Supplementary Tables

Table S1. Demographic information of twins in predicted models.

Clinical characteristics	Age at scan	Number (male / female)
HCP	22-35 years (29.17 ± 3.31)	184 pairs (158 / 210)
QTIM	12-21 years (18.11 ± 2.85)	245 pairs (205 / 285)

Table S2. Statistical results of multidimensional behavioral similarity across twins and unrelated cohorts.

Domains	MZ vs. UN_m (Permutation test)	DZ vs. UN_d (Permutation test)	MZ vs. DZ (Two sample t test)
Alertness	$z = 2.98$, $p < 0.01$	$z = 2.18$, $p < 0.05$	$t = 0.21$, $p = 0.83$
Cognition	$z = 5.67$, $p < 0.001$	$z = 2.46$, $p < 0.01$	$t = 1.76$, $p = 0.07$
Emotion	$z = 4.42$, $p < 0.001$	$z = 4.24$, $p < 0.001$	$t = 0.06$, $p = 0.95$
Motor	$z = 6.64$, $p < 0.001$	$z = 3.44$, $p < 0.001$	$t = 1.63$, $p = 0.10$
Personality	$z = 10.69$, $p < 0.001$	$z = 5.03$, $p < 0.001$	$t = 3.36$, $p < 0.001$
Sensory	$z = 3.91$, $p < 0.001$	$z = 3.03$, $p < 0.01$	$t = 0.09$, $p = 0.93$

z represents the permutation-based standardized score derived from 5,000 iterations. *t* denotes the parametric independent two-sample t-test. MZ, monozygotic twins; DZ, dizygotic twins; UN_m, unrelated pairs generated matched to the MZ group; UN_d, unrelated pairs matched to the DZ group.

Table S3. Associations between ISS and performances difference during N-back task in two independent cohorts.

Dataset	Task performance measures	MZ	Correlation with ISS	DZ	Correlation with ISS
HCP	2-back ACC(%)	6.96 ± 5.60	**$r_{MZ} = -0.27$, $p < 0.01$**	8.49 ± 8.13	$r_{DZ} = -0.22$, $p = 0.08$
	0-back ACC(%)	9.73 ± 10.23	$r_{MZ} = -0.01$, $p = 0.94$	8.87 ± 9.46	$r_{DZ} = 0.12$, $p = 0.33$
	2-back RT(ms)	105.48 ± 87.59	**$r_{MZ} = -0.23$, $p < 0.05$**	131.82 ± 94.85	$r_{DZ} = -0.06$, $p = 0.62$
	0-back RT(ms)	111.48 ± 105.46	$r_{MZ} = -0.17$, $p = 0.07$	116.34 ± 136.89	$r_{DZ} = -0.12$, $p = 0.35$
QTIM	2-back ACC	0.15 ± 0.14	$r_{MZ} = -0.01$, $p = 0.96$	0.19 ± 0.16	$r_{DZ} = -0.01$, $p = 0.93$
	0-back ACC	0.13 ± 0.19	$r_{MZ} = 0.03$, $p = 0.73$	0.16 ± 0.20	$r_{DZ} = -0.05$, $p = 0.50$
	2-back RT(s)	0.16 ± 0.14	$r_{MZ} = 0.16$, $p = 0.13$	0.14 ± 0.12	$r_{DZ} = 0.12$, $p = 0.14$
	0-back RT(s)	0.07 ± 0.09	$r_{MZ} = 0.08$, $p = 0.45$	0.08 ± 0.09	$r_{DZ} = -0.05$, $p = 0.48$

ISS, inter-subject synchronization; MZ, monozygotic twins; DZ, dizygotic twins.

Table S4. Prediction accuracy across twins from two independent cohorts.

Dataset	Contrasts	Twin A predict twin B (Average ± SEM)		Twin B predict twin A (Average ± SEM)	
		Average correlation	MSE	Average correlation	MSE
HCP	0-back vs. fixation	**0.63 ± 0.02**	0.53 ± 0.04	**0.61 ± 0.03**	0.81 ± 0.11
	2-back vs. fixation	0.61 ± 0.02	0.79 ± 0.07	0.60 ± 0.02	0.85 ± 0.12
	2-back vs. 0-back	0.37 ± 0.03	0.61 ± 0.14	0.32 ± 0.03	0.54 ± 0.08
	Faces vs. baseline	0.53 ± 0.03	0.32 ± 0.04	0.48 ± 0.03	0.33 ± 0.04
	Shapes vs. baseline	0.47 ± 0.03	0.38 ± 0.05	0.45 ± 0.03	0.31 ± 0.04
	Math vs. baseline	0.50 ± 0.02	0.29 ± 0.06	0.46 ± 0.02	0.36 ± 0.08
	Story vs. baseline	0.62 ± 0.02	0.28 ± 0.05	0.60 ± 0.02	0.44 ± 0.14
	Average motor	0.43 ± 0.02	0.49 ± 0.06	0.44 ± 0.03	0.47 ± 0.06
QTIM	0-back vs. fixation	**0.34 ± 0.03**	0.78 ± 0.10	**0.32 ± 0.04**	0.72 ± 0.13
	2-back vs. fixation	0.30 ± 0.02	1.24 ± 0.18	0.34 ± 0.02	1.27 ± 0.19
	2-back vs. 0-back	0.17 ± 0.03	1.65 ± 0.25	0.15 ± 0.03	1.63 ± 0.27

SEM, standard error of the mean; MSE, mean square error. Twin A and twin B are a pair of twin. Bold fonts represents the best result.

Supplementary Figures

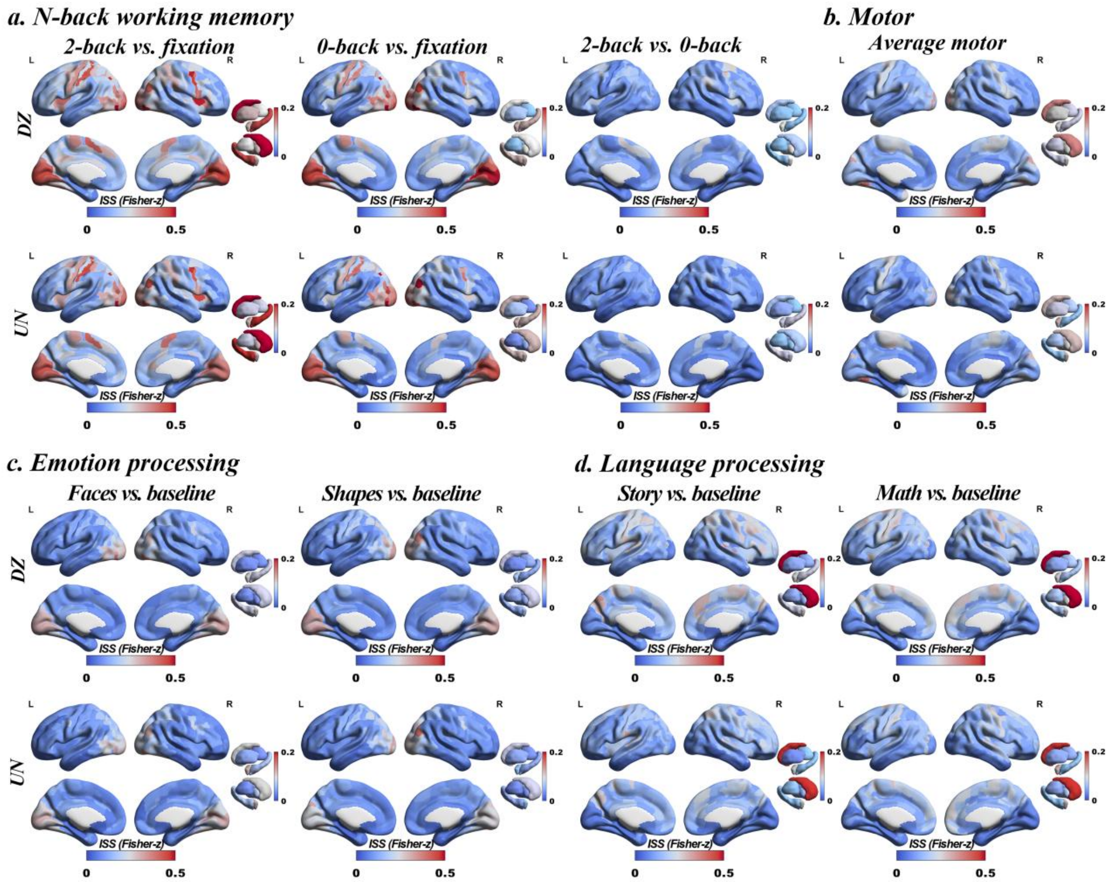

Figure S1. Task-evoked ISS in DZ twins and UN pairs. (a) N-back working memory task. (b) Motor task. (c) Emotion processing task. (d) Language processing task. DZ, dizygotic twins; UN, unrelated pairs.

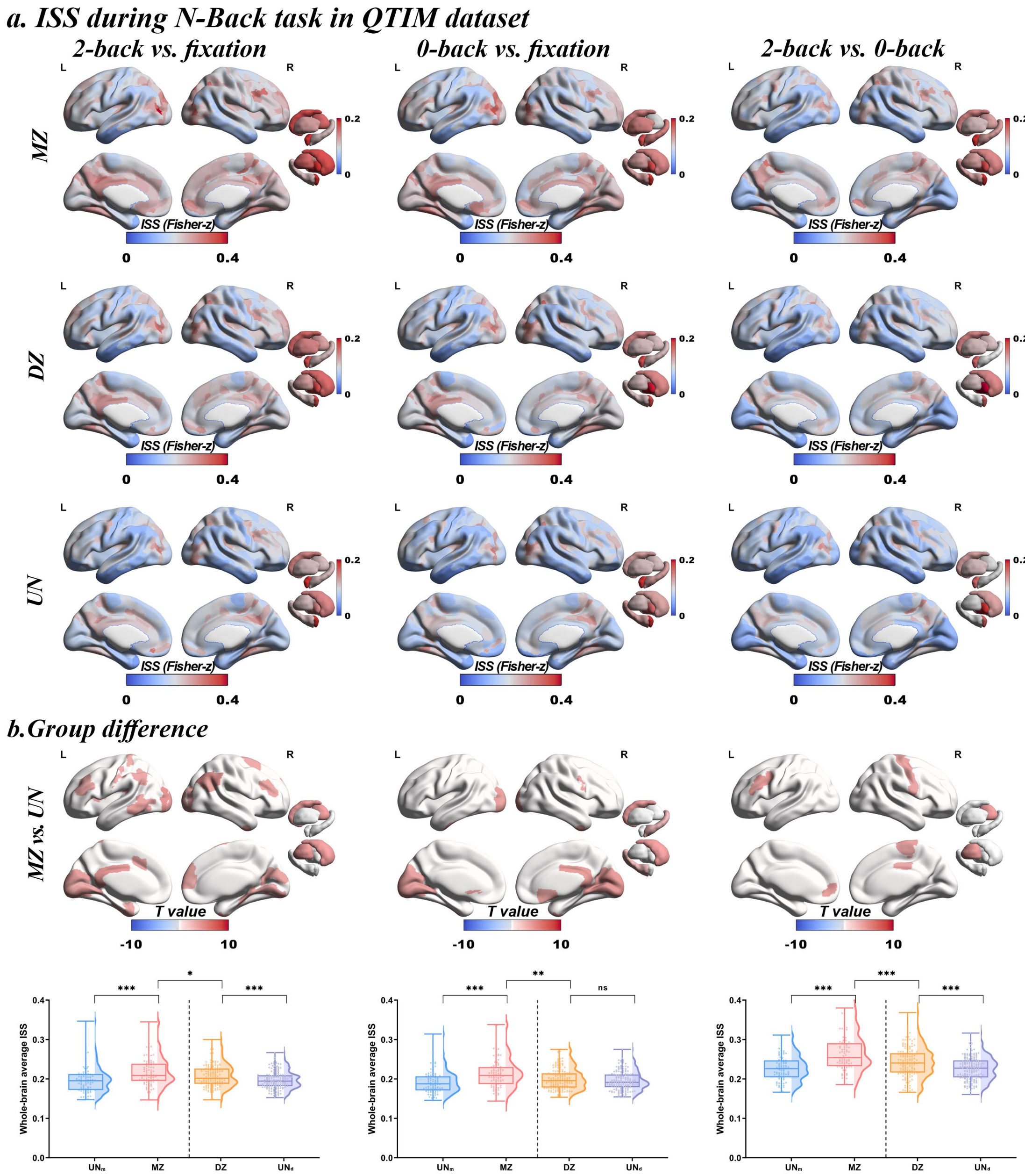

Figure S2. Task-evoked group difference in ISS between twins and UN pairs from QTIM dataset. (a) Regional ISS in MZ twins, DZ twins, and age- and sex-matched UN pairs. (b) Region-specific group comparisons were corrected for multiple comparisons using the FDR approach with a significance threshold of $q < 0.05$. Group differences in whole-brain ISS among MZ, DZ, and UN dyads. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ (two sample T-test). MZ, monozygotic twins; DZ, dizygotic twins; UN_m, unrelated pairs generated matched to the MZ group; UN_d, unrelated pairs matched to the DZ group.

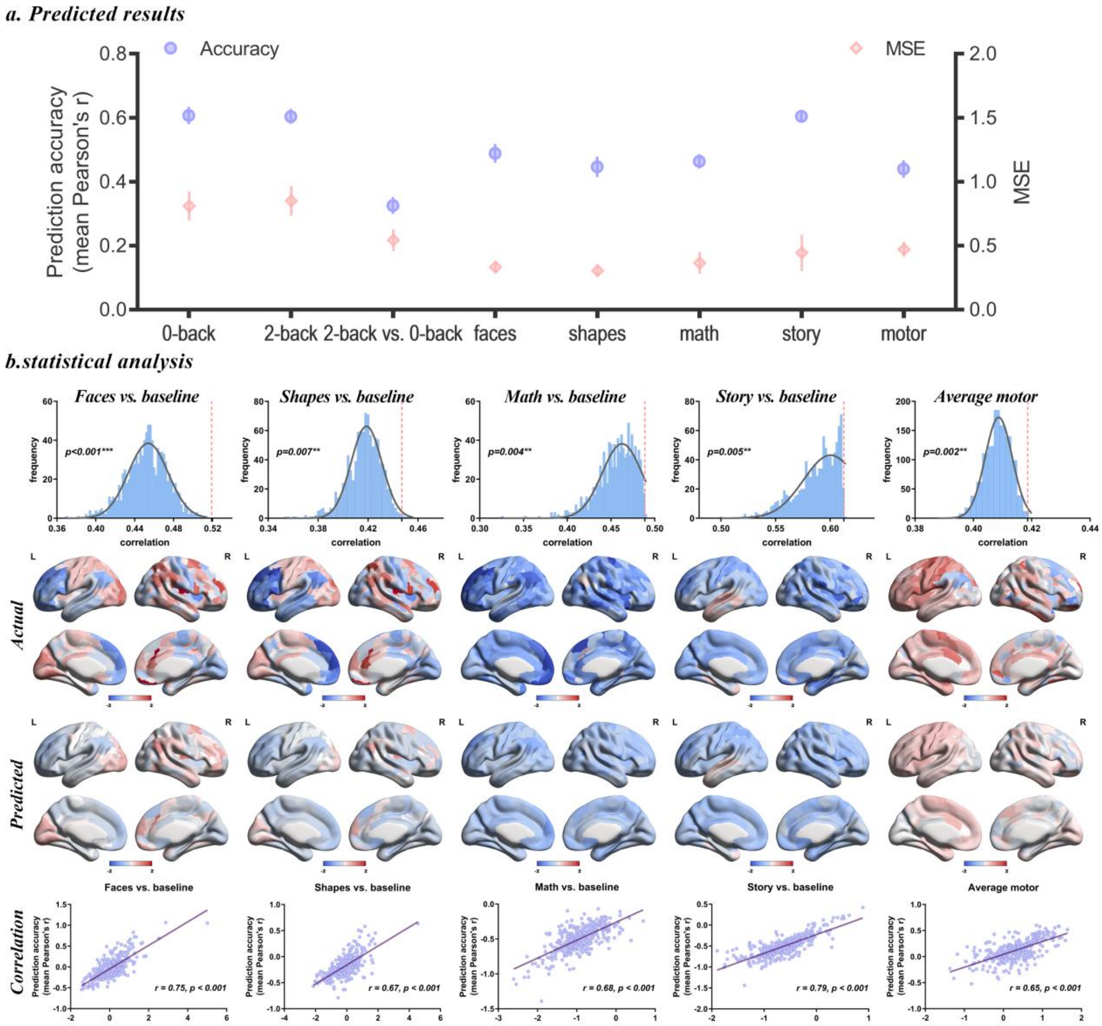

Figure S3. Prediction accuracy of task-evoked activation across twins in main dataset. (a) Average prediction accuracy and mean squared error (MSE) across all task contrasts, using one twin (Twin B) as the reference for model training. Twin A and Twin B are a pair of twins. (b) Top, Permutation test assessing statistical significance of prediction accuracy. Bottom, An example showing the observed and predicted task-evoked activation maps, along with the correlation analysis for the twin with the highest prediction accuracy in remaining five contrasts during adulthood.

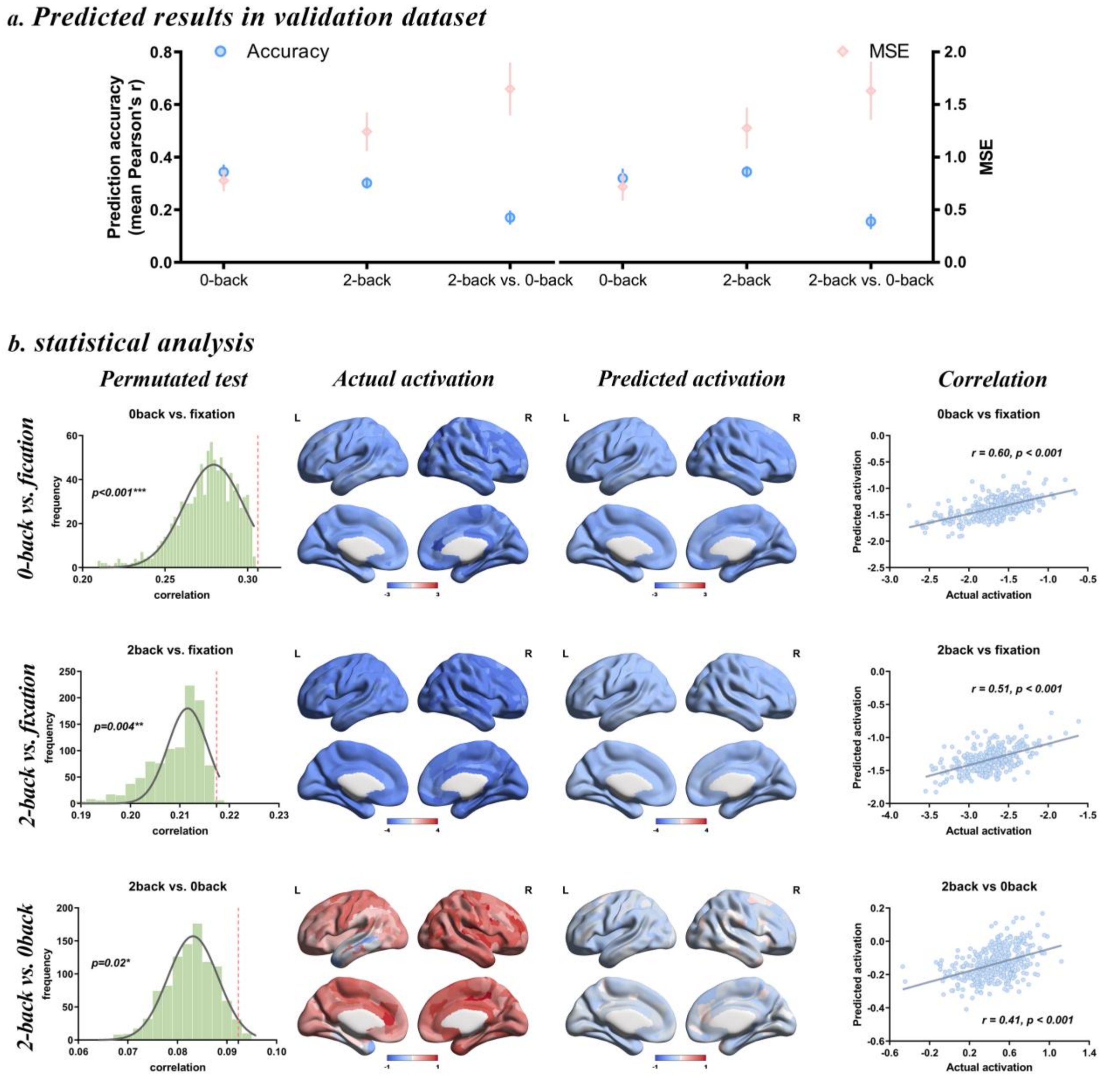

Figure S4. Prediction accuracy of N-back task-evoked activation across twins in validation dataset. (a) Average prediction accuracy and mean squared error (MSE) across all task contrasts. Left, using Twin A as the reference for model training. Right, reversing the source-target order (using twin B as the reference). Twin A and Twin B are a pair of twins. (b) Left, Permutation test assessing statistical significance of prediction accuracy. Right, An example showing the observed and predicted task-evoked activation maps, along with the correlation analysis for the twin with the highest prediction accuracy in the N-back task.

Supplementary methods

1. Behavioral performances assessment

Alertness: we incorporated the seven constituent components scores of the Pittsburgh Sleep Quality Index (PSQI_Comp1 to Comp7), deliberately omitting the PSQI total score and item-level metrics (e.g., specific bedtimes) to minimize collinearity and stochastic behavioral noise.

Cognition: rather than utilizing individual tasks prone to task-specific trade-offs, we selected the Age-Adjusted Scale Scores of the NIH Toolbox Fluid Cognition Composite and Crystallized Cognition Composite, supplemented by the List Sorting Working Memory Test. This triad captures core cognitive control, accumulated knowledge, and basic processing infrastructure while avoiding the mathematical redundancy introduced by the Global Total Composite.

Emotion: To eliminate subjective self-report biases and total-score mathematical dependency, the emotional profile was constrained to social-cognitive processing accuracy via the five specific emotion subscales (Anger, Fear, Happiness, Sadness, and Neutral) of the Penn Emotion Recognition Test (ER40).

Motor: Four age-adjusted or computed metrics indexing distinctive muscular and neuro-locomotive systems were consolidated: 2-minute Walk Endurance, 4-Meter Gait Speed, 9-hole Pegboard Dexterity, and Grip Strength.

Personality: Stable trait profiles were quantified using the five domains extracted from the NEO Five-Factor Inventory (NEO-FFI): Neuroticism, Extraversion, Openness to Experience, Agreeableness, and Conscientiousness.

Sensory: Five continuous, objective metrics representing disparate primary perceptual systems were selected, including Words-In-Noise (audition), Age-Adjusted Odor Identification (olfaction), Regional Taste Intensity (taste), Pain Interference T-score (pain), and the Mars Final Contrast Sensitivity Score (vision).

2. Cross-Twin Prediction

2.1 Model architecture

For each twin pair, rs-fMRI time series were extracted according to the HCP-MMP atlas. Two 1D convolutional layers with interleaved pooling and normalization were utilized to capture dynamical temporal features. The convolution operation is defined as:

$$F^l = \sigma(S \cdot W^l + b^l) \qquad S \in \mathbb{R}^{T \times R}$$

where F^l refers to the activation matrix at the l^{th} layer, and σ denotes the activation function. S is the input time series of length T across R brain regions, and W^l and b^l are the trainable weight and bias matrix, respectively. The BiGRU captures long-range temporal dependencies bidirectionally: forward GRU (GRU_f) processes the sequence in chronological order, whereas backward GRU (GRU_b) processes the reversed sequence.

$$\overrightarrow{h_t} = GRU_f(f_t, \overrightarrow{h}_{t-1})$$

$$\overleftarrow{h_t} = GRU_b(f_t, \overleftarrow{h}_{t-1})$$

The concatenated output is:

$$H^{(GRU)} = [\overrightarrow{H} \| \overleftarrow{H}]$$

where f_t denotes the input feature vector at time t, with $F = (f_1, f_2, \cdots f_{T'}) \in \mathbb{R}^{T' \times D}$, T' and D is the time length features dimensions after CNN pooling, respectively. GRU_f Global average pooling was conducted to integrate global feature into $z \in \mathbb{R}^{2D}$, followed by a linear output layer to predict activation may ŷ:

$$\hat{y} = z \cdot W^{(o)} + b^{(o)}$$

where $W^{(o)}$ and $b^{(o)}$ is the output layer's weight and bias parameters, respectively.

Prediction accuracy was evaluated by correlating the predicting activation maps $\hat{y}$ with the actual task-evoked activation maps y. The composite loss function combines mean squared error (MSE) and negative Pearson correlation between $\hat{y}$ and ground y:

$$\mathcal{L} = \lambda_1 \cdot MSE(\hat{y}, y) + \lambda_2 \cdot (1 - p(\hat{y}, y))$$

2.2 Experimental setup

The model integrates multi-scale temporal feature extraction with bidirectional contextual modeling. Prior to model training, all features were normalized across the whole-brain. For each task contrasts, predictions were generated using cosine-decay learning schedules with an initial learning rate of 2e-4, batch size of 24, and 150 training epochs. We used the AdamW optimizer with a weight decay of 1e-5 and a composite loss function (λ_1= 0.7 and λ_2= 0.3). The dataset was randomly divided into training and testing sets in the ratio of 8:2, with the random seed set to 42. The prediction procedures was repeated 10 times to ensure robustness, and the average predicted activation map for each contrast was computed.