of signal alone. While white matter signals indeed had smaller amplitudes than grey matter signals (Fig. 2g), they were not pure noise but contained substantial information about emotion (Fig. 2h and Supp. Fig. 14): they decoded valence and arousal significantly better than chance-level across all subjects (valence cvCC= 0.60±0.02 and 0.47±0.02, arousal cvCC= 0.46±0.02 and 0.43±0.02 in video-task and image-task, respectively, $P < 0.0002$ in all cases), although the performance was slightly worse than grey matter (valence cvCC, 0.60±0.02 vs. 0.64±0.02 and 0.47±0.02 vs. 0.52±0.02; arousal cvCC= 0.46±0.02 vs. 0.49±0.02 and 0.43±0.02 vs. 0.47±0.03 in video-task and image-task, respectively). However, the decoding performance of combined grey and white matter signals (detailed in the previous paragraph) significantly

outperformed either signal type alone (Fig. 2h, $P<0.0005$ in all cases). These results indicate that signals recorded along white matter tracts provide complementary functional information that augments grey matter signals, boosting the decoding performance, whose neural structural basis will be explained in later sections.

In addition, we conducted several control analyses to confirm the effectiveness of decoding (Fig. 2i). First, since prior iEEG studies have exclusively used linear dimensionality reduction and regression models (see Discussion), we replaced the MLP module with regularized linear regression and then replaced the LSTM autoencoder module with linear principal component analysis, leading to a decrease of video-task valence decoding cvCC from 0.70±0.02 to 0.52±0.03, and further to 0.40±0.04 (a similar decrease held for image-task and arousal results; Supp. Note 1 and Supp. Fig. 12). This confirms the necessity of nonlinear dimensionality reduction and nonlinear learning for more precise decoding. Second, we compared with the recent deep learning-based CEBRA algorithm[76] designed for general decoding purposes and found that our decoding model achieved significantly better performance (video-task valence 0.70±0.02 vs. 0.59±0.02, $P = 7.63\times10^{-9}$; similar results held for image-task and arousal; Supp. Note 1 and Supp. Fig. 12), showing the necessity of explicit temporal neural dynamics prediction in our LSTM encoder. Third, to exclude the influence of epileptic activity[77], we extracted interictal discharge rates as the neural features, repeated the same decoding analyses, and found that decoding failed for valence or arousal in either task (mean population-pooled cvCC < 0.14, P >0.06 in all cases; Supp. Note 2; Supp. Fig. 13). Fourth, to exclude the possibility that the larger number of included neural features purely contributed to the improvement of combined grey and white matter decoding, we randomly removed iEEG channels in the combined case to match that of grey or white matter decoding alone and still found significantly improved decoding performance (Supp. Note 3 and Supp. Fig. 14).

Interchangeable, transferable, and temporally-stable decoding across distinct emotion-eliciting tasks. We have demonstrated that emotion states can be effectively decoded from iEEG separately for the video-task and image-task. With the assumption that different emotion stimuli activate a common central brain state, we next investigated whether the neural representations of the central brain state enable interchangeable, transferable, and temporally-stable decoding across the two tasks. The cohort here comprised the 14 subjects who were included in both tasks.

We first investigated whether a model trained by data from one task (source-task) could be directly interchanged to decode emotion states in the other task (target-task). We trained three models (Fig. 3a): (1) a model trained on target-task data (intra-task model), (2) a model trained on source-task data (inter-task model), and (3) a permutation model trained with randomly permuted target-task labels (random model). We evaluated these models using the same target-task test data with 10-fold leave-trials-out cross-validation (see Methods). Taking image-task as the source-task and video-task as the target-task as an example, the inter-task model achieved a population-level decoding cvCC of 0.34±0.03 for valence (N =14, n=643), significantly outperforming the random model (cvCC=0.03±0.02; $P=1.22\times10^{-4}$; Fig. 3b), though the intra-task model, as expected, achieved the best performance (cvCC=0.70±0.02; Fig. 3b). Consistently, t-SNE visualization of the hidden layer embedding of the classification MLP in each subject demonstrated enhanced cluster separation of the inter-task model compared to the random model (Fig. 3c and Supp. Fig. 15) with significantly higher clustering efficacy values (Calinski-Harabasz index, CHI, Fig. 3d), while the intra-task model still had the best cluster separation. Similar results held for arousal decoding (Supp. Fig. 16) and for interchanging the roles of video-task and image-task (Supp. Fig. 17). Such interchangeable decoding indicates common neural representations across different tasks and the possibility of using part of the target-task data to fine-tune the inter-task model for better decoding the target-task.

Therefore, we explored whether the inter-task model can be used as a 'pretrained source-task' model to enable rapid fine-tuning by the target-task data within a transfer learning framework (Fig. 3e). Taking image-task as the source-task and video-task as the target-task as an example, across the population, the pretrained source-task model fine-tuned with only 30% of target-task trials achieved a sufficiently good cvCC of 0.63±0.02 for valence, then saturated at 50% of target-task trials with a cvCC of 0.68±0.02, which was already comparable to the optimal intra-task performance of 0.70±0.02 (N = 14, n = 570, P=0.05; Fig. 3f). The performance was significantly better than fine-tuning the random model at each data amount below 50% (FDR-corrected P<0.036; Fig. 3f). t-SNE visualization of the fine-tuning process showed a clear progressive

refinement pattern of embedding clusters for the pretrained source-task models (Fig. 3g and Supp. Fig. 18), outperforming random models with higher cluster efficacy values (FDR-corrected $P < 0.0003$; Fig. 3h). Similar results held for arousal (Supp. Fig. 16) and for interchanging the roles of video-task and image-task (Supp. Fig. 17). To confirm the temporal stability of transfer decoding, we additionally performed a causal transfer analysis. The model was first pre-trained on the source-task, then fine-tuned using the initial round of target-task data (approximately 20% of target-task data), and finally tested on subsequent rounds of target-task data (time elapsed from initial round: 15.03 ± 3.03 hours; range: 1–44 hours). The decoding performance remained high and stable across test rounds (FDR-corrected $P<0.0042$ at each elapsed time in all cases; Supp. Fig. 19), showing the feasibility of rapid and causal fine-tuning and the multi-hour transfer stability.

To further confirm the existence of shared neural representation patterns across different tasks, we conducted a control analysis by mixing video-task and image-task data for training a common mixed decoding model and testing it for separate decoding of video-task and image-task with 10-fold leave-trials-out cross-validation (see Methods). We found that the mixed model significantly decoded valence for the video-task and for the image-task, performing similarly with the intra-task models ($P > 0.19$ in both cases; Fig. 3i); by contrast, mixing video-task (or image-task) and random data for training led to significantly worse decoding ($P < 2.44\times10^{-4}$ in both cases; Fig. 3j). Similar results held for arousal decoding (Supp. Fig. 20).

Spatial encoding explanation: Shared and preferred grey matter regions (or white matter tracts) encoding valence and arousal. Having demonstrated effective cross-task decoding, we used the personalized decoding models to summarize population-level spatial encoding patterns of valence and arousal. Note that since our results so far had indicated common neural representations across tasks, we trained/validated mixed-task decoding models for 14 subjects by using data from both the video-task and image-task and trained/validated video-task decoding models for the other 4 subjects excluded in the image-task by using data from only the video-task (in total, $N = 18$; single-subject $n = 143.33\pm13.08$ and 150.00 ± 13.65 for valence and arousal, respectively).

First, in each subject, we identified top-decoding grey matter subnetworks from all implanted regions. Previous studies have indicated multi-region subnetworks that reflect different aspects of emotion and mood, we thus identified subject-specific top-decoding subnetworks of valence (or arousal) by using a greedy algorithm that starts with a single region and gradually adds regions one by one to form best-performing subnetworks with increasing sizes (with a rigorous nested cross-validation procedure to prevent overfitting to the region selection process, see Methods). In each subject, the typical pattern is that as the number of included regions increased, the decoding performance gradually increased and saturated at around four regions, reaching sufficiently good decoding compared with best-possible decoding of using all implanted regions (Supp. Fig. 21). Such observations were confirmed by the average pattern across subjects ($N = 18$, valence: $n = 2581$, four-region cvCC= 0.55±0.02 vs. all-region cvCC = 0.60±0.01, $P=0.06$; arousal: $n = 2702$, four-region cvCC= 0.47±0.02 vs. all-region cvCC = 0.52±0.02, $P=0.13$; Fig. 4a and 4f). Consistently, in terms of the region selection frequency computed across all cross-validation folds for each subject (along each column of Fig. 4b and 4g), a few regions (again consistently around four) had higher selection frequency than other regions, indicating the robust selection of these regions across folds. However, the selected regions for each subject were quite different, which might be due to the heterogeneity of implantation locations or the inherent difference of personalized emotion processing circuits (see Discussion).

Second, despite some inter-subject variability, the overall distribution of all subjects' top-decoding subnetworks revealed common encoding patterns at the population level. Specifically, the rows in Fig. 4b and 4g showed that some regions were more frequently selected across subjects than others. We computed the frequency of each region being selected as part of the top-decoding subnetworks across all cross-validation folds among the population (Fig. 4c and 4h) and found that this selection frequency was positively correlated with single-region decoding performance (Fig. 4d-e and 4i-j), confirming the decoding stability of these regions. Therefore, we estimated each region's population-level encoding strength of valence and arousal by linearly combining its population-averaged selection frequency and single-region decoding performance (see Methods). We defined a population-level set of key-encoding regions as the regions whose encoding strength for valence or arousal ranked top 50% among all available regions.

Third, at the population level, we explored in detail whether certain regions jointly encoded the valence and arousal emotion primitives and whether some regions exhibited a stronger encoding preference towards one emotion primitive than the other. We divided the key-encoding regions into three groups according to each region's relative encoding strength: (1) a shared group that had comparable strong valence and arousal encoding, (2) a valence-preferred group that had relatively stronger valence encoding than arousal, and (3) an arousal-preferred group that had relatively stronger arousal encoding than valence (Fig. 4k). Our results showed that the shared group mainly included mesolimbic structures such as the amygdala (AMYG), hippocampus (HIP), insula (INS), lateral/medial orbitofrontal cortex (lOFC and mOFC), their densely connected temporal cortex (MTG and FFG), and the typical fronto-parietal network (rostral middle frontal cortex, RMF, and precuneus, PCUN). The valence-preferable group largely situated in the more posterior and lateral parts of the frontal and parietal cortices, including the inferior frontal gyrus (IFG), the caudal middle frontal gyrus (CMF), and the inferior parietal lobe (IPL), as well as the caudal anterior cingulate (cACC) cortex. The arousal-preferable group consisted of more posterior part of the cingulate (isthmus cingulate cortex, IC), the primary somatosensory cortex (the postcentral gyrus, PoG), and especially, the thalamus (THA) (Fig. 4l and Supp. Fig. 23). Critically, these are typical regions that have been implicated in emotion, valence and arousal processing by extensive neuroimaging and animal studies, showing the strong biological consistency of our encoding explanation (see Discussion).

The same analytical pipeline identified population-level key-encoding white matter tracts (Fig. 5a-j; Supp. Fig. 22; Supp. Note 4), also revealing shared and preferred patterns for valence and arousal encoding (Fig. 5k-l; Supp. Fig. 24). The shared group mainly comprised tracts that connect within limbic structures and connect limbic structures to widespread cortical regions. For example, the cingulum parahippocampal (C_PH), as the core white matter pathway of the limbic system, connects the anterior/posterior cingulate cortex and the parahippocampal gyrus; the inferior longitudinal fasciculus (ILF) and uncinate fasciculus (UF) connect the amygdala and hippocampus to widespread temporal, frontal, and occipital regions. Also, the arcuate fasciculus (AF) is the key tract bundle that directly connects the frontal and parietal lobes and forms the structural basis of the fronto-parietal network. The valence-preferred group mainly included the parahippocampal parietal cingulum (C_PHP), connecting the limbic system to more localized posterior parts of the parietal cortex, and the anterior corticostriatal tract (CS_A) and superior longitudinal fasciculus II (SLF2), widely connecting the frontal regions to the striatum and parietal regions. The arousal-preferred group especially encompassed the anterior thalamic radiation (TR_A), connecting the thalamus to prefrontal and parietal regions, and the superior longitudinal fasciculus I (SLF1) and posterior corticostriatal tract (CS_P), connecting the postcentral gyrus to limbic, parietal and frontal regions.

Network encoding explanation: structural connectivity pattern explaining grey and white matter integration. The boosted decoding performance by integrating grey and white matter signals is likely facilitated by network-level structural connectivity. It should be noted that the key-encoding grey matter regions were identified at the population level. Ideally, in a given subject, all key-encoding regions would have implanted iEEG contacts to maximize decoding performance. However, in practice, only a subset of these key-encoding regions was implanted according to the subject's clinical need (Supp. Note 5). On the other hand, white matter tracts interconnect distributed grey matter regions; indeed, we found that in any given subject, the subject-specific top-decoding white matter tracts had connections with not only key-encoding grey matter regions implanted in that subject but also non-implanted key-encoding regions (termed 'Non-implanted key-encoding Grey matter regions Connected by subject-specific top-decoding White matter tracts', $\mathrm{N_{grey}C_{white}}$ for short; Supp. Fig. 25). We thus hypothesized that white matter signals contain network information that originates from $\mathrm{N_{grey}C_{white}}$ regions, which complement the signals directly recorded from localized, implanted regions, subsequently leading to boosted decoding performance (Fig. 6a).

To test the hypothesis, we first conducted case studies of four representative subjects with varying implantation configurations (Fig. 6b). Subjects 4 and 5 exhibited substantial valence decoding performance enhancement after incorporating white matter signals (0.13 and 0.08 increase of cvCC). For them, the implanted iEEG contacts did not directly cover many of the critical key-encoding grey matter regions such as the amygdala, hippocampus and thalamus in subject 4, and hippocampus and caudal anterior cingulate in subject 5. However, the implanted white matter tracts such as the ILF, C_PHP and CST connect to these

critical non-implanted key-encoding grey matter regions. Accordingly, our decoding models effectively aggregated these white matter signals to complement the limited grey matter signals, leading to substantial decoding enhancement. By contrast, subject 1 and 13 only demonstrated a small decoding performance enhancement (0.017 and 0.016 increase in cvCC), likely due to the already comprehensive implantation coverage of key-encoding grey matter regions (covered 10 and 13 regions accordingly) alongside electrode placements in white matter tracts that have few connection to non-implanted key-encoding regions (only 3 and 2 regions), leading to little decoding enhancement. The above findings generalized to other subjects, as revealed by the population-level positive correlation between the percentage of $\mathrm{N_{grey}C_{white}}$ regions and the cvCC enhancement from grey-matter-only decoding to all-contact decoding (*N* = 18, Pearson's correlation coefficient = 0.54, *P*=0.03; Fig. 6c and 6d; Supp. Fig. 26; see Supp. Note 5 for more details).

We further simulated an ablation study where for the 6 subjects that had many implanted key-encoding grey matter regions (thus a small percentage of $\mathrm{N_{grey}C_{white}}$ regions<20%), 20%-50% of implanted key-encoding regions were artificially and randomly removed. Thus, these subjects would have a larger number of $\mathrm{N_{grey}C_{white}}$ regions and we would expect a larger cvCC enhancement for them. Indeed, as an example, for subject 13, the simulation led to a much larger number of $\mathrm{N_{grey}C_{white}}$ regions (5 vs. 2) and accordingly, a much larger decoding enhancement was observed (0.075 vs. 0.016; Fig. 6e). Across subjects involved in the simulation, when compared with the original data, we observed a significantly larger percentage of $\mathrm{N_{grey}C_{white}}$ regions (37.44±2.29% vs. 16.19±0.77%, *P<0.0002*) and a significantly larger cvCC enhancement (0.11±0.02 vs. 0.02±0.003, *P<0.0002*). Consistently, the population-level positive correlation between the percentage of $\mathrm{N_{grey}C_{white}}$ regions and the cvCC enhancement was still preserved in the simulation (Fig. 6f). The results together showed that white matter signals contained information from non-implanted key-encoding grey matter regions, which was effectively aggregated by our model for boosted decoding.

Spectral and temporal interpretation. We further used our neural decoding models to explore how spectral and temporal neural dynamics facilitate emotion decoding. In terms of single spectral band decoding, the performance using the γ band features consistently outperformed the other lower frequency bands (FDR-corrected *P* < 0.043; Fig. 7a and 7d for grey and white matter, respectively). Such higher γ-band performance was consistently observed across key-encoding grey matter regions and white matter tracts (Fig. 7b and 7e). It was further confirmed by a more detailed time-frequency analysis time-locked to the starting of each trial showing that the γ time-frequency power significantly correlated with valence/arousal ratings across all trials and subjects (FDR-corrected cluster-based *P* < 0.042; Fig. 7c and 7f; Supp. Figs. 27 and 28). Moreover, the decoding performance when using all six frequency bands was significantly higher than using each single frequency band alone (FDR-corrected *P* < 0.008; Fig. 7a and 7d), suggesting complementary emotion-related information may be encoded by different frequency bands.

Our model can take neural feature time-series as input and output decoded valence/arousal time-series with a time-step of 1 second, allowing for a detailed investigation of the temporal neural feature dynamics contributing to decoding. We quantified the timescale of decoded time-series for each subject by computing its frequency-domain power density (FPD, quantifying how fast the time-series varies over time, see Methods). Across the population, the FPD of decoded emotion time-series was significantly higher than random-decoded time-series at timescales larger than 16 seconds for the video-task and image-task (FDR-corrected *P*<0.032 at each timescale; Fig. 7g and 7h; Supp. Fig. 29), indicating neural feature dynamics slower than tens of seconds mainly contributed to emotion decoding. The peaking FPD value was at 128 seconds for the video-task and 28 seconds for the image-task, which roughly aligned with the period of showing the emotion-eliciting material in these two tasks, further confirming our identification of emotion-related temporal neural feature dynamics.

Online implementation of neural decoding models and robust real-time emotion decoding. We finally implemented the neural decoding model online and demonstrated a robust real-time emotion decoding system in 4 new subjects not included in the above offline analyses (Fig. 8a; Supp. Tables 1,3,4,5 and 8). The online decoding system consisted of four core components: (1) an emotion-eliciting task presentation computer, (2) a camera for continuous and synchronized monitoring of the subject's facial expression, (3) a neural interface device for real-time iEEG signal streaming, and (4) a laptop workstation dedicated to signal

processing, neural decoding model implementation, and real-time decoding result visualization. For each subject, four personalized decoding models were trained in a separate training experiment (including four rounds) one day before the online experiment: one video-task valence decoding model, one video-task arousal decoding model, one image-task valence decoding model, and one image-task arousal decoding model. Each model was then fixed and implemented in the online experiment while the subject was doing the video-task and the image-task (three rounds), respectively (Fig. 8b and 8c). Note that the subject was not informed of the neural signals being decoded or given feedback of the real-time decoding results; the online system only passively decoded the emotion states from neural activity and computed the online decoding performance measure. The measure, denoted as onlineCC, was defined as the CC between model-predicted valence/arousal ratings and the ground-truth scores rated up to the current time point. Note that the ground-truth scores were ensured to be unseen by the neural decoding model (see Methods).

In terms of the video-task, our real-time system enabled online valence and arousal decoding in each subject with the final onlineCC ranging from 0.32 to 0.62 and 0.25 to 0.45, respectively, which was significantly better than chance level (N=4, n = 32 to 34 per subject; FDR-corrected $P < 0.05$ for each subject; Fig. 8d; Supp. Fig. 30). The real-time decoded traces of valence and arousal closely followed the ground-truth ratings in each subject, qualitatively confirming the above results (Fig. 8d). The population-pooled onlineCC reached 0.51±0.07 and 0.37±0.04 for valence and arousal, respectively (n =132, P<0.001; Fig. 8e), which was comparable with the cvCC achieved in the offline training experiment (n=182, cvCC = 0.53±0.03 and 0.31±0.03 for valence and arousal, respectively; t-test $P > 0.06$ in both cases). Similar results held for online image-task decoding (Fig. 8e; Supp. Figs. 31 and 32). The average latency from streaming iEEG signals to outputting decoded emotion states was 376 ± 11.87 milliseconds across all tasks, subjects, and trials (n=564, see Supp. Table 11 for detailed computational complexity and time). These results demonstrate low-latency, high-performance online decoding of emotion states from iEEG.

Discussion

We developed self-supervised learning models that effectively decoded fine-grained continuous values of the two major dimension of emotion (valence and arousal), whose performance was boosted by integrating grey and white matter iEEG network activity. Such decoding was interchangeable and transferable across two distinct emotion-eliciting tasks, suggesting common neural representations of emotion states elicited by different stimuli. The decoding models further revealed shared and preferred mesolimbic-thalamo-cortical subnetworks encoding valence and arousal and explained the structural connectivity pattern underlying grey and white matter integration. Finally, we implemented the models online and demonstrated robust real-time emotion decoding. Together, these advancements form an integrated framework for intracranial neural activity-based human emotion decoding, explanation, and application.

We used task-based emotion-eliciting paradigms to elicit rapid and controlled emotion responses that allowed for the collection of a much larger number of emotion state measurements than spontaneous emotion assessment paradigms[78]. In terms of stimulus type selection, compared to alternative methods (e.g., recall[79] and hypnosis[80]), audiovisual stimuli offer better standardization and reproducibility[81]. Previous studies have suggested that video-task and image-task had different effects in eliciting emotion—the video-task, with audiovisual inputs and narrative engagement, evokes stronger time-varying valence and arousal responses[82], whereas the image-task, with pure visual input, allows more standardized and faster emotion elicitation[72]—thus, they are suitable for studying cross-task decoding. Overall, the quantity, quality, and diversity of our behavioral data allowed for the training of high-performance cross-task decoding models.

In terms of decoding performance, our model surpassed several related lines of studies. First, several recent iEEG investigations into the spectral dynamics of emotion processing have reported modest linear correlation between spectral features and valence/arousal scores[35,36,59] (on average 0.42±0.02, Supp. Table 12) from an encoding perspective, pointing to both the possibility and challenge of decoding nuanced dynamics of emotion. Second, a series of recent studies have used iEEG to directly decode mood states or depression symptoms with rigorous cross-validation[40–46,60], with a cohort-size-weighted average cvCC=0.57±0.02 (Supp. Table 13). Comparing our decoding performance (e.g., video-task valence cvCC of 0.70±0.02) to these studies should take caution due to the differences in experimental design, behavioral

measurements, and temporal resolution. Task-evoked emotions may exhibit larger and more obvious variations than naturalistic mood fluctuations; on the other hand, our emotion measurement is a single-item rating of valence or arousal, potentially introducing more noise compared to the multi-item questionnaires typically employed in mood or symptom assessments (e.g., Immediate Mood Scaler[83] and Computerized Adaptive Test–Depression Inventory[84]). Mood decoding captures more stable, enduring affective states; by contrast, we decoded more rapid emotion fluctuations, providing distinct insights into the faster dynamics of affect processing as revealed by our temporal interpretation analyses. Third, in terms of the coarse-grained discrete emotion classification, our three-class cvACC of 0.75 (two-class cvACC of 0.81) outperformed previous iEEG emotion classification studies with three-class cvACC being up to 0.68[39] (two-class cvACC of 0.73[38] and 0.62[37]). Finally, consistent with prior studies[11,19], γ band mainly contributed to our decoding, suggesting high-performance emotion decoding may require intensive use of intracranial signals given the limited capacity of EEG and fMRI in capturing high-frequency neural dynamics. Overall, the above comparisons show the high performance of our neural decoding models.

In terms of cross-task decoding, previous iEEG and EEG decoding studies have focused on cross-time and cross-subject generalization, all within a single task[11–13,19,23–25], which did not investigate or sufficiently leverage the central integrative nature of emotion states across diverse tasks. Our capability of interchangeably decoding valence and arousal across the image-task and video-task suggests common neural representations of emotion states across tasks, which motivates transfer-learning across tasks for more efficient deployment of models. Our transferable decoding not only reduced the risk of overfitting by leveraging more data but also indicated the practical utility for ecological emotion decoding applications requiring rapid deployment across diverse contexts[67–69].

Recent advances in intracranial recordings and stimulation have significantly expanded our understanding of the distributed grey matter networks critical for mood and emotion processing[11,12,19]. However, they have been limited by cohort size, region coverage, and decoding accuracy, precluding detailed analyses of the size and spatial distribution of emotion-encoding subnetworks and how they represent distinct dimension of emotion. Our study partially closed the gap by showing that subnetworks of roughly four regions were sufficient for decoding in each subject, consistent with existing studies showing that small subnetworks can encode different aspects of emotion, mood, and depression symptoms[10,11,19,85] and carefully-designed computational methods can efficiently aggregate such information to achieve decoding[39,43–46,60]. The spatial distribution of these key subnetworks for each subject was quite different, which may be due to the different iEEG implantation locations but may also be due to the possible heterogeneity of affect-regulating circuits in each subject, as evidenced by a recent study that had consistent implantations across three depression patients but still found different key regions for decoding symptoms[43].

Nevertheless, our relatively large subject cohort enabled the aggregation of subject-specific top-decoding subnetworks to reveal population-level encoding patterns: common mesolimbic-thalamo-cortical key-encoding regions emerged, which differentially encoded valence and arousal. The shared group mainly consisted of mesolimbic structures (AMYG, HIP, INS, lOFC, mOFC)—a well-known emotion processing network[86,87], and the fronto-parietal network (RMF, PCUN) for task-directed, domain-general cognitive control of emotion regulation[88–90]; alternations of these core circuits have been reported across various affective disorders such as depression and anxiety[85]. The valence-preferred group included more posterior and lateral parts of the frontal and parietal lobes (CMF, IFG, IPL)—likely related to social valence[91], and the anterior cingulate cortex (cACC)—in particular consistent with its role in processing valence-specific information about value and uncertainty[92]. The arousal-preferred group pointed to the outstanding role of the thalamus—a well-known structure critical in awake-sleep[93] and consciousness regularization[94], the posterior part of the cingulate cortex (IC)—which has been implicated as a key structure for both arousal and awareness[95], and the postcentral gyrus known for somatosensory information processing[96]. These findings together delineate a detailed spatial distribution of key regions supporting valence and arousal processing.

Our results further emphasize the importance of white matter networks in human emotion processing by expanding conventional grey matter-focused approaches. Grey matter field potentials have been ascribed to synaptic-induced current flow of local neuron ensembles[97]. By contrast, electrical activity in the white matter is proposed to be generated by the voltage-gated channel current flow in axons connecting widely separated

neuronal populations[98]. Also, white matter signals have smaller amplitudes possibly due to the denser myelin sheaths asymmetrically constraining current flow and signal propagation[99,100]. Therefore, conventional approaches[35–60] have typically treated white matter signals as noise to reduce neural feature complexity. However, in theory, white matter signals can contain useful information as they represent a mixture of activities from nearby and distant grey matter[98]. Consequently, white matter signals have recently garnered increasing attention in fMRI studies[101,102] but remain underexplored in iEEG decoding of brain states, with only a few studies for decoding movements or speech[103–106]. Given our iEEG dataset scale and model design, we were able to show that white matter signals effectively decoded human emotion states, and their combination with grey matter signals significantly improved decoding. Furthermore, we identified both shared and valence/arousal-preferred white matter tracts, mainly involving the uncinate fasciculus, superior/anterior thalamic radiation and cingulum bundle, which have been shown to be associated with affective disorders in prior neuroimaging studies[107]. In addition, these key-encoding white matter tracts mainly connect the limbic system, thalamus, postcentral gyrus, and the fronto-parietal network, which are consistent with the key-encoding grey matter regions identified in our analyses and align with the emerging theory that emotion is encoded by distributed brain networks[108].

The ability to decode human emotion states in real time serves as a translational foundation for next-generation engineering and clinical applications but existing emotion decoding studies have been largely limited to offline analyses, with only a few recent DBS studies achieving real-time computation of single-channel, single-band power feature that correlates with affective disorder symptoms[34,60,109]. Our work moved a step forward by realizing online emotion state decoding, which has the potential of providing real-time feedback signals for guiding interventions of affective disorders. Our task-based decoding paradigm naturally fits into the neurofeedback training framework[110], where the decoded emotion states can directly provide real-time feedback signals for the subject to perform emotion regulation. Our focus on intracranial decoding also closely relates to the development of aDBS therapy for affective disorders such as treatment-resistant depression (TRD)[60] and post-traumatic stress disorder (PTSD)[34]. aDBS aims to use invasive local field potentials to decode naturalistic or event-triggered affective symptom variations as real-time feedback to guide stimulation delivery[67–69,111]. Our transferable decoding framework can provide useful 'pre-trained' models that can be further optimized for such applications. Note that the offline training performance from our online experiments was lower than our main offline experiments because the online recording device had fewer channels and larger noise (Supp. Fig. 33). Thus, incorporating our methods and systems with neurofeedback training and aDBS are promising future directions but require further optimization of model structure, decoding accuracy, implementation stability, and operating latency (see Supp. Note 6).

Our work has several limitations. First, our study cohort consisted exclusively of subjects with epilepsy; while our control analyses excluded the influence of epileptic activity on decoding, the subjects' pathological brain network and electrode implantation heterogeneity can still influence the interpretability of the results. Directly recruiting patients with primary affective disorders and using consistent electrode implantations[37,40,43,59,112,113] offers a possible solution. Second, we only included two emotion-eliciting tasks; cross-task decoding should be validated across a broader range of emotion-eliciting methods, e.g., imagined emotions[79] and immersive virtual reality paradigms[114], to enhance generalizability. Third, we decoded and interpreted valence and arousal as the two major dimension of emotion, but other dimension such as dominance and expectation[115,116] should be considered in the future to formulate a multi-dimensional decoding framework for emotion. Finally, while our subject cohort size and electrode coverage were relatively large when comparing previous iEEG mood and emotion decoding studies, future work should consider incorporating more subjects (especially for the online experiment) and augmenting coverage by simultaneously recording high-density scalp EEG to further improve decoding performance and encoding explainability.

Outlook

Taken as a whole, our results demonstrate high-performance, task-transferable, explainable, and real-time decoding of human emotion states from intracranial neural activity, which can advance emotion decoding and modulation neurotechnology. Moreover, our findings suggest shared and preferred mesolimbic-thalamo-cortical grey matter regions and white matter tracts for representing distinct dimension of emotion, which have implications for understanding network-level neural encoding mechanisms of emotion processing.

Data availability

The main data supporting the results in this study are available within the paper and its Supplementary Information. The collected raw neural and behavioural data are a modified version of clinical recordings for the purpose of seizure localization and clinical decisions. Thus, the minimum de-identified dataset used to generate the findings of this study will be available upon reasonable request to the corresponding author. Processed data for generating the figures will be available upon publication.

Code availability

The iEEG data preprocessing was implemented by MATLAB 2020b custom scripts mainly relying on the Fiel dTrip toolbox. The decoding model construction, fitting, and evaluation were implemented by Python 3.0 custom scripts. All codes will be available upon publication.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grants 62306269, 62476239, and 62336007, the National Key Research and Development Program of China under Grant 2023YFC2506200, the Key R&D Program of Zhejiang under Grant 2025C01112, the Zhejiang Provincial Natural Science Foundation of China under Grant LD24H090001, and the Nanhu Brain-computer Interface Institute under Grant 010901008.

Author contributions

Y.Y., S.H., J.Z. and Y.W. supervised the project. Y.Y. and Y.W. conceived the emotion decoding framework, and Y.Y. and W.C. developed it. Y.Y., W.C., and S.H. designed the experiment. W.C. and Y.C. implemented the experimental system. W.C., Y.C., L.D., and C.Z. performed the experiments and data collection. H.J., Z.Z., X.G., and S.W. identified the electrode position using imaging data and assessed the neurological conditions of the subjects. N.W. and S.H. performed the cognitive and psychiatric assessment. Y.Y. and W.C. implemented and performed the modeling and analyses. Y.Y. and Y.W. supervised all the modelling and analysis work. S.H. and J.Z. supervised all clinical aspects of the study. Y.Y. and W.C. wrote the manuscript with input from Y.W., S.H., J.Z., and G.P.

References

1. Bach, D. R. & Dayan, P. Algorithms for survival: a comparative perspective on emotions. *Nat. Rev. Neurosci.* **18**, 311–319 (2017).
2. Lincoln, T. M., Schulze, L. & Renneberg, B. The role of emotion regulation in the characterization, development and treatment of psychopathology. *Nat. Rev. Psychol.* **1**, 272–286 (2022).
3. Ekman, P. An argument for basic emotions. *Cogn. Emot.* **6**, 169–200 (1992).
4. Panksepp, J. *Affective Neuroscience: The Foundations of Human and Animal Emotions*. (Oxford university press, 2004).
5. Schachter, S. & Singer, J. E. Cognitive, social, and physiological determinants of emotional state. *Psychol. Rev.* **69**, 379–399.
6. Arnold, M. B. *Emotion and Personality*. (1960).
7. Russell, J. A. A circumplex model of affect. *J. Pers. Soc. Psychol.* **39**, 1161–1178 (1980).
8. Barrett, L. F. The theory of constructed emotion: an active inference account of interoception and categorization. *Soc. Cogn. Affect. Neurosci.* **12**, 1–23 (2016).
9. Anderson, D. J. & Adolphs, R. A Framework for Studying Emotions across Species. *Cell* **157**, 187–200 (2014).
10. Malezieux, M., Klein, A. S. & Gogolla, N. Neural Circuits for Emotion. *Annu. Rev. Neurosci.* **46**, 211–231 (2023).
11. Kabotyanski, K. E., Provenza, N. R. & Sheth, S. A. Intracranial neural biomarkers of psychiatric symptoms and their utility for guiding neuromodulation therapy: a systematic review. *Biol. Psychiatry* https://doi.org/10.1016/j.biopsych.2025.10.005 (2025) doi:10.1016/j.biopsych.2025.10.005.
12. Shanechi, M. M. Brain–machine interfaces from motor to mood. *Nat Neurosci* **22**, 1554–1564 (2019).

13. Lim, R. Y., Lew, W.-C. L. & Ang, K. K. Review of EEG Affective Recognition with a Neuroscience Perspective. *Brain Sci.* **14**, 364 (2024).
14. Kragel, P. A. & LaBar, K. S. Decoding the Nature of Emotion in the Brain. *Trends Cogn. Sci.* **20**, 444–455 (2016).
15. Tang, W. *et al.* Emerging Themes and Future Directions in Neurodesign and Human-Computer Interaction: A Systematic Review. *Int. J. Human–Computer Interact.* **41**, 1652–1673 (2025).
16. Rossi, S. Emotion Recognition for Human-Robot Interaction: Recent Advances and Future Perspectives. *Front. Robot. AI* **7**, (2020).
17. Mao, K., Wu, Y. & Chen, J. A systematic review on automated clinical depression diagnosis. *Npj Ment. Health Res.* **2**, 20 (2023).
18. Scangos, K. W., State, M. W., Miller, A. H., Baker, J. T. & Williams, L. M. New and emerging approaches to treat psychiatric disorders. *Nat. Med.* **29**, 317–333 (2023).
19. Ramos Lopez, C. G. *et al.* Stereotactic electroencephalography in epilepsy patients for mapping of neural circuits related to emotional and psychiatric behaviors: a systematic review. *Neurosurg. Focus* **54**, E4 (2023).
20. Siedlecka, E. & Denson, T. F. Experimental Methods for Inducing Basic Emotions: A Qualitative Review. *Emot. Rev.* **11**, 87–97 (2019).
21. Underwood, R., Tolmeijer, E., Wibroe, J., Peters, E. & Mason, L. Networks underpinning emotion: A systematic review and synthesis of functional and effective connectivity. *NeuroImage* **243**, 118486 (2021).
22. Logothetis, N. K. What we can do and what we cannot do with fMRI. *Nature* **453**, 869–878 (2008).
23. Alarcao, S. M. & Fonseca, M. J. Emotions Recognition Using EEG Signals: A Survey. *IEEE Trans. Affect. Comput.* **10**, 374–393 (2019).
24. Wu, D., Lu, B.-L., Hu, B. & Zeng, Z. Affective Brain–Computer Interfaces (aBCIs): A Tutorial. *Proc. IEEE* **111**, 1314–1332 (2023).
25. Wu, D., Xu, Y. & Lu, B.-L. Transfer Learning for EEG-Based Brain–Computer Interfaces: A Review of Progress Made Since 2016. *IEEE Trans. Cogn. Dev. Syst.* **14**, 4–19 (2022).
26. Ding, Y., Zhang, S., Tang, C. & Guan, C. MASA-TCN: Multi-Anchor Space-Aware Temporal Convolutional Neural Networks for Continuous and Discrete EEG Emotion Recognition. *IEEE J. Biomed. Health Inform.* **28**, 3953–3964 (2024).
27. Ding, Y. *et al.* EmT: A Novel Transformer for Generalized Cross-Subject EEG Emotion Recognition. *IEEE Trans. Neural Netw. Learn. Syst.* 1–13 (2025).
28. Boucher, O. *et al.* Spatiotemporal dynamics of affective picture processing revealed by intracranial high-gamma modulations: Intracranial EEG and Affective Pictures. *Hum. Brain Mapp.* **36**, 16–28 (2015).
29. Kauvar, I. *et al.* Conserved brain-wide emergence of emotional response from sensory experience in humans and mice. *Science* **388**, eadt3971 (2025).
30. Fan, X. *et al.* Brain mechanisms underlying the emotion processing bias in treatment-resistant depression. *Nat. Ment. Health* **2**, 583–592 (2024).
31. Pan, N. C. *et al.* Left Insula and Right Middle Temporal Gyrus Dominate Cortical Network Discriminating Arousal‐dependent Emotions. *Adv. Sci.* **12**, 2411790 (2025).
32. Metzger, B. A. *et al.* Intracranial stimulation and EEG feature analysis reveal affective salience network specialization. *Brain* **146**, 4366–4377 (2023).
33. Buot, A. *et al.* Emotions Modulate Subthalamic Nucleus Activity: New Evidence in Obsessive-Compulsive Disorder and Parkinson's Disease Patients. *Biol. Psychiatry Cogn. Neurosci. Neuroimaging* **6**, 556–567 (2021).
34. Gill, J. L. *et al.* A pilot study of closed-loop neuromodulation for treatment-resistant post-traumatic stress disorder. *Nat. Commun.* **14**, 2997 (2023).
35. Huebl, J. *et al.* Processing of emotional stimuli is reflected by modulations of beta band activity in the subgenual anterior cingulate cortex in patients with treatment resistant depression. *Soc. Cogn. Affect. Neurosci.* **11**, 1290–1298 (2016).
36. Sonkusare, S. *et al.* Frequency dependent emotion differentiation and directional coupling in amygdala, orbitofrontal and medial prefrontal cortex network with intracranial recordings. *Mol Psychiatry* 1–11 (2022).
37. Merk, T. *et al.* Invasive neurophysiology and whole brain connectomics for neural decoding in patients with brain implants. *Nat. Biomed. Eng.* 1–18 (2025) doi:10.1038/s41551-025-01467-9.

38. Azadian, E. *et al.* Decoding Happiness from Neural and Video Recordings for Better Insight Into Emotional Processing in the Brain. in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* 6747–6750 (IEEE, Mexico, 2021). doi:10.1109/EMBC46164.2021.9629972.
39. Bijanzadeh, M. *et al.* Decoding naturalistic affective behaviour from spectro-spatial features in multiday human iEEG. *Nat. Hum. Behav.* **6**, 823–836 (2022).
40. Alagapan, S. *et al.* Cingulate dynamics track depression recovery with deep brain stimulation. *Nature* **622**, 130–138 (2023).
41. Frank, A. *et al.* Identification of a personalized intracranial biomarker of depression and response to DBS therapy. *Brain Stimulat.* **16**, 383 (2021).
42. Cao, D. State-specific modulation of mood using intracranial electrical stimulation of the orbitofrontal cortex. *Brain Stimulat.* **16**, 1112–1122 (2023).
43. Xiao, J. *et al.* Decoding Depression Severity From Intracranial Neural Activity. *Biol. Psychiatry* **94**, 445–453 (2023).
44. Sani, O. G. *et al.* Mood variations decoded from multi-site intracranial human brain activity. *Nat Biotechnol* **36**, 954–961 (2018).
45. Kirkby, L. A. *et al.* An Amygdala-Hippocampus Subnetwork that Encodes Variation in Human Mood. *Cell* **175**, 1688-1700.e14 (2018).
46. Scangos, K. W. *et al.* Pilot Study of An Intracranial Electroencephalography Biomarker of Depressive Symptoms in Epilepsy. *J. Neuropsychiatry Clin. Neurosci.* **32**, 185–190 (2020).
47. Voon, V. *et al.* Bed Nucleus of the Stria Terminalis-Nucleus Accumbens Deep Brain Stimulation for Depression: A Randomized Controlled Trial and an Intracranial Physiological Biomarker Predictor. Preprint at https://doi.org/10.21203/rs.3.rs-4854344/v1 (2024).
48. Veerakumar, A. *et al.* Field potential 1/ *f* activity in the subcallosal cingulate region as a candidate signal for monitoring deep brain stimulation for treatment-resistant depression. *J. Neurophysiol.* **122**, 1023–1035 (2019).
49. Sendi, M. S. E. *et al.* Intraoperative neural signals predict rapid antidepressant effects of deep brain stimulation. *Transl. Psychiatry* **11**, 551 (2021).
50. Iravani, B. *et al.* Intracranial Recordings of the Human Orbitofrontal Cortical Activity during Self-Referential Episodic and Valenced Self-Judgments. *J. Neurosci.* **44**, e1634232024 (2024).
51. Hacker, C. *et al.* Aperiodic (1/f) Neural Activity Robustly Tracks Symptom Severity Changes in Treatment-Resistant Depression. *Biol. Psychiatry Cogn. Neurosci. Neuroimaging* **10**, 186–194 (2025).
52. Sonkusare, S. *et al.* Power signatures of habenular neuronal signals in patients with bipolar or unipolar depressive disorders correlate with their disease severity. *Transl. Psychiatry* **12**, 72 (2022).
53. Clark, D. L., Brown, E. C., Ramasubbu, R. & Kiss, Z. H. T. Intrinsic Local Beta Oscillations in the Subgenual Cingulate Relate to Depressive Symptoms in Treatment-Resistant Depression. *Biol. Psychiatry* **80**, e93–e94 (2016).
54. Neumann, W.-J. *et al.* Different patterns of local field potentials from limbic DBS targets in patients with major depressive and obsessive compulsive disorder. *Mol. Psychiatry* **19**, 1186–1192 (2014).
55. Merkl, A. *et al.* Modulation of Beta-Band Activity in the Subgenual Anterior Cingulate Cortex during Emotional Empathy in Treatment-Resistant Depression. *Cereb. Cortex* **26**, 2626–2638 (2016).
56. De Hemptinne, C. *et al.* Prefrontal Physiomarkers of Anxiety and Depression in Parkinson's Disease. *Front. Neurosci.* **15**, 748165 (2021).
57. Rao, V. R. *et al.* Direct Electrical Stimulation of Lateral Orbitofrontal Cortex Acutely Improves Mood in Individuals with Symptoms of Depression. *Curr. Biol.* **28**, 3893-3902.e4 (2018).
58. Zhang, C. *et al.* Bilateral Habenula deep brain stimulation for treatment-resistant depression: clinical findings and electrophysiological features. *Transl. Psychiatry* **12**, 52 (2022).
59. Kabotyanski, K. E. *et al.* Towards objective, temporally resolved neurobehavioral predictors of emotional state. *Brain Stimulat.* **17**, 1208–1212 (2024).
60. Scangos, K. W. *et al.* Closed-loop neuromodulation in an individual with treatment-resistant depression. *Nat. Med.* **27**, 1696–1700 (2021).
61. Landré, E., Chipaux, M., Maillard, L., Szurhaj, W. & Trébuchon, A. Electrophysiological technical procedures. *Neurophysiol. Clin.* **48**, 47–52 (2018).
62. Zhang, S. *et al.* Disruption of superficial white matter in the emotion regulation network in bipolar disorder. *NeuroImage Clin.* **20**, 875–882 (2018).

63. Forkel, S. J., Friedrich, P., Thiebaut De Schotten, M. & Howells, H. White matter variability, cognition, and disorders: a systematic review. *Brain Struct. Funct.* **227**, 529–544 (2022).
64. Silversmith, D. B. *et al.* Plug-and-play control of a brain–computer interface through neural map stabilization. *Nat. Biotechnol.* **39**, 326–335 (2021).
65. Spitale, M. *et al.* Past, Present, and Future: A Survey of The Evolution of Affective Robotics For Well-being. *IEEE Trans. Affect. Comput.* 1–17 (2025).
66. Fernández-Alvarez, J. *et al.* Efficacy of bio- and neurofeedback for depression: a meta-analysis. *Psychol. Med.* **52**, 201–216 (2022).
67. Oganesian, L. L. & Shanechi, M. M. Brain–computer interfaces for neuropsychiatric disorders. *Nat. Rev. Bioeng.* **2**, 653–670 (2024).
68. Sellers, K. K. *et al.* Closed-loop neurostimulation for the treatment of psychiatric disorders. *Neuropsychopharmacology* **49**, 163–178 (2024).
69. Herron, J. *et al.* Challenges and opportunities of acquiring cortical recordings for chronic adaptive deep brain stimulation. *Nat. Biomed. Eng.* **9**, 606–617 (2024).
70. Johnson, K. A., Okun, M. S., Scangos, K. W., Mayberg, H. S. & De Hemptinne, C. Deep brain stimulation for refractory major depressive disorder: a comprehensive review. *Mol. Psychiatry* **29**, 1075–1087 (2024).
71. Oehrn, C. R. *et al.* Chronic adaptive deep brain stimulation versus conventional stimulation in Parkinson's disease: a blinded randomized feasibility trial. *Nat. Med.* **30**, 3345–3356 (2024).
72. Branco, D., Gonçalves, Ó. F. & Badia, S. B. I. A Systematic Review of International Affective Picture System (IAPS) around the World. *Sensors* **23**, 3866 (2023).
73. Klein, A. & Tourville, J. 101 Labeled Brain Images and a Consistent Human Cortical Labeling Protocol. *Front. Neurosci.* **6**, (2012).
74. Fischl, B. *et al.* Whole Brain Segmentation: Automated Labeling of Neuroanatomical Structures in the Human Brain. *Neuron* **33**, 341–355 (2002).
75. Van Essen, D. C. *et al.* The Human Connectome Project: A data acquisition perspective. *NeuroImage* **62**, 2222–2231 (2012).
76. Schneider, S., Lee, J. H. & Mathis, M. W. Learnable latent embeddings for joint behavioural and neural analysis. *Nature* **617**, 360–368 (2023).
77. Gunnarsdottir, K. M. Source-sink connectivity: a novel interictal EEG marker for seizure localization. *Brain* **145**, 3901–3915.
78. Ciuk, D., Troy, A. K. & Jones, M. C. Measuring Emotion: Self-Reports vs. Physiological Indicators. *SSRN Electron. J.* (2015).
79. Chanel, G., Kierkels, J. J. M., Soleymani, M. & Pun, T. Short-term emotion assessment in a recall paradigm. *Int. J. Hum.-Comput. Stud.* **67**, 607–627 (2009).
80. Houghton, L. A. Visceral sensation and emotion: a study using hypnosis. *Gut* **51**, 701–704 (2002).
81. Gerdes, A. B. M., Wieser, M. J. & Alpers, G. W. Emotional pictures and sounds: a review of multimodal interactions of emotion cues in multiple domains. *Front. Psychol.* **5**, (2014).
82. Uhrig, M. K. *et al.* Emotion Elicitation: A Comparison of Pictures and Films. *Front. Psychol.* **7**, (2016).
83. Nahum, M. *et al.* Immediate Mood Scaler: Tracking Symptoms of Depression and Anxiety Using a Novel Mobile Mood Scale. *JMIR MHealth UHealth* **5**, e6544 (2017).
84. Gibbons, R. D. *et al.* Development of a Computerized Adaptive Test for Depression. *Arch. Gen. Psychiatry* **69**, 1104 (2012).
85. Williams, L. M. Precision psychiatry: a neural circuit taxonomy for depression and anxiety. *Lancet Psychiatry* **3**, 472–480 (2016).
86. Rolls, E. T. Limbic systems for emotion and for memory, but no single limbic system. *Cortex* **62**, 119–157 (2015).
87. Manssuer, L. *et al.* Reward recalibrates rule representations in human amygdala and hippocampus intracranial recordings. *Nat. Commun.* **15**, 9518 (2024).
88. Cole, M. W., Repovš, G. & Anticevic, A. The Frontoparietal Control System: A Central Role in Mental Health. *The Neuroscientist* **20**, 652–664 (2014).
89. Taylor, C. *et al.* An Exploratory Study of Large-Scale Brain Networks during Gambling Using SEEG. *Brain Sci.* **14**, 773 (2024).
90. Hoy, C. W. *et al.* Beta and theta oscillations track effort and previous reward in the human basal ganglia and prefrontal cortex during decision making. *Proc. Natl. Acad. Sci.* **121**, e2322869121 (2024).

91. Padilla-Coreano, N., Tye, K. M. & Zelikowsky, M. Dynamic influences on the neural encoding of social valence. *Nat. Rev. Neurosci.* **23**, 535–550 (2022).
92. Monosov, I. E. Anterior cingulate is a source of valence-specific information about value and uncertainty. *Nat. Commun.* **8**, 134 (2017).
93. Gent, T. C., Bandarabadi, M., Herrera, C. G. & Adamantidis, A. R. Thalamic dual control of sleep and wakefulness. *Nat. Neurosci.* **21**, 974–984 (2018).
94. Fang, Z. *et al.* Human high-order thalamic nuclei gate conscious perception through the thalamofrontal loop. *Science* **388**, eadr3675 (2025).
95. Leech, R. & Sharp, D. J. The role of the posterior cingulate cortex in cognition and disease. *Brain* **137**, 12–32 (2014).
96. Anders, S., Lotze, M., Erb, M., Grodd, W. & Birbaumer, N. Brain activity underlying emotional valence and arousal: A response‐related fMRI study. *Hum. Brain Mapp.* **23**, 200–209 (2004).
97. Buzsáki, G., Anastassiou, C. A. & Koch, C. The origin of extracellular fields and currents — EEG, ECoG, LFP and spikes. *Nat. Rev. Neurosci.* **13**, 407–420 (2012).
98. Mercier, M. R. *et al.* Evaluation of cortical local field potential diffusion in stereotactic electro-encephalography recordings: A glimpse on white matter signal. *NeuroImage* **147**, 219–232 (2017).
99. Kajikawa, Y. & Schroeder, C. E. Generation of field potentials and modulation of their dynamics through volume integration of cortical activity. *J. Neurophysiol.* **113**, 339–351 (2015).
100. Greene, P., Li, A., González-Martínez, J. & Sarma, S. V. Classification of Stereo-EEG Contacts in White Matter vs. Gray Matter Using Recorded Activity. *Front. Neurol.* **11**, 605696 (2021).
101. Berluti, K., Ploe, M. L. & Marsh, A. A. Emotion processing in youths with conduct problems: an fMRI meta-analysis. *Transl. Psychiatry* **13**, 105 (2023).
102. Seyedmirzaei, H. *et al.* White matter tracts associated with alexithymia and emotion regulation: A diffusion MRI study. *J. Affect. Disord.* **314**, 271–280 (2022).
103. Li, G. *et al.* Detection of human white matter activation and evaluation of its function in movement decoding using stereo-electroencephalography (SEEG). *J. Neural Eng.* **18**, 0460c6 (2021).
104. Bouton, C. *et al.* Decoding Neural Activity in Sulcal and White Matter Areas of the Brain to Accurately Predict Individual Finger Movement and Tactile Stimuli of the Human Hand. *Front. Neurosci.* **15**, 699631 (2021).
105. Ottenhoff, M. C. *et al.* Decoding continuous goal-directed movement from human brain-wide intracranial recordings. Preprint at https://doi.org/10.1101/2025.02.05.636287 (2025).
106. Soroush, P. Z. *et al.* Contributions of Stereotactic EEG Electrodes in Grey and White Matter to Speech Activity Detection. in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* 4789–4792 (IEEE, Glasgow, Scotland, United Kingdom, 2022).
107. Koshiyama, D. *et al.* White matter microstructural alterations across four major psychiatric disorders: mega-analysis study in 2937 individuals. *Mol. Psychiatry* **25**, 883–895 (2020).
108. LeDuke, D. O., Borio, M., Miranda, R. & Tye, K. M. Anxiety and depression: A top‐down, bottom‐up model of circuit function. *Ann. N. Y. Acad. Sci.* **1525**, 70–87 (2023).
109. Nho, Y.-H. *et al.* Responsive deep brain stimulation guided by ventral striatal electrophysiology of obsession durably ameliorates compulsion. *Neuron* **112**, 73-83.e4 (2024).
110. Sitaram, R. *et al.* Closed-loop brain training: the science of neurofeedback. *Nat. Rev. Neurosci.* **18**, 86–100 (2017).
111. Yang, Y. *et al.* Modelling and prediction of the dynamic responses of large-scale brain networks during direct electrical stimulation. *Nat Biomed Eng* **5**, 324–345 (2021).
112. Liu, S. *et al.* A Habenula Neural Biomarker Simultaneously Tracks Weekly and Daily Symptom Variations during Deep Brain Stimulation Therapy for Depression. *IEEE J. Biomed. Health Inform.* 1–14 (2025) doi:10.1109/JBHI.2025.3566057.
113. Sellers, K. Towards individualized deep brain stimulation: A stereoencephalography-based workflow for unbiased neurostimulation target identification. *bioRxiv* (2025).
114. Somarathna, R., Bednarz, T. & Mohammadi, G. Virtual Reality for Emotion Elicitation – A Review. *IEEE Trans. Affect. Comput.* **14**, 2626–2645 (2023).
115. Mehrabian, A. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in Temperament. *Curr. Psychol.* **14**, 261–292 (1996).

116. Fontaine, J. R. J., Scherer, K. R., Roesch, E. B. & Ellsworth, P. C. The World of Emotions is not Two-Dimensional. *Psychol. Sci.* **18**, 1050–1057 (2007).
117. Fischl, B. FreeSurfer. *NeuroImage* **62**, 774–781 (2012).
118. Tadel, F., Baillet, S., Mosher, J. C., Pantazis, D. & Leahy, R. M. Brainstorm: A User-Friendly Application for MEG/EEG Analysis. *Comput. Intell. Neurosci.* **2011**, 1–13 (2011).
119. The Qt Company. Qt6. (2022).
120. Kleiner, M. What's new in Psychtoolbox-3? (2007).
121. Lang, P.J., Bradley, M.M., & Cuthbert, B.N. International affective picture system (IAPS): Technical manual and affective ratings. *NIMH Cent. Study Emot. Atten.* **3**, 39–58 (1997).
122. Dan-Glauser, E. S. & Scherer, K. R. The Geneva affective picture database (GAPED): a new 730-picture database focusing on valence and normative significance. *Behav. Res. Methods* **43**, 468–477 (2011).
123. Pengfei, X., Yuxia, H. & Yuejia, L. Establishment and assessment of native Chinese affective video system. *Chin. Ment. Health J.* **24**, (2010).
124. Chen, W., Wang, Y. & Yang, Y. Efficient Estimation of Directed Connectivity in Nonlinear and Nonstationary Spiking Neuron Networks. *IEEE Trans. Biomed. Eng.* **71**, 841–854 (2024).
125. Ian Goodfellow, Yoshua Bengio, & Aaron Courville. *Deep Learning*. (MIT Press, 2016).
126. Cawley, G. C. & Talbot, N. L. C. On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *The Journal of Machine Learning Research* 2079–2107 (2010).
127. Maaten, L. & Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
128. Maulik, U. & Bandyopadhyay, S. Performance evaluation of some clustering algorithms and validity indices. *IEEE Trans. Pattern Anal. Mach. Intell.* **24**, 1650–1654 (2002).
129. Yeh, F.-C. Population-based tract-to-region connectome of the human brain and its hierarchical topology. *Nat. Commun.* **13**, 4933 (2022).
130. Yeh, F.-C., Verstynen, T. D., Wang, Y., Fernández-Miranda, J. C. & Tseng, W.-Y. I. Deterministic Diffusion Fiber Tracking Improved by Quantitative Anisotropy. *PLoS ONE* **8**, e80713 (2013).
131. Cohen, M. X. *Analyzing Neural Time Series Data: Theory and Practice*. (The MIT Press, Cambridge, Massachusetts, 2014).
132. FFmpeg Developers. FFmpeg. (2024).

Figures and captions

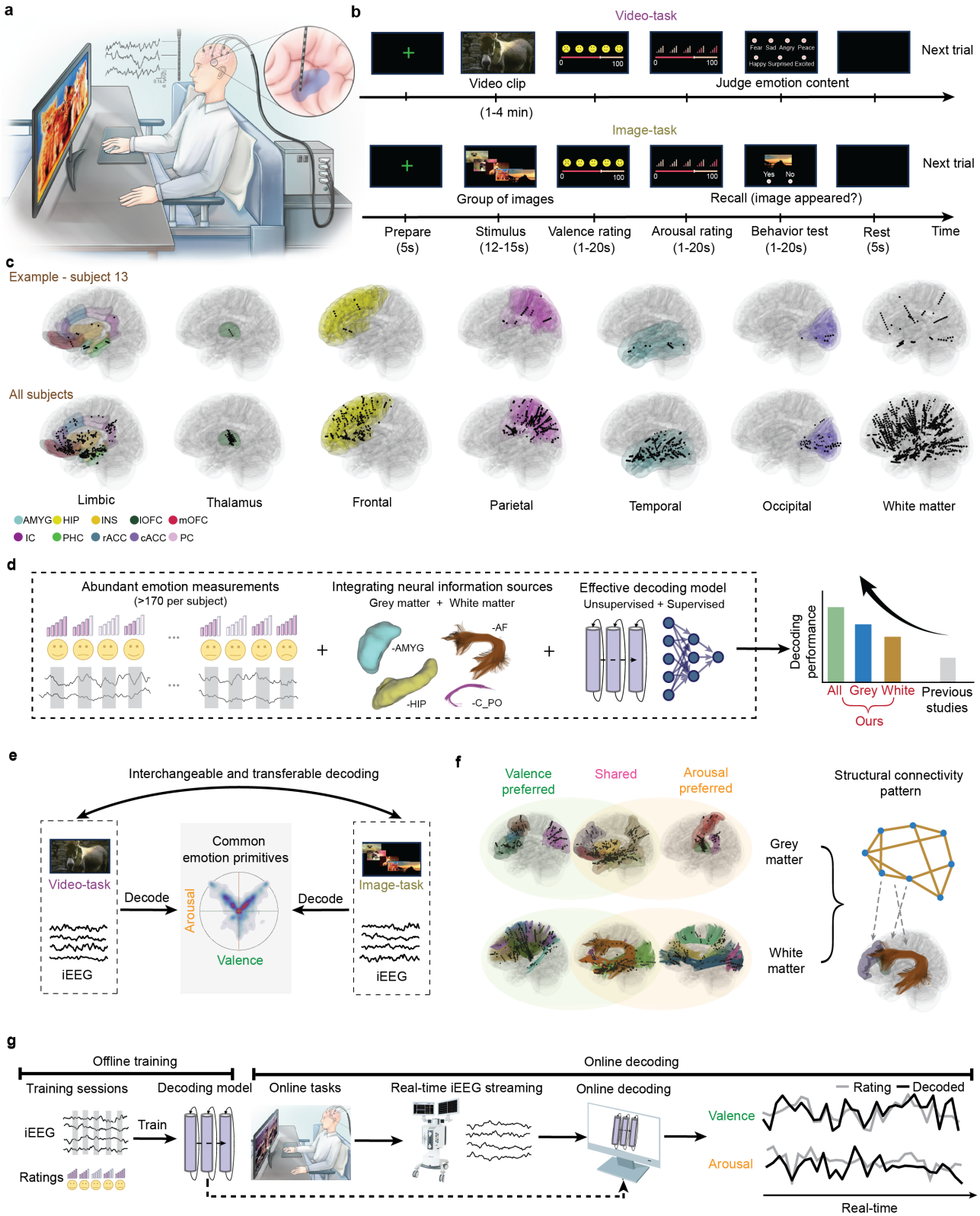

Fig. 1 | Overview of experimental design, neural recording, and decoding analyses. a, Experimental setup for emotion elicitation and iEEG recording. Subjects are instructed to view image or video stimuli and self-report valence and arousal ratings. iEEG signals are continuously recorded. **b**, Schematic diagrams of example trials of the video-task and image-task (see Methods for details). **c**, Top, iEEG contact locations in a representative subject. Bottom, summary of all iEEG contact locations across the population. Each black dot represents a grey matter recording contact overlaid on the DKT and ASEG atlases with different colors representing different grey matter regions. Each black dot without color overlay represents a white matter recording contact. See Supp. Tables 6 and 7 for region name abbreviations. **d-g**, Summary of the main

analyses and findings. **d,** We collected abundant emotion measurements as self-report valence and arousal ratings, took both the grey matter and white matter as useful neural information sources, and designed specialized deep learning models to integrate such information, together leading to high-performance decoding. **e**, Our model enabled interchangeable and transferable decoding of the common neural representations of emotion primitives across the two tasks. **f**, Our decoding models had sufficient encoding explainability, revealing shared and preferred grey and white matter subnetworks encoding valence and arousal and explaining the structural connectivity pattern that supports grey and white matter decoding coordination. **g**, New subjects were instructed to perform image-tasks and video-tasks online while iEEG data were streamed to a fixed, offline-trained decoding model for real-time prediction of valence and arousal states.

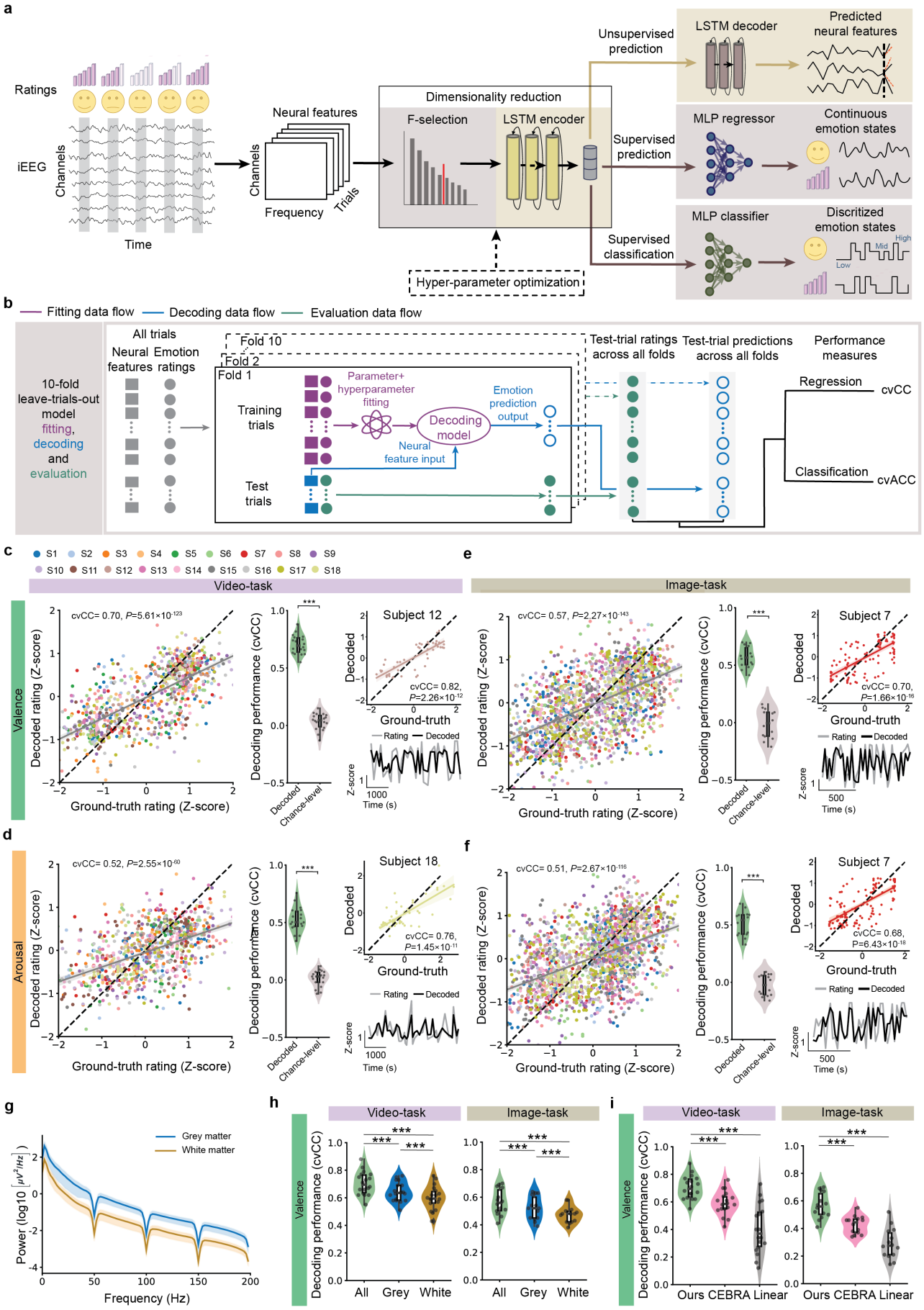

Fig. 2 | High-performance single-task decoding of valence and arousal states by personalized self-supervised learning models. a, The data processing pipeline and structure of the neural decoding model,

which mainly consists of unsupervised dimensionality reduction by an LSTM autoencoder and supervised learning by an MLP regressor/classifier. **b**, The 10-fold leave-trials-out model fitting, decoding, and evaluation procedure. Note that the model hyperparameters are also rigorously cross-validated. **c**, Cross-validated valence decoding results in the video-task. Left, the population-pooled cross-validated regression decoding results, where distinct colors represent different subjects (the colors remain consistent for all following analyses), each dot represents the decoded rating of a single trial (n=845), the grey line represents linear fitting of all decoding results with the shaded area representing the 95% confidence interval, and the dotted black line indicates the 45-degree reference line of perfect decoding. Same plotting format holds for subsequent linear fitting. Note that to pool the results across subjects, the ground-truth ratings (horizontal axis) and decoded ratings (vertical axis) are both z-scored by the distribution of the ground-truth ratings within each subject. Middle, the pooled decoding performance significantly outperformed chance-level (N=18). In each violin plot, each dot represents a subject. The width of the shape represents the data distribution density. The inner central line represents the median, the box spans the interquartile range (IQR), and whiskers extend to 1.5×IQR. n.s.: P>0.05; *: P<0.05; **: P<0.005; ***: P<0.0005. Subsequent violin plots follow the same format. Right top, example decoding results of a high-performance subject (see Supp. Fig. 11 for other subjects). Right bottom, the corresponding temporal traces of decoded ratings and ground-truth ratings for the example subject. **d,** Same as **c** but for cross-validated arousal decoding results in the video-task. **e-f**, Same as **c-d** but for the image-task (N=14, n=1736). **g**, The average spectral power of pre-processed grey matter and white matter iEEG signals across all contacts and all subjects. **h**, Population-level valence decoding performance comparison in the video-task and image-task of (1) using all iEEG contacts; (2) using only iEEG contacts in grey matter; (3) using only iEEG contacts in white matter (see Supp. Fig. 14 for arousal results). **i**, Valence decoding performance comparison of our model with the deep learning-based CEBRA decoding model and the linear control model (see Supp. Fig. 12 for arousal results).

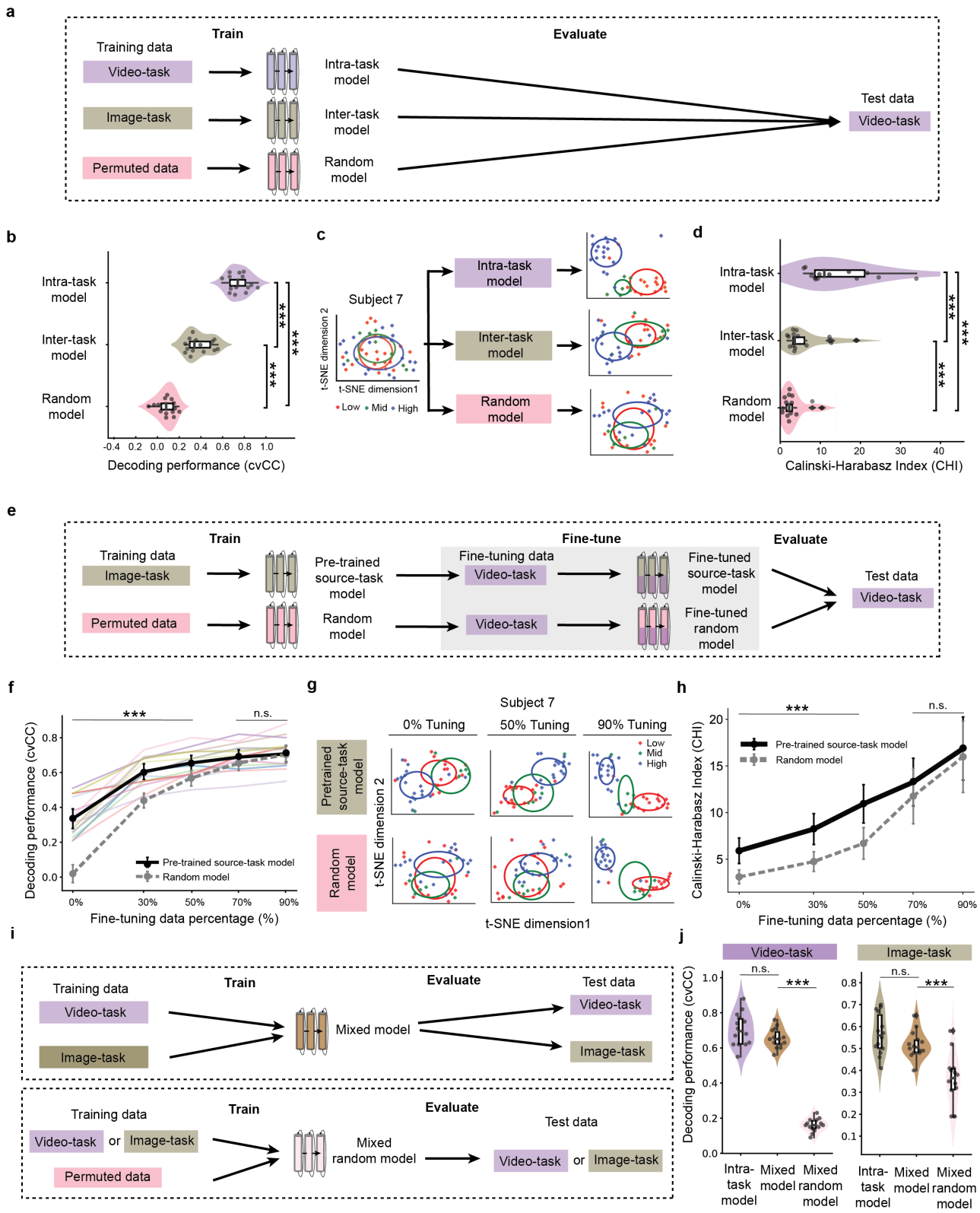

Fig. 3 | Interchangeable and transferable decoding across distinct emotion-eliciting tasks. a, The interchangeable decoding paradigm where image-task is taken as the source-task and video-task is taken as the target-task for an example. An intra-task model, an inter-task model, and a random model are separately trained and then evaluated using the same target-task test data. **b**, Decoding performance comparison of the three models in **a** across the population. **c**, t-SNE visualization comparison of the three models in **a** for an example subject (see Supp. Fig. 15 for other subjects). Left, 2-dimensional t-SNE visualization of input neural features before decoding, with dots representing neural feature samples, colors representing low, median, and high valence groups, and ellipses representing Gaussian clustering boundaries. Right, t-SNE visualization of the neural embeddings extracted by the three models. **d**, Quantitative comparison of

clustering efficacy. Higher CHI values represent better cluster separation. **e**, The transferable decoding paradigm where image-task is taken as the source-task and video-task is taken as the target-task as an example. A 'pretrained source-task' model and a random model are fine-tuned by different amount of target-task training data and then evaluated by the same target-task test data. **f**, Decoding performance comparison between fine-tuning the two models in **e** with varying proportions of target-task training data. Each colored line represents a single subject, and the solid black line represents the mean with whiskers representing the 95% confidence interval. Subsequent error bar plots follow the same format. **g**, 2-dimensional t-SNE visualizations of the fine-tuning process in a single example subject (see Supp. Fig. 18 for other subjects). **h**, CHI clustering efficacy of the fine-tuning process. **i**, The mixed decoding paradigm where the mixed model is trained using combined video-task and image-task training data and evaluated separately by the video-task or image-task test data. The bottom shows the mixed random control analysis. **j,** Decoding performance comparison between the intra-task model, the mixed model, and mixed random control in **i**.

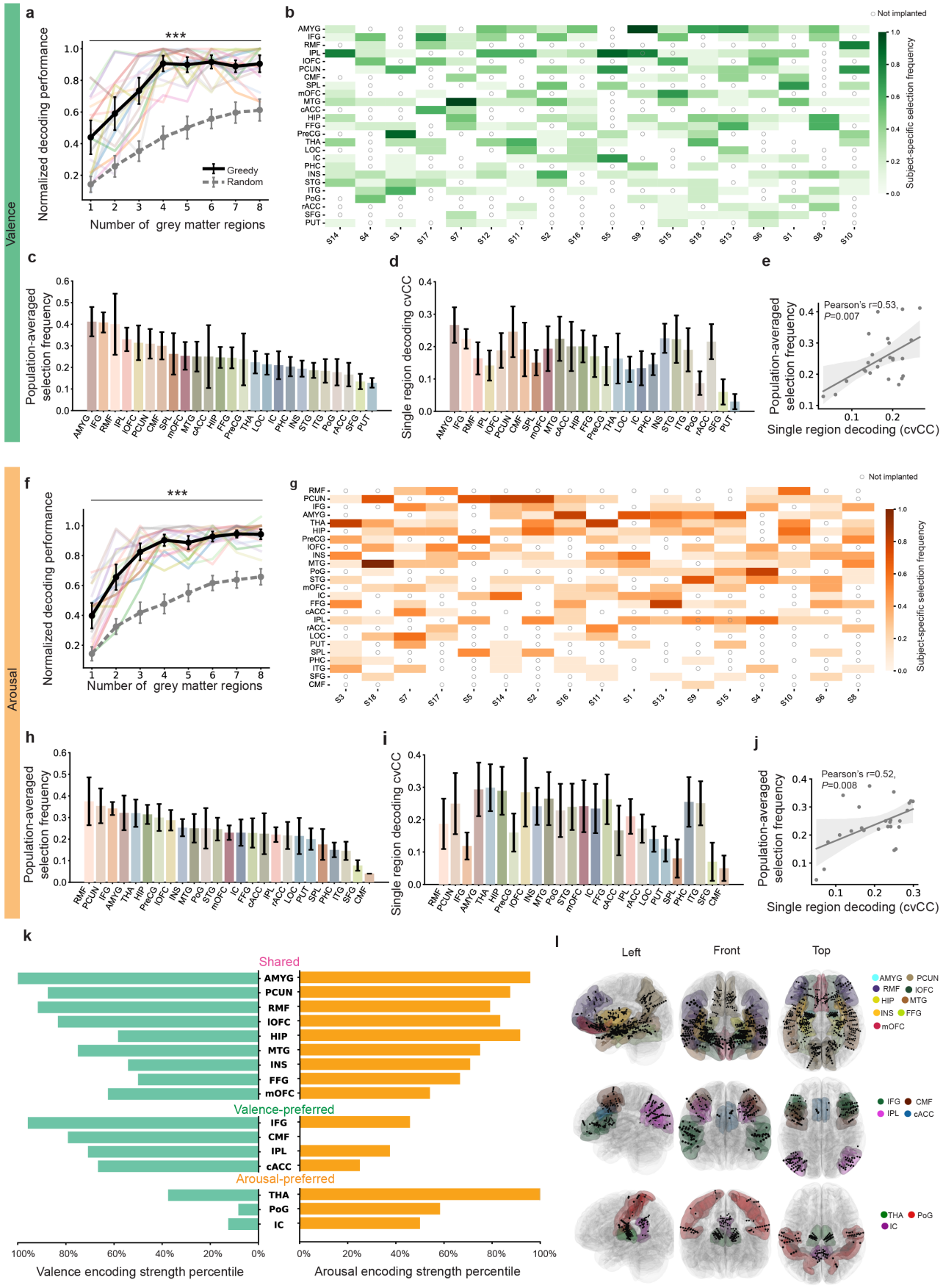

Fig. 4 | Grey matter regions differentially encoding valence and arousal. a, Valence decoding performance of models trained with an increasing number of grey matter regions as selected by our nested cross-validation procedure and greedy algorithm. Each color represents a single subject where the decoding

cvCC is normalized by dividing the cvCC of using all implanted grey matter regions. The solid black line represents population average. The grey line represents chance-level performance of randomly selecting regions for comparison. **b**, The distribution of subject-specific top-decoding subnetworks (size of four) across all cross-validation folds. Each region's frequency being selected as part of the top-decoding subnetwork across all cross-validation folds is visualized separately for each subject. The regions were ranked by population-averaged selection frequency (consistent with panel **c** for better comparative visualization) from top to bottom. The subjects were ranked by valence decoding performance from left to right. The white color represents other implanted brain regions not being selected in the top-decoding subnetworks in any fold. The circle represents brain regions not being implanted in the subject. See Supp. Table 6 for region name abbreviations. **c**, Selection frequency of each grey matter region averaged across all subjects. **d**, Single-region valence decoding performance across all subjects. **e**, Pearson correlation between the population-averaged region selection frequency in **c** and the single-region valence decoding performance in **e**, with each dot representing a grey matter region and the solid line representing the linear fitting. Population-level key-encoding grey matter regions are defined by combining **c** and **d** as detailed in Methods. **f–j**, Subject-specific top-decoding subnetworks and population-level key-encoding grey matter regions for arousal, where panel captions are similar with panels **a**–**e**. **k**, Division of the population-level key-encoding grey matter regions into the shared group, the valence-preferred group, and the arousal-preferred group. **l**, Anatomical visualization of the corresponding key-encoding grey matter regions (shaded colors) and electrode implantation locations (black dots) for each group in panel **k** across all subjects. From top to bottom: shared, valence-preferred, and arousal-preferred.

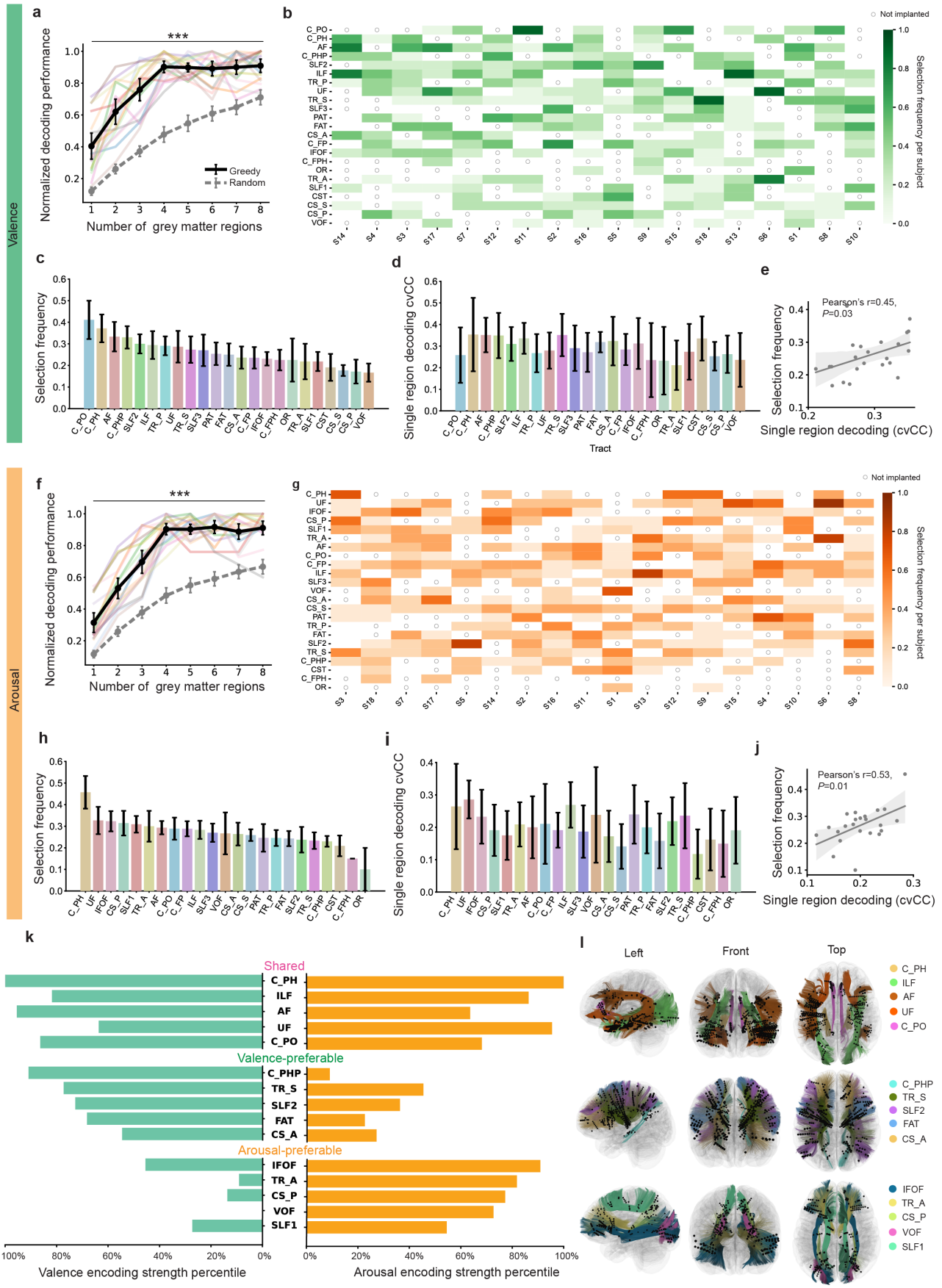

Fig. 5 | White matter tracts differentially encoding valence and arousal. a-l Panel captions are the same as Fig. 4 but for key white matter tract distributions. See Supp. Table 7 for white matter tract name abbreviations.

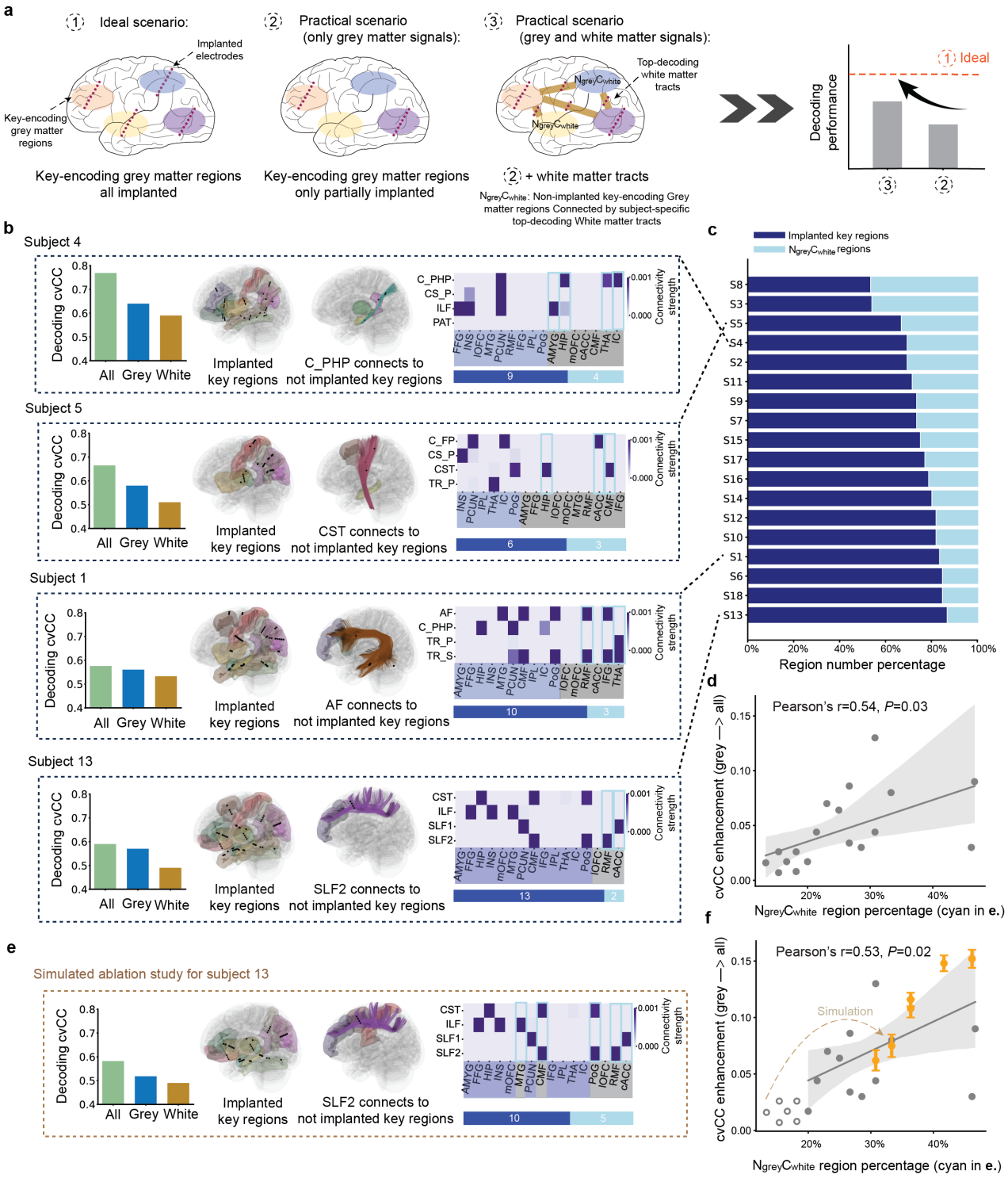

Fig. 6 | Structural connectivity pattern explaining grey and white matter integration. a, The illustration of the hypothesis that the boosted decoding performance by integrating grey and white matter is facilitated by network-level structural connectivity: (1) the ideal scenario that all key-encoding grey matter regions were implanted in a given subject; (2) the practical scenario (if only consider grey matter signals) that only a subset of key-encoding grey matter regions were implanted; (3) the practical scenario (if consider both grey matter and white matter signals) that subject-specific top-decoding white matter tracts had connections with not only key-encoding grey matter regions implanted in that subject but also non-implanted key grey matter regions ($N_{grey}C_{white}$ regions), bringing in complementary information that benefits decoding. **b**, Each row represents a case study, from left to right: (1) valence decoding performance comparison; (2) visualization of the key-encoding grey matter regions implanted in that subject; (3) visualization of one example subject-specific top-decoding white matter tract that connects to not-implanted key-encoding grey matter regions; (4)

structural connectivity strength among the four top-decoding white matter tracts identified within the example subject (vertical axis) and the population-level key-encoding grey matter regions (horizontal axis). The dark-blue horizontal axis labels represent the implanted key-encoding grey matter regions (the number of such regions are denoted underneath). The grey horizontal axis labels represent the not-implanted key-encoding grey matter regions. The light-blue boxes represent the 'Not-implanted key-encoding Grey matter regions that are Connected to the subject-specific top-decoding White matter tracts' ($\mathrm{N_{grey}C_{white}}$ regions, the number of such regions is denoted underneath). **c**, The percentage of $\mathrm{N_{grey}C_{white}}$ regions among all involved regions (implanted key-encoding regions plus $\mathrm{N_{grey}C_{white}}$ regions) for each subject. **d**, Pearson's correlation between the percentage of $\mathrm{N_{grey}C_{white}}$ regions and the cvCC enhancement from grey-matter-only decoding to all-contact decoding, with each dot representing a subject. **e**, same as **b**, but for the simulated ablation study. **f**, same as **d**, but for the simulated ablation study. Open circles denote original data replaced by simulation and the yellow dots indicate corresponding results from simulation.

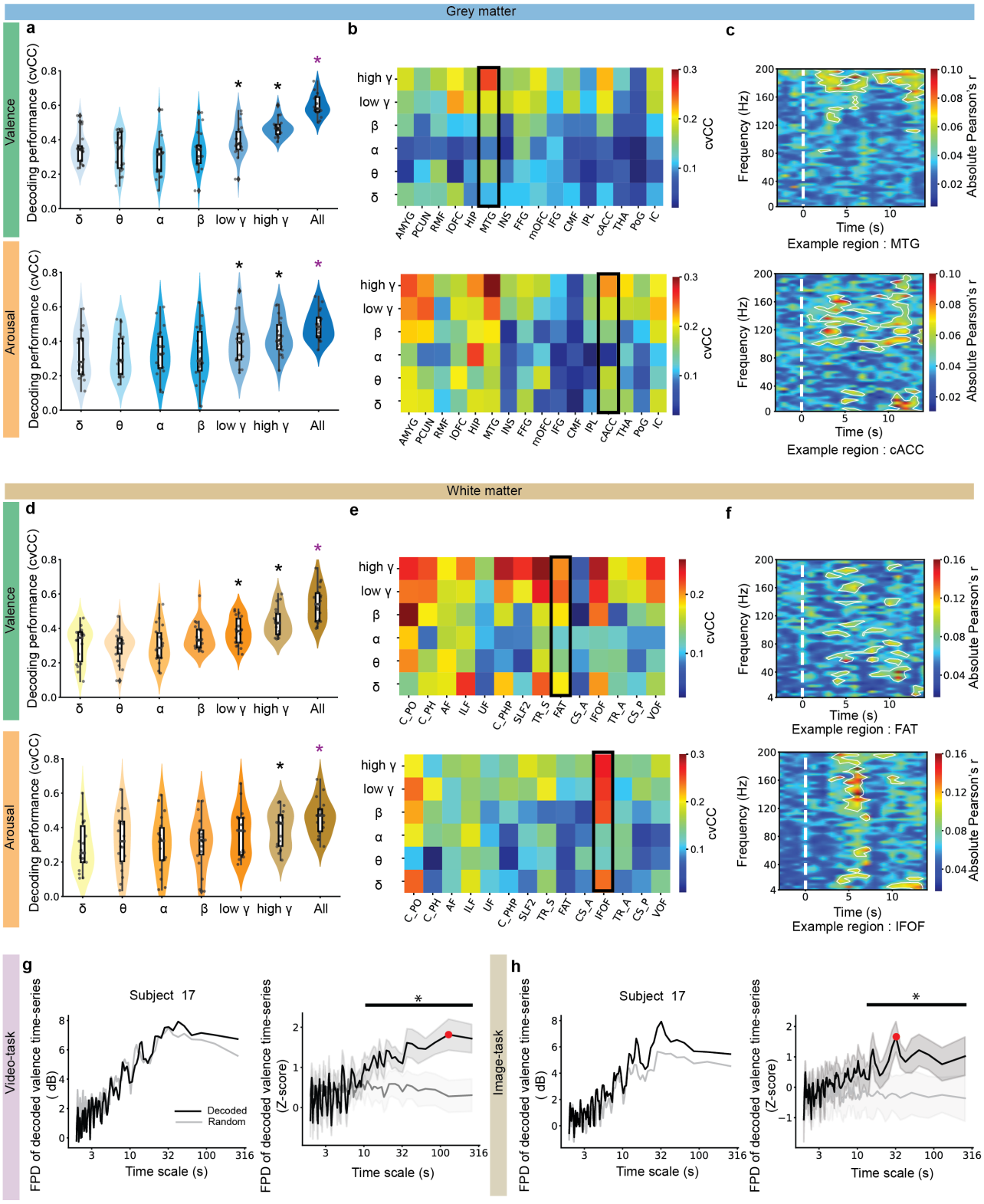

Fig. 7 | Spectral and temporal interpretation of emotion decoding. a-f, Spectral interpretation of grey matter and white matter decoding. **a**, Top: valence decoding performance of only using a single frequency band, where each dot represents a subject. The black asterisks represent that the decoding performance of γ bands significantly outperform the other lower frequency bands (FDR-corrected $P < 0.043$). The purple asterisk represents that the decoding performance when using all frequency bands significantly outperforms using each single frequency band alone (FDR-corrected $P < 0.008$). Bottom: same as top but for arousal decoding. **b**, Top: The valence decoding performance of each frequency band for each key-encoding grey matter region as in Fig. 4k, where the black box highlights the example region shown in **c**. Bottom: same as

top but for arousal decoding. **c**, Top: Correlation between the time-frequency power and valence ratings time-locked to the starting of each image-task trial for an example brain region. The color represents the absolute value of Pearson's correlation coefficient between the spectral power and the valence rating across all trials and subjects. The dotted white line represents the time that emotion-eliciting material appeared. The white curves highlight the significant Pearson's correlation coefficient clusters compared with the permutation baseline (FDR-corrected $P<0.05$). An example region of obvious γ activation is shown, see Supp. Fig. 27 for all regions. Bottom: same as top but for arousal decoding. **d-f,** Same as **a-c**, but for white matter decoding, and see Supp. Fig. 28 for all tracts. **g,** Left: the frequency-domain power density (FPD) of decoded valence time-series (black line) and random-decoded valence time-series (grey line) during the video-task for one example subject. The horizontal axis represents the timescale (inverse of frequency). Right: the normalized FPD of decoded valence time-series (each subject's FPD is *z*-scored by the subject's own random-decoding distribution), with the solid black line representing the average across all subjects. The grey represents random decoding. The red dot represents the peaking FPD value. **h,** Same as **g** but for the image-task.

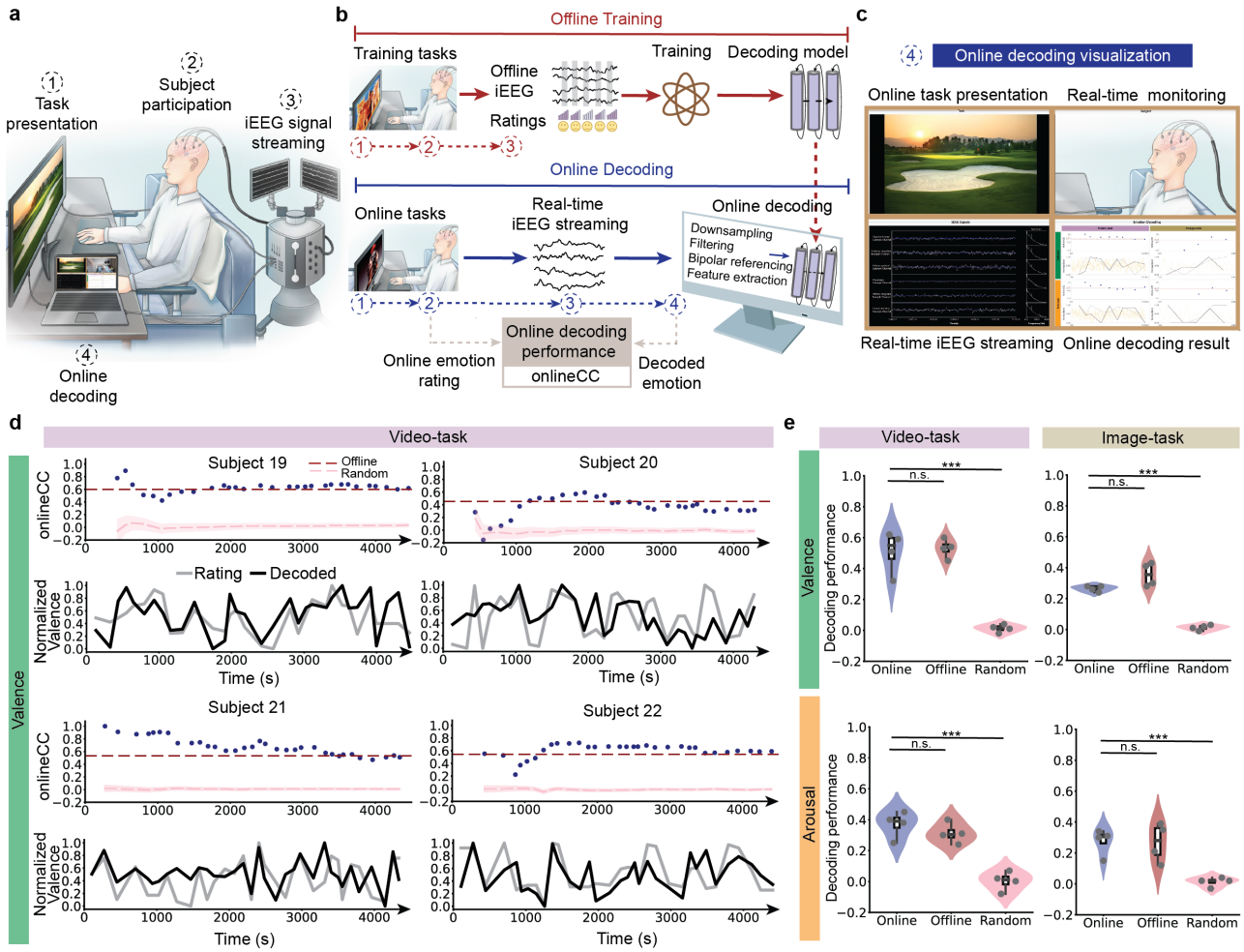

Fig. 8 | Online implementation of neural decoding models and robust real-time emotion decoding. a, The online emotion decoding experiment setup, which involves four key components as indicated by dotted circles. **b**, In each subject, the neural decoding model is trained in a separate offline training experiment (top) one day before the online decoding experiment (bottom). **c**, The online decoding visualization software streams the online task presentations, monitors the subject's facial expressions, displays the real-time iEEG signals, and shows the online decoding results, which are all synchronized (see Supp. Video 1). **d**, The online valence decoding performance for each subject in the video-task (see Supp. Figs. 30-32 for other conditions). Top: the onlineCC values over time. The blue points indicate the onlineCC computed at each time point once the subject rated a valence score. Note that the onlineCC was updated by computing the correlation coefficient between all completed online ratings and decoding values up to the current time point. The onlineCC became stable after completing around 10 ratings, where enough samples had been accumulated for computing a reasonable correlation coefficient. The red dotted lines show the cvCC achieved by the offline training data. The pink dotted lines display the cvCC achieved by the offline training data with permuted labels, serving as the chance-level. Bottom: the decoded valence traces shown at each time point when the subject rated a valence score during the online experiments. **e**, The population pooled decoding performance comparison of online decoding, offline training, and chance-level. Each grey dot represents a subject (*N*=4).