frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 04 May 2023
DOI 10.3389/fnins.2023.1183145

OPEN ACCESS

EDITED BY
Xi Jiang,
University of Electronic Science and Technology of China, China

REVIEWED BY
Lei Xie,
Zhejiang University of Technology, China
Lu Zhang,
University of Texas at Arlington, United States

*CORRESPONDENCE
Bao Ge
✉ bob_ge@snnu.edu.cn

RECEIVED 09 March 2023
ACCEPTED 17 April 2023
PUBLISHED 04 May 2023

CITATION
He M, Hou X, Ge E, Wang Z, Kang Z, Qiang N, Zhang X and Ge B (2023) Multi-head attention-based masked sequence model for mapping functional brain networks. *Front. Neurosci.* 17:1183145. doi: 10.3389/fnins.2023.1183145

Multi-head attention-based masked sequence model for mapping functional brain networks

Mengshen He[1,2], Xiangyu Hou[2], Enjie Ge[2], Zhenwei Wang[2], Zili Kang[2], Ning Qiang[2], Xin Zhang[3] and Bao Ge[1,2]*

[1]Key Laboratory of Modern Teaching Technology, Ministry of Education, Shaanxi Normal University, Xi'an, China, [2]School of Physics and Information Technology, Shaanxi Normal University, Xi'an, China, [3]Institute of Medical Research, Northwestern Polytechnical University, Xi'an, Shaanxi, China

The investigation of functional brain networks (FBNs) using task-based functional magnetic resonance imaging (tfMRI) has gained significant attention in the field of neuroimaging. Despite the availability of several methods for constructing FBNs, including traditional methods like GLM and deep learning methods such as spatiotemporal self-attention mechanism (STAAE), these methods have design and training limitations. Specifically, they do not consider the intrinsic characteristics of fMRI data, such as the possibility that the same signal value at different time points could represent different brain states and meanings. Furthermore, they overlook prior knowledge, such as task designs, during training. This study aims to overcome these limitations and develop a more efficient model by drawing inspiration from techniques in the field of natural language processing (NLP). The proposed model, called the Multi-head Attention-based Masked Sequence Model (MAMSM), uses a multi-headed attention mechanism and mask training approach to learn different states corresponding to the same voxel values. Additionally, it combines cosine similarity and task design curves to construct a novel loss function. The MAMSM was applied to seven task state datasets from the Human Connectome Project (HCP) tfMRI dataset. Experimental results showed that the features acquired by the MAMSM model exhibit a Pearson correlation coefficient with the task design curves above 0.95 on average. Moreover, the model can extract more meaningful networks beyond the known task-related brain networks. The experimental results demonstrated that MAMSM has great potential in advancing the understanding of functional brain networks.

KEYWORDS

masked sequence modeling, multi-head attention, functional brain networks, feature selection, task fMRI

1. Introduction

Research into the function of the human brain has garnered significant attention and has been a popular field of study for several decades. One pivotal research direction in this field is the mapping of functional brain networks (FBNs), which has become a useful way to study the working mechanisms of the brain. By providing insight into the underlying neural mechanisms of such networks, FBNs hold the potential to unravel the working of the brain

(Power et al., 2010; Park and Friston, 2013; Sporns and Betzel, 2016; Jiang et al., 2021), as well as the pathogenesis of several diseases (Canario et al., 2021). Therefore, exploring FBNs is crucial for comprehending the complex dynamics of the brain and can offer an avenue for further understanding the neural processes underlying different functions.

In traditional methods, generalized linear models (GLM) (Beckmann et al., 2003; Barch et al., 2013), independent component analysis (ICA) (McKeown, 2000; Beckmann et al., 2005; Calhoun and Adali, 2012), and sparse dictionary learning (SDL) (Lv et al., 2014; Ge et al., 2016; Lee et al., 2016; Zhang et al., 2016; Shen et al., 2017; Zhang et al., 2018) have been utilized to construct functional brain networks. Moreover, other machine learning techniques have been effectively applied to fMRI data analysis, such as support vector machines (SVM) (LaConte et al., 2005; Mourao-Miranda et al., 2006) for fMRI analysis and classification, and principal component analysis (PCA) (Thirion and Faugeras, 2003; Smith et al., 2014) for fMRI data dimensionality reduction. With the advancement of deep learning technology, numerous deep learning models have been applied to fMRI data analysis and functional brain network construction. For instance, Huang et al. (2017) proposed a deep convolutional autoencoder (DCAE) to extract hierarchical features from fMRI data; Zhao et al. (2018) proposed a spatiotemporal convolutional neural network (ST-CNN) to learn temporal and spatial information from fMRI data simultaneously; Qiang et al. (2020) proposed a spatiotemporal self-attention mechanism (STAAE) (Dong et al., 2020b) for brain functional network modeling and ADHD disease classification. Additionally, Qiang et al. (2020) proposed a residual autoencoder (RESAE) (Dong et al., 2020a) for constructing task related functional brain networks. Jiang et al. (2023) introduce a Spatio-Temporal Attention 4D Convolutional Neural Network (STA-4DCNN) model to characterize individualized spatio-temporal patterns of FBNs. Yan et al. (2022) proposed a Multi-Head Guided Attention Graph Neural Network (Multi-Head GAGNN) to simultaneously model both spatial and temporal patterns of holistic functional brain networks. Experimental results have indicated that deep learning methods are effective in fMRI data modeling and brain network construction tasks, which demonstrate the significant advantages of deep learning models.

Although the methods mentioned above have shown promising results, there are still certain limitations that need to be addressed. Firstly, the current design and parameterization of models do not fully account for the characteristics of fMRI data. For instance, the same signal value at different time points may have different meanings depending on the task or state, and thus, it is crucial to exploit this information for improving model performance. Secondly, the model training process disregards some prior knowledge, such as task design curves, which could potentially enhance the efficacy and efficiency of the model. These limitations underscore the need for more advanced techniques that can tackle these challenges and improve the accuracy and applicability of fMRI analysis.

Recent research has revealed the exceptional capabilities of Transformer models (Vaswani et al., 2017) in tasks such as text analysis and prediction. One of key mechanisms of transformer is to use multi-head attention to do the processing of sequence data. By leveraging multi-head attention mechanisms, the distinctive semantics of a single word in different language contexts can be analyzed. For instance, the term "apple" could signify either a fruit or a mobile phone brand in various language contexts. Given the similarity between fMRI time series and text sequences, multi-head attention mechanisms can be employed to extract features from fMRI data. Furthermore, the growing popularity of the masked language modeling (MLM) training method in the Bert model (Devlin et al., 2018) suggests that masking-based training techniques are remarkably effective at capturing contextual information. Since there are similarities between fMRI time series and sentences, the multi-head attention mechanism and mask training method can be extended to fMRI feature extraction.

So, this manuscript proposed a novel model called the Multi-head Attention-based Masked Sequence Model (MAMSM) which utilizes a multi-head attention mechanism to scrutinize different states of voxel signals at various locations while also implementing the Masked Sequence Model (MSM) method to analyze and process the fMRI time series. Furthermore, MAMSM employs both randomly discrete and continuous masks in the masking operation to enhance the model's learning capacity and training effectiveness. In addition to that, this study leverages prior knowledge of the task design curves and cosine similarity to construct a new loss function, resulting in improved outcomes in model training.

In order to demonstrate the effectiveness of our proposed model, we utilized data from the Human Connectome Project (HCP) (Van Essen et al., 2013) and analyzed the seven task-state datasets of 10 individuals using both individual and group average approaches. To evaluate the performance of our model, we compared it with the SDL and STAAE methods. The experimental results indicate that the FBNs extracted by our proposed model outperformed those extracted by the other methods across various task datasets. Notably, our model also detected several brain networks that were distinct from the task-state-corresponding FBNs, and we subsequently identified some networks as similar to the known resting-state brain networks. Specifically, our experimental results demonstrate that our model is highly effective in extracting features from a small amount of data, which is particularly important in the context of brain imaging research where data acquisition is often difficult, costly, and resource-intensive. A brief version of the study has been published as a conference paper in the MICCAI 2022 (He M. et al., 2022).

2. Materials and methods

2.1. Overview

As shown in **Figure 1**, the proposed method consists of three main steps: (1) four-dimensional fMRI data is pre-processed and mapped to two-dimensional space; (2) the pre-processed two-dimensional fMRI time series is input into the MAMSM, composed of multiple headed attention mechanisms, and trained with a mask-based approach; (3) all the latent features extracted from the pre-training are input into the feature selection layer, which are trained with a loss function by leveraging the prior task designs. Finally, the features output by the encoder of the feature selection layer are regressed by lasso and mapped back to the original brain space, resulting in the visualization of FBNs.

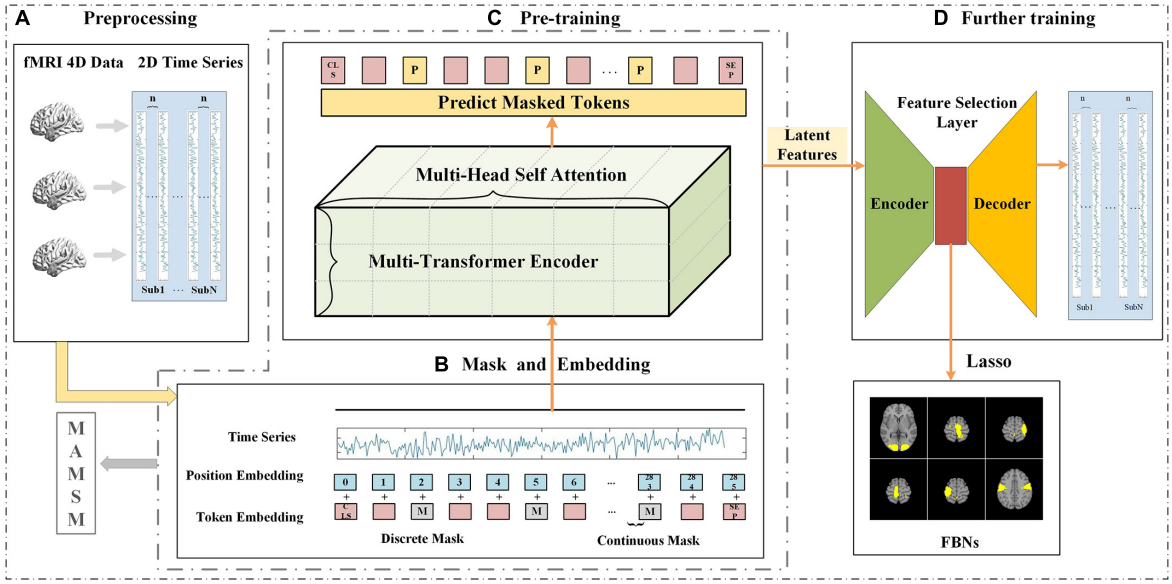

FIGURE 1
The framework of MAMSM **(A)** preprocessing, which involves mapping all subjects' tfMRI data to 2D space; **(B)** masking and embedding operations of time-series data; **(C)** pre-training of the model, consisting of three layers of transformer encoders; **(D)** further training is performed using the feature selection layer, and the obtained features are used to map FBNs.

TABLE 1 Summary of used datasets.

	H	W	D	Time points	Voxels	Training subjects
Motor	46	55	46	284	28,546	10
Emotion	46	55	46	176	28,546	10
Gambling	46	55	46	253	28,546	10
Language	46	55	46	316	28,546	10
Relational	46	55	46	232	28,546	10
Social	46	55	46	274	28,546	10
WM	46	55	46	405	28,546	10

2.2. Materials and pre-processing

The dataset from the Human Connectome Project Q3 was used in this work, which is publicly available on the website.[1] We selected randomly the 10 subjects from HCP dataset. To evaluate the temporal features and spatial features obtained by the MAMSM, we chose 24 task designs from seven tasks. The corresponding hemodynamic response function (HRF) responses, which are the convolution of the task paradigm and HRF function, are utilized as temporal templates and the group-wise functional brain networks (FBNs) derived from the GLM are utilized as spatial templates (Güçlü and Van Gerven, 2017). For the sake of description, 24 distinct symbols were used to represent each of the selected task designs. For emotion task, E1 is for emotional faces, and E2 is for simple shapes. For gambling task, G1 is for punishment over baseline, and G2 is for reward over baseline. For language task, L1 is for math over story, and L2 is for story over math. For social task, S1 is for social over baseline, and S2 is for random over baseline. For relational task, R1 is for match over baseline, and R2 is for relational over baseline. For motor task, M1-M6 are for cue, left foot movement, left hand movement, right foot movement, right hand movement, and tongue movement, respectively. For working memory task, W1-W8 are for the 2-back and 0-back task events of body parts, places, faces, and tools, respectively.

The parameters of data collection used in this text is as follows: a 90 × 104 matrix, 220 mm FOV, 72 slices, TR = 0.72 s, TE = 33.1 ms, Flip angle = 52°, BW = 2,290 Hz/Px, in-plane FOV = 208 mm × 180 mm. For the tfMRI data, the pre-processing operations included skull stripping, motion correction, slice timing correction, spatial smoothing, global drift removal (high pass filtering) and registration to MNI space. Table 1 provides an overview of the pre-processed task functional magnetic resonance imaging (tfMRI) datasets used in this study. After pre-processing of the tfMRI data, the four-dimensional tfMRI data was transformed into a two-dimensional matrix by using Nilearn tools (available at https://nilearn.github.io/) and the MNI-152 mask. Data for each time point comprised 28,546 voxels.

1 https://db.humanconnectome.org

TABLE 2 The results of training with different mask operations.

Mask strategies	Training loss	Predict loss
Discrete	0.043	4.73
Continuous	0.047	4.739
Discrete and continuous	0.04	4.719

2.3. MAMSM

2.3.1. MSM

In recent years, Masked Language Modeling (MLM) and Masked Image Modeling (MIM) approaches have been widely employed in Natural Language Processing (NLP) (Devlin et al., 2018; Chung et al., 2021; Sinha et al., 2021) and Computer Vision (CV) (Zhou et al., 2021; He K. et al., 2022; Tong et al., 2022; Xie et al., 2022) due to their demonstrated efficacy in extracting contextual information through mask training. This work utilized Masked Sequence Modeling (MSM) to process fMRI sequence data. MSM is a self-supervised training method in which a portion of the tokens in the sequence are replaced with [mask] symbols and the remaining tokens and location information are used to predict the tokens replaced with [mask]. This training method allows the model to learn more about the relationships between contexts.

In the BERT model proposed by Devlin et al. (2018), the [CLS] (Classification Token) serves to create a compact representation of the entire input sequence. This condensed representation can be used for tasks such as text classification and similarity computation. Specifically, for each input fMRI time series, the proposed model is designed to generate a vector representation for each input. By adding the special [CLS] tag at the beginning of the sequence, this vector representation of the tag serves as a summary of the entire sequence, compressing and integrating the information from the entire input. As a result, the [CLS] tag provides a comprehensive representation for subsequent feature extraction and similarity calculation processes.

Before the mask processing process, the fMRI data was normalized to a range of (0, 1). After normalization, we retained three decimal places for the values, resulting in a maximum of 1,001 distinct values (from 0, 0.001, 0.002, . . . to 1) for the whole-brain signals. In the subsequent model training process, we treat these 1,001 different values as 1,001 classes, simplifying the model training process into a multi-classification problem. That is, if we want to predict the value of fMRI signals at a certain time point, we converted it into categories with a total of 1,001 values for classification. The prediction range of the model is also within these 1,001 classes of values. When predicting the value of a masked position, the model only needs to determine the class to which it belongs. To facilitate the prediction of token values, a multi-classification task was employed, where in a cross-entropy loss function was utilized to compute the error between the model's predicted value and the actual value. As shown below, where $\mathbf{y_i}$ is the true probability distribution, $\widehat{\mathbf{y}}_\mathbf{i}$ is the predicted probability distribution, and n is the number of categories:

$$\mathbf{C}E\left(\mathbf{y_i}, \widehat{\mathbf{y}}_\mathbf{i}\right) = -\sum_{\mathbf{i}=\mathbf{1}}^{\mathbf{n}} \mathbf{y_i} log\left(\widehat{\mathbf{y}}_\mathbf{i}\right)$$

In the mask processing process, for each fMRI sequence input, a certain proportion of positions on the fMRI time series will be randomly covered, with the original signal values replaced by [mask]. Here, taking the tfMRI sequence of Motor task as an example, we selected roughly 10% time points as mask locations for each input with the length of 284 time steps, as illustrated in Figure 1B. After the Mask operation was performed, the pre-training stage in the proposed model employed an unsupervised training process to predict the token values of the masked locations, as shown in Figure 1C.

In order to enhance the learning capability of the model and achieve optimal training outcomes, this study employs a combination of continuous and discrete masking techniques. When using only discrete masking, the model may be able to predict the values of the masked regions through simple methods such as averaging the values of its previous and subsequent time steps. This may lead to the model failing to learn deeper features. To avoid this issue, we designed more sophisticated methods of masking, such as continuous mask, etc., Table 2 presents the outcomes of the training with different masking modes, where 90% of the voxels in the same subject are allocated for the training set, 10% for the test set, and the same training parameters are utilized. We adopt a uniform sampling strategy for voxel selection, wherein every ten voxels, the first nine are assigned to the training set, and the last one is designated for the testing set. By comparing the

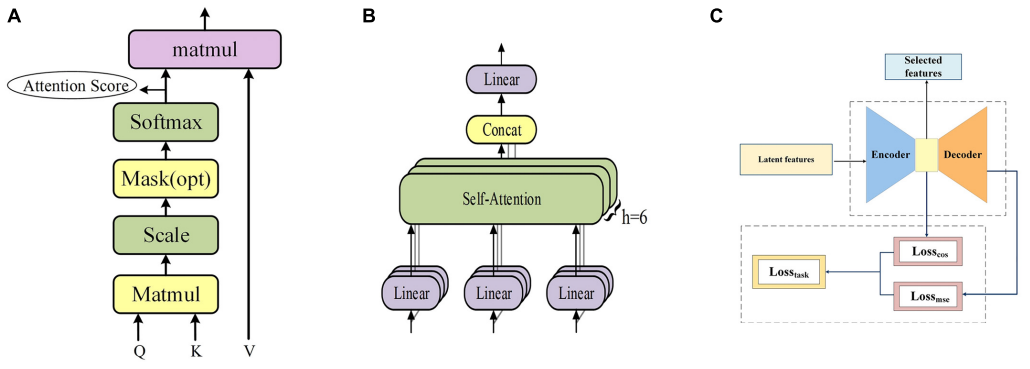

FIGURE 2
(A) The frame of self-attention. **(B)** The frame of multi-head attention. **(C)** The frame of feature selection layer.

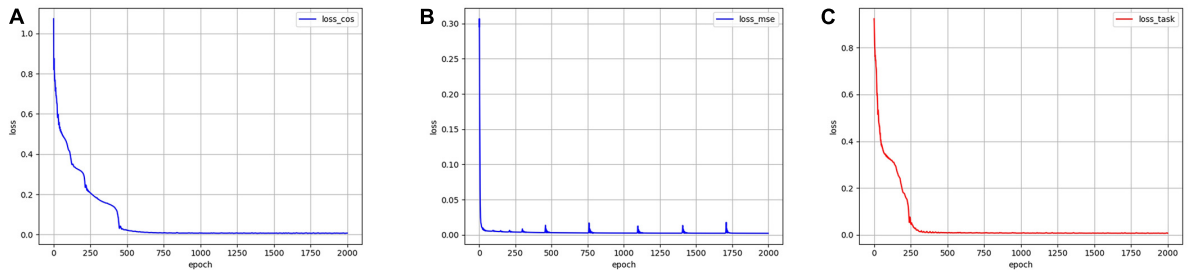

FIGURE 3
The training errors by using three different loss functions. **(A)** Error variation by using $\boldsymbol{Loss_{cos}}$. **(B)** Error variation by using $\boldsymbol{Loss_{mse}}$. **(C)** Error variation by using $\boldsymbol{Loss_{task}}$.

minimum loss on the training set and the test set, it can be seen that the combination of the two mask operations can achieve better results.

2.3.2. Multi transformer encoder layers

The Transformer model is a sophisticated deep neural network that is based on an attention mechanism, originally introduced by Vaswani et al. (2017) for machine translation. The model is structured according to the seq2seq paradigm and comprises two primary components: an encoder that encodes the input sequence and a decoder that generates the output sequence. Unlike traditional Recurrent Neural Network (RNN) models (Hochreiter and Schmidhuber, 1997; Schuster and Paliwal, 1997; Graves et al., 2005; Cho et al., 2014), the transformer model utilizes multi-head attention mechanism for computation. This mechanism can represent information from multiple semantic spaces, capturing different meanings of the same words in different contexts, similar to the same signal values in fMRI data may represent different states and meanings.

Therefore, in this manuscript, each fMRI sequence is embedded and masked as the input of the transformer encoder, and then the input is linearly transformed to obtain three matrices, namely **Q** (Query), **K** (Key) and **V** (Value). Subsequently, **Q** and **K** are dot-multiplied and then normalized by dividing by $\sqrt{\mathbf{d_k}}$ to stabilize the gradient. Subsequently, a softmax operation was used to obtain the attention score, which represents the importance of each position of the fMRI sequence, and then multiplied by V to obtain the output of self-attention, as shown in Figure 2A. Eventually, the output of multiple self-attentions is superimposed as the output of multi-headed attention, as shown in Figure 2B. The formulae of self-attention and multi-head attention can be expressed as follows, where $\mathbf{h}ead_\mathbf{i}$ denotes the $\boldsymbol{i}$-th self-attention mechanism.

$$\mathbf{M}ultiHead\,(\mathbf{Q},\mathbf{K},\mathbf{V}) \;=\; \mathbf{C}oncat\left(\mathbf{h}ead_1,,\mathbf{h}ead_\mathbf{n}\right)\mathbf{W^O}$$

$$\mathbf{h}ead_\mathbf{i} \;=\; \mathbf{A}ttention(\mathbf{Q},\mathbf{K},\mathbf{V})$$

$$\mathbf{A}ttention\,(\mathbf{Q},\mathbf{K},\mathbf{V}) \;=\; \mathbf{s}oftmax(\frac{\mathbf{QK^T}}{\sqrt{\mathbf{d_k}}})\mathbf{V}$$

Upon completion of the pre-training of the model, the attention score was extracted as a feature matrix, which represents the weights at various time points within an fMRI time series. After the model pre-training was completed, the attention scores were extracted as the features representing the weights of each time point in the fMRI time series. We use the sliding average operation to smooth the attention scores, and then use the average results as latent features of the pre-trained model. We set the size of the sliding average window to 10 and the step size of the sliding window is 1.

2.3.3. Feature selection layer

Here we propose a novel loss function, $\boldsymbol{Loss_{task}}$, for the training of a feature selection layer in autoencoders, as illustrated in Figure 2C. By combining mean squared loss function ($\boldsymbol{Loss_{mse}}$) and cosine similarity loss function ($\boldsymbol{Loss_{cos}}$), this loss function is more conducive to the task of tfMRI data compared to the other methods (Dong et al., 2020b; Qiang et al., 2020), which often focus solely on reconstruction error such as MSE, disregarding the latent feature distribution and the relationship with the task curves, both of which are indispensable to characterize fMRI time series. The latent feature matrix obtained from pre-training serves as the input for the encoder, which, after training, produces the final feature matrix as its output. Through this process, the feature selection layer also facilitates the reduction of dimensionality of the latent feature matrix, thus contributing to more efficient and effective features. The $\boldsymbol{Loss_{task}}$ function is formulated as the combination of $\boldsymbol{Loss_{cos}}$ and $\boldsymbol{Loss_{mse}}$ and we experimentally chose the value of k to 1 in this work, as follows:

$$\boldsymbol{Loss}_{\mathbf{mse}} \;=\; \mathbf{M}SE \;=\; \frac{\mathbf{1}}{\mathbf{n}}\sum_{\mathbf{i=1}}^{\mathbf{m}}\mathbf{w_i}(\mathbf{y_i}-\widehat{\mathbf{y}}_\mathbf{i})^\mathbf{2}$$

$$\boldsymbol{Loss}_{\mathbf{cos}} \;=\; \mathbf{l_i} \;=\; \mathbf{1}-cos\left(\mathbf{x_i},\mathbf{y_i}\right)$$

$$\boldsymbol{Loss}_{\mathbf{task}} \;=\; \mathbf{loss}\,(\mathbf{mse})+\mathbf{k}*\mathbf{l}oss(\mathbf{c}os)$$

The actual value $\mathbf{y_i}$ and the predicted value $\widehat{\mathbf{y}}_\mathbf{i}$ are compared by calculating the cosine similarity between the n sequences of the

TABLE 3 The final training errors of three different loss functions.

	$Loss_{cos}$	$Loss_{mse}$	$Loss_{task}$
Cos-error	0.0074	1.0632	**0.0067**
Mse-error	0.2955	0.0022	**0.0022**

The bold values represent the minimum values of each row.

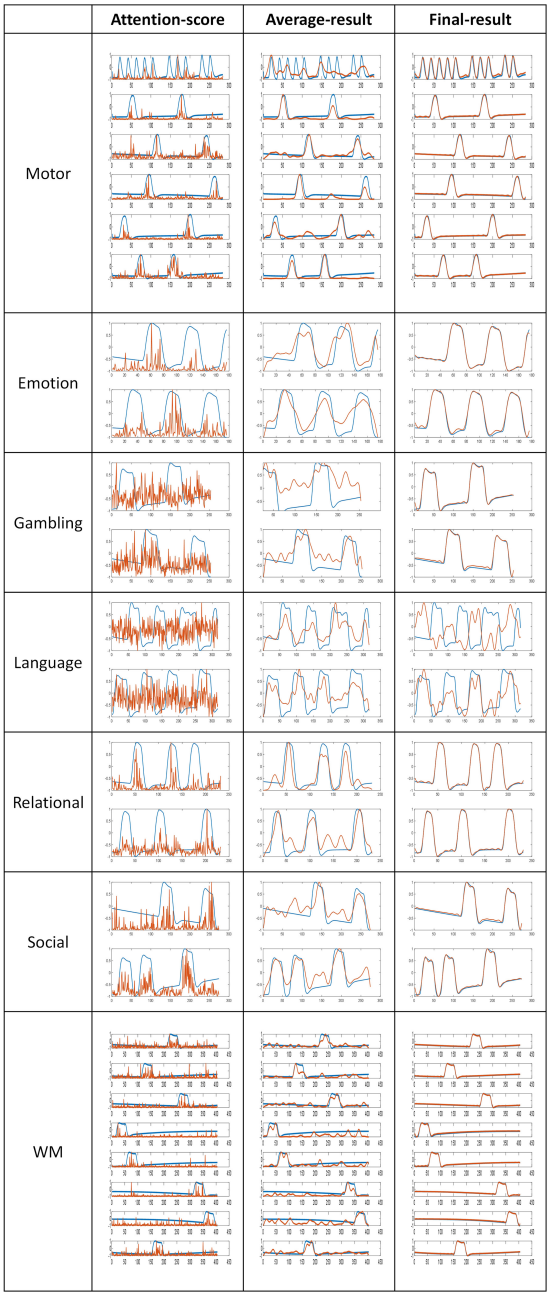

FIGURE 4
Comparison of features and task designs. The blue curves represent the task designs, the red curves depict the features.

feature$\mathbf{x_i}$ and the n task design curves $\mathbf{y_i}$, and the cosine similarity calculation formula is:

$$cos\left(\mathbf{x_i}, \mathbf{y_i}\right) = \frac{\sum_{i=1}^{n}\left(\mathbf{x_i} \times \mathbf{y_i}\right)}{\sqrt{\sum_{i=1}^{n}\left(\mathbf{x_i}\right)^2} \times \sqrt{\sum_{i=1}^{n}\left(\mathbf{y_i}\right)^2}}$$

Mse-error is the MSE reconstruction error of the decoder output and the original data; Cos-error is the cosine similarity error between the n sequences of the encoder output feature and n task design curves. To demonstrate the effectiveness of the proposed new loss function, an ablation experiment was conducted using the same data and parameters. The model was trained using $\boldsymbol{Loss_{cos}}$, $\boldsymbol{Loss_{task}}$, and $\boldsymbol{Loss_{mse}}$, respectively. As illustrated in Figure 3, when $\boldsymbol{Loss_{task}}$ was used, the convergence rate was faster and more stable than when only $\boldsymbol{Loss_{cos}}$ or $\boldsymbol{Loss_{mse}}$ was used. Quantitatively, Table 3 shows that when $\boldsymbol{Loss_{task}}$ was employed, the final Cos-error and Mse-error were lower.

2.3.4. Mapping FBNs

To obtain the spatial distribution of the functional network, lasso regression is applied to the feature matrix and the original two-dimensional input data to get the sparse coefficient matrix, which represents the spatial distribution of the functional network The calculation formula of LASSO regression (Pedregosa et al., 2011) is as follows:

$$\min_{\mathbf{w}} \frac{1}{2\mathbf{T}} \| \mathbf{Z} - \mathbf{X}W \|_2^2 + \lambda \| \mathbf{W} \|_1$$

$\mathbf{Z}$ is the original 2D input data, $\boldsymbol{T}$ represents the total number of time points, $\boldsymbol{X}$ is the feature matrix, and $\boldsymbol{W}$ is the regressed sparse coefficient matrix. The coefficient matrix $\boldsymbol{W}$, which captures the spatial distribution information of the underlying functional network, was then mapped back to the original 3D brain image space, the result was finally visualized as FBNs.

3. Results

The work reports its findings in terms of two primary dimensions: temporal and spatial features. To evaluate temporal features, the final feature matrix was utilized to obtain partial task-related features, which were subsequently evaluated for similarity with the task design curves. Spatial features were assessed by computing the similarity between the derived FBNs and the templates derived from the GLM. Besides task-related FBNs, we also identified additional FBNs, including those resting-state FBNs.

TABLE 4 Pearson correlation coefficient between the features and the task designs.

	E1	E2	G1	G2	W1	W2	W3	W4	W5	W6	W7	W8	/
Attention-score	0.331	0.343	0.321	0.333	0.287	0.268	0.270	0.267	0.276	0.262	0.275	0.276	/
Average-result	0.851	0.894	0.792	0.792	0.813	0.799	0.870	0.804	0.845	0.845	0.789	0.856	/
Final-result	0.999	0.998	0.998	0.998	0.999	0.999	0.999	0.994	0.999	0.999	0.998	0.999	/
	L1	L2	S1	S2	R1	R2	M1	M2	M3	M4	M5	M6	Ave
Attention-score	0.262	0.250	0.319	0.583	0.419	0.337	0.336	0.423	0.430	0.408	0.427	0.567	0.345
Average-result	0.737	0.785	0.796	0.916	0.880	0.889	0.451	0.903	0.913	0.880	0.873	0.961	0.831
Final-result	0.727	0.737	0.998	0.997	0.999	0.999	0.973	0.997	0.998	0.997	0.997	0.997	0.975

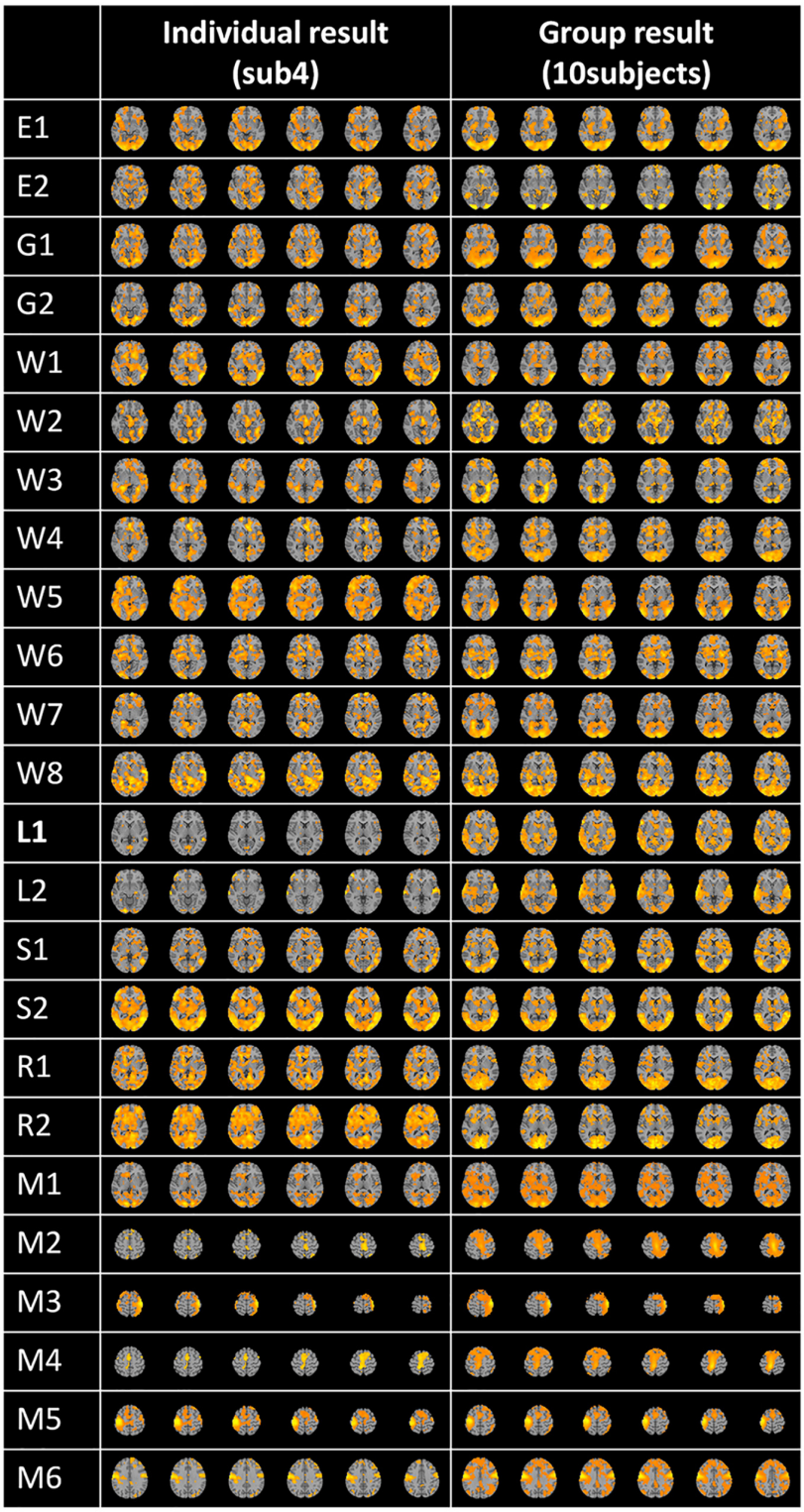

FIGURE 5
Individual and group averaged FBNs.

3.1. Temporal features

The proposed model generated three different temporal feature matrices, namely the intermediate "attention-score" feature, which is obtained immediately after model pre-training; the "average-result" feature, calculated by computing a sliding average of the attention-score feature; and the "Final-result" feature, obtained after training the feature selection layer. The dimension of attention-score, average-result, and final-result are [6*28,546,t], [6*28,546,t], and [256,t]. In this work, "t"

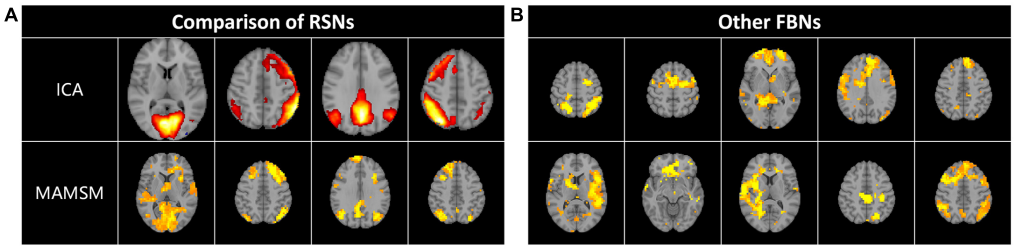

FIGURE 6
(A) Some resting-state FBNs. **(B)** Other FBNs.

represents the length of the fMRI sequence's time dimension corresponding to different tasks, while "6" denotes the number of attention heads we have set for the multi-head attention mechanism. To evaluate the significance of the three kinds of features selected in this study, a comparative analysis is conducted between these features and the task design curves. As illustrated in Figure 4, a graphical representation of the three kinds of features and the correspondingly relevant task design curves are presented. The blue curves represent the task design curves and serve as the baseline, the red curves depict the features.

Based on the results of the comparison, it is evident that the attention-score and task curves display an obvious fitting trend, with their highest peak approximately coinciding with the peak of the task design curves. Furthermore, the application of a sliding average filter results in an even higher similarity between the average-result and task design curves. These outcomes provide evidence that the latent features derived from the pre-training module are both meaningful and interpretable.

In order to quantitatively compare the similarity between the feature matrices and the task design curves, the Pearson correlation coefficient was calculated in this work, the formula for the Pearson correlation coefficient is presented below:

$$\rho(\mathbf{X}, \mathbf{Y}) = \frac{\sum_{i=1}^{n}(\mathbf{X_i}-\mu_X)(\mathbf{Y_i}-\mu_Y)}{\sqrt{\sum_{i=1}^{n}(\mathbf{X_i}-\mu_X)^2}\sqrt{\sum_{i=1}^{n}(\mathbf{Y_i}-\mu_Y)^2}}$$

where $\mathbf{X}$, $\mathbf{Y}$ are the features and task design curves, μ_X, μ_Y are the mean of the them and$\mathbf{X_i}$, $\mathbf{Y_i}$ are the samples of them.

The Pearson correlation coefficient values serve as an indicator of the strength of the correlation, with higher values indicating stronger correlations. As shown in Table 4, all Pearson's correlation coefficients achieved statistical significance at the level of $P < 0.05$. These results demonstrate that the features extracted by the proposed pre-training model were significantly correlated with the design curves. Specifically, the initially extracted attention-score feature exhibited a certain degree of similarity with the task design curves. With the application of the sliding average technique, the Average-result feature approached the task design curves more. Finally, the incorporation of a feature selection layer and a new loss function as a guide led to the generation of the Final-result feature. The Pearson correlation coefficient for the task design curves was significantly improved from 0.831 to 0.975 as a result. These findings underscore the importance of the pre-training model and feature selection layer, and provide further support for the efficacy and interpretability of the proposed model in this study.

3.2. Spatial features

3.2.1. Task FBNs

Following the feature selection process, the feature matrix was remapped to the original 3D brain space for the visualization of FBNs using lasso regression, as shown in Figure 5. This figure displays a randomly selected individual FBN for 24 tasks and group-averaged FBNs from 10 subjects. As demonstrated in Figure 5, each task-related FBN can be accurately identified, and the FBNs becomes even more pronounced after group averaging.

3.2.2. Other FBNs

Multi-head Attention-based Masked Sequence Model can not only acquire the known activated networks, but also enable the identification of other brain networks with specific patterns. In this work, we also selected and displayed a part of them. After comparison and analysis, we found some resting-state networks, which were compared and displayed with the corresponding resting-state brain network templates obtained by the ICA method, as shown in Figure 6A. In addition, this manuscript also displays other brain networks with certain patterns, as shown in Figure 6B.

3.3. Comparative experiments

To further evaluate the effectiveness of the proposed MAMSM, it is compared with SDL (Lv et al., 2015)and STAAE (Dong et al., 2020b). SDL is the traditional way to build FBNs. STAAE has been proposed as a deep learning method recently. All three methods are applied to the same dataset and their temporal and spatial characteristics are compared in this section.

3.3.1. Comparison of temporal features

In this study, three different methods were employed for comparison purposes. In order to ensure fairness in our comparison analysis, we adopted the "average-result" features instead of the "final result" features for comparison with the features obtained from SDL and STAAE, as our proposed model leveraged prior knowledge (task designs) to train the model in the feature selection layer. Figure 7 displays the task design curves, with the blue curves representing specific task design curves used as comparison benchmarks and the red curves representing the task-related features. Our qualitative and quantitative comparison analysis aimed to assess the degree of correlation between these two curves. For quantitative comparison, the Pearson correlation

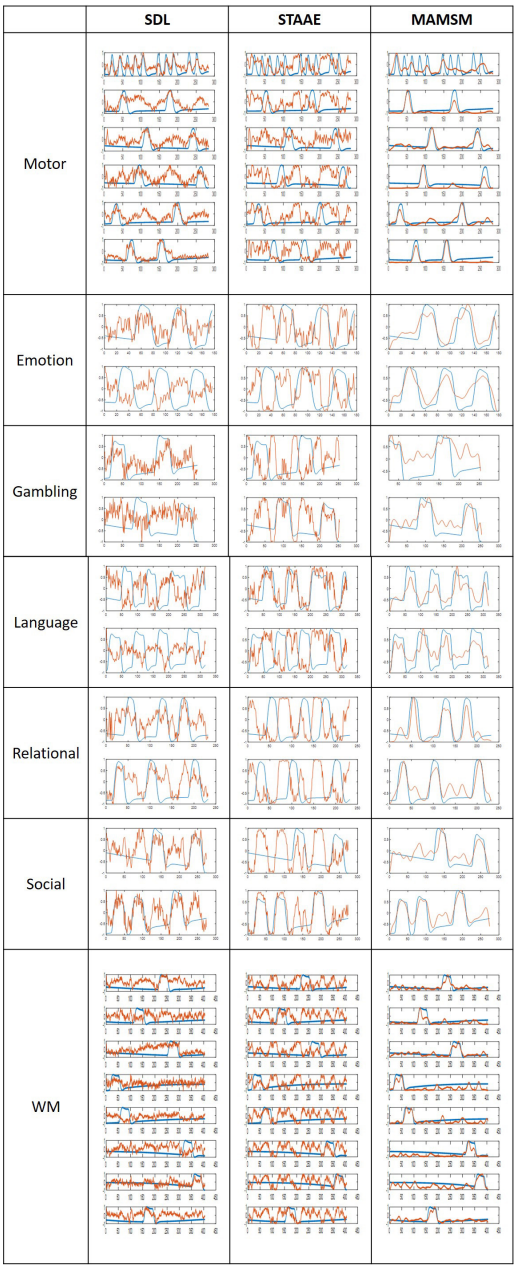

FIGURE 7
Comparison of features and task designs obtained by different methods. The blue curves represent the task designs, the red curves depict the features.

coefficient was employed to assess the similarity between the extracted features and the task curves, as presented in Table 5. It should be noted that Figure 7 shows the results of an individual. Table 5 is the average result of ten individuals.

As shown in Figure 7, the correlation between the features generated by MAMSM and the task design curves was found to be significantly higher compared to that between the features generated by SDL/STAAE and the task design curves. Quantitatively, the results presented in Table 5 demonstrate that the proposed MAMSM achieved a higher averaged Pearson correlation coefficient (0.824) compared to that from SDL (0.527) or STAAE (0.306). Overall, the results of our experiment demonstrate the effectiveness of MAMSM for constructing FBNs based on tfMRI.

In terms of individual-level performance, our results indicate that the deep learning method STAAE performed slightly worse than SDL and MAMSM. It is worth noting that according to the description of the STAAE (Dong et al., 2020b), the method can achieve better results when applied to larger datasets. However, the inherent requirement of deep learning methods for large volumes of data may limit their advantage over traditional methods in cases where data availability is limited. Our proposed method, on the other hand, demonstrates good performance on individual data, suggesting that it can effectively learn temporal features from small datasets.

3.3.2. Comparison of spatial features

In order to qualitatively compare the spatial features from the three methods, this work applies SDL, STAAE, and MAMSM to the same dataset and obtains the group averaged results, as shown in Figure 8. The GLM templates were derived by summarizing a large amount of individual data and were subsequently employed for the purpose of comparing the performance of FBNs generated through various methods. Our results demonstrate that the activation maps obtained through MAMSM exhibit greater resemblance to the GLM templates.

Quantitatively, we also used the spatial overlap rate as an indicator to compare the FBNs from the three methods and the GLM template. The spatial overlap rate can be used to compare the similarity between two different networks, which is defined as follows:

$$\mathbf{OR}(\mathbf{N^1}, \mathbf{N^2}) = \frac{\sum_{\mathbf{i}=1}^{\mathbf{n}} |\mathbf{N_i}^1 \cap \mathbf{N_i}^2|}{\sum_{\mathbf{i}=1}^{\mathbf{n}} |\mathbf{N_i}^1 \cup \mathbf{N_i}^2|}$$

$N1$, $N2$ are the two brain networks to be compared, n is the number of voxel points of the brain network. The spatial overlap

TABLE 5 Pearson correlation coefficient obtained by SDL, STAAE, and MAMSM.

	E1	E2	G1	G2	W1	W2	W3	W4	W5	W6	W7	W8	
SDL	0.631	0.624	0.483	0.515	0.390	0.356	0.395	0.443	0.419	0.369	0.453	0.379	/
STAAE	0.322	0.246	0.351	0.385	0.195	0.128	0.259	0.272	0.088	0.069	0.197	0.155	/
MAMSM	0.830	0.867	0.848	0.821	0.864	0.870	0.869	0.803	0.849	0.819	0.799	0.869	/
	L1	L2	S1	S2	R1	R2	M1	M2	M3	M4	M5	M6	Ave
SDL	0.603	0.622	0.523	0.673	0.514	0.564	0.658	0.603	0.586	0.493	0.603	0.738	0.527
STAAE	0.606	0.619	0.302	0.651	0.440	0.422	0.429	0.218	0.203	0.189	0.226	0.383	0.306
MAMSM	0.760	0.777	0.812	0.835	0.863	0.868	0.500	0.836	0.862	0.838	0.850	0.875	0.824

TABLE 6 The spatial overlap rate obtained by SDL, STAAE, and MAMSM.

	E1	E2	G1	G2	W1	W2	W3	W4	W5	W6	W7	W8	
SDL	0.150	0.102	0.231	0.200	0.231	0.236	0.266	0.236	0.203	0.225	0.226	0.262	/
STAAE	0.188	**0.234**	0.265	0.210	0.186	0.247	0.209	0.172	0.200	0.263	0.241	0.200	/
MAMSM	**0.221**	0.171	**0.321**	**0.320**	**0.274**	**0.262**	**0.302**	**0.288**	**0.213**	**0.307**	**0.256**	**0.293**	/
	L1	L2	S1	S2	R1	R2	M1	M2	M3	M4	M5	M6	Ave
SDL	0.209	0.177	0.272	0.273	0.244	0.201	0.133	0.146	0.122	0.143	0.146	0.206	0.202
STAAE	0.210	0.265	0.161	0.199	0.205	0.206	0.302	0.293	0.257	0.272	0.273	0.293	0.231
MAMSM	**0.305**	**0.272**	**0.352**	**0.374**	**0.374**	**0.258**	**0.343**	**0.345**	**0.299**	**0.297**	**0.314**	**0.322**	**0.295**

The bold values represent the maximum values of each column.

FIGURE 8
Comparison of FBNs obtained from SDL, STAAE, and MAMSM.

rate of the FBNs obtained from each method and GLM templates are shown in **Table 6**. We can see that the average OR value (0.295) of the brain network obtained by MAMSM is larger than that of STAAE (0.231) and SDL (0.202), which proves that the MAMSM proposed in this manuscript is superior to STAAE and SDL.

4. Discussion and conclusion

In this study, the multi-head attention mechanism and mask training method were applied to the analysis of tfMRI data, and a new loss function was constructed by task design curves for the mapping of functional brain networks. The multi-head attention mechanism helps the model better understand the situation where the same signal value in tfMRI signals may represent different states. Meanwhile, a mask training method was adopted to learn the relationship between the contexts of input sequences, and by combining a continuous mask and a discrete mask, deeper-level features were learned. The experimental results demonstrated that these techniques can improve the model's performance. By analyzing the comparison results of the intermediate features (attention-score, average-result) outputted from the model and the task design curves, it can be seen that the proposed model can better understand the tfMRI signals and the derived features are interpretable. The attention-score extracted after the model was trained represented the weight scores of different locations in each tfMRI sequence. The region with the highest score in the attention-score bears close resemblance to the area with the most significant alteration in the task design curves. The average-result obtained by simply sliding the attention-score achieved higher similarity with the task design curves than the results obtained by other methods.

We also leveraged prior knowledge (Task designs) to guide the model to learn the more efficient features, the task designs were introduced to build a new loss function which optimizes the model by cosine similarity error and MSE error. By analyzing the results, we found that this new loss function can improve the performance of the model. Other methods usually ignored the prior knowledge in their model, and experimental results show that MAMSM achieves better results than other methods when using the new loss function.

The experimental results show that the proposed method can achieve better generalization performance on smaller sample size, compared to other deep learning methods which require large amounts of data to achieve better results, such as STAAE (Dong et al., 2020b), ResAE (Dong et al., 2020a), Dvae (Qiang et al., 2020) and so on. Due to the characteristics of medical image data, such as high confidentiality and small sample size, the method proposed in this manuscript can have better development prospects in the future.

It is important to note that this study has certain limitations. Firstly, the relatively small size of the dataset employed may introduce noise when aggregating across groups, potentially impacting the outcomes of the brain network analyses. Furthermore, the present methodology places greater emphasis on temporal features of tfMRI data, and future investigations may benefit from incorporating a combination of convolutional neural network (CNN) models (Ronneberger et al., 2015; Liu et al., 2022) and visual transformer (VIT) models (Dosovitskiy et al., 2020; Liu et al., 2021) to extract spatial features, which may achieve better results. Additionally, the precise functional significance of some brain networks identified in the results is not fully understood at present, and hence, further research is warranted to explore the functional areas and meanings attributed to these networks.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://humanconnectome.org/study/hcp-young-adult/document/q3-data-release.

Ethics statement

The studies involving human participants were reviewed and approved by the dataset is public and has been approved by its Ethics Committee. The patients/participants provided their written informed consent to participate in this study.

Author contributions

MH: methodology, formal analysis, software, visualization, writing—original draft, and writing—review and editing. XH: visualization and writing—original draft. EG, NQ, and XZ: writing—review and editing. ZW: software and visualization. ZK: validation and visualization. BG: conceptualization, methodology, writing—review and editing, funding acquisition, resources, and supervision. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the National Natural Science Foundation of China (no. 61976131) and the Fundamental Research Funds for Central Universities (JK202205022).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Barch, D. M., Burgess, G. C., Harms, M. P., Petersen, S. E., Schlaggar, B. L., Corbetta, M., et al. (2013). Function in the human connectome: Task-fMRI and individual differences in behavior. *Neuroimage* 80, 169–189.

Beckmann, C. F., DeLuca, M., Devlin, J. T., and Smith, S. M. (2005). Investigations into resting-state connectivity using independent component analysis. *Philos. Trans. R. Soc. B Biol. Sci.* 360, 1001–1013. doi: 10.1098/rstb.2005.1634

Beckmann, C. F., Jenkinson, M., and Smith, S. M. (2003). General multilevel linear modeling for group analysis in FMRI. *Neuroimage* 20, 1052–1063. doi: 10.1016/S1053-8119(03)00435-X

Calhoun, V. D., and Adali, T. (2012). Multisubject independent component analysis of fMRI: A decade of intrinsic networks, default mode, and neurodiagnostic discovery. *IEEE Rev. Biomed. Eng.* 5, 60–73. doi: 10.1109/RBME.2012.2211076

Canario, E., Chen, D., and Biswal, B. (2021). A review of resting-state fMRI and its use to examine psychiatric disorders. *Psychoradiology* 1, 42–53.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., et al. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv* arXiv: 1406.1078. doi: 10.3115/v1/D14-1179 [Preprint].

Chung, Y.-A., Zhang, Y., Han, W., Chiu, C.-C., Qin, J., Pang, R., et al. (2021). "W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training," in *Proceedings of the 2021 IEEE automatic speech recognition and understanding workshop (ASRU)* (Cartagena: IEEE), 244–250. doi: 10.1109/ASRU51503.2021.9688253

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* arXiv: 1810.04805. [Preprint].

Dong, Q., Qiang, N., Lv, J., Li, X., Liu, T., and Li, Q. (2020b). "Spatiotemporal attention autoencoder (STAAE) for ADHD classification," in *Proceedings of the 23rd international conference, medical image computing and computer assisted intervention–MICCAI 2020* (Lima: Springer), 508–517. doi: 10.1007/978-3-030-59728-3_50

Dong, Q., Qiang, N., Lv, J., Li, X., Liu, T., and Li, Q. (2020a). "Discovering functional brain networks with 3D residual autoencoder (ResAE)," in *Proceedings of the 23rd international conference, medical image computing and computer assisted intervention–MICCAI 2020* (Lima: Springer), 498–507. doi: 10.1007/978-3-030-59728-3_49

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* arXiv: 2010.11929. [Preprint].

Ge, B., Makkie, M., Wang, J., Zhao, S., Jiang, X., Li, X., et al. (2016). Signal sampling for efficient sparse representation of resting state FMRI data. *Brain Imaging Behav.* 10, 1206–1222. doi: 10.1007/s11682-015-9487-0

Graves, A., Fernández, S., and Schmidhuber, J. (2005). "Bidirectional LSTM networks for improved phoneme classification and recognition," in *Proceedings of the 15th international conference, artificial neural networks: Formal models and their applications–ICANN* (Warsaw: Springer), 799–804.

Güçlü, U., and Van Gerven, M. A. (2017). Modeling the dynamics of human brain activity with recurrent neural networks. *Front. Comput. Neurosci.* 11:7. doi: 10.3389/fncom.2017.00007

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. (2022). "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (New Orleans, LA: IEEE), 16000–16009.

He, M., Hou, X., Wang, Z., Kang, Z., Zhang, X., Qiang, N., et al. (2022). "Multi-head attention-based masked sequence model for mapping functional brain networks," in *Proceedings of the 25th international conference, medical image computing and computer assisted intervention–MICCAI* (Singapore: Springer), 295–304. doi: 10.1007/978-3-031-16431-6_28

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.* 9, 1735–1780. doi: 10.1162/neco.1997.9.8.1735

Huang, H., Hu, X., Zhao, Y., Makkie, M., Dong, Q., Zhao, S., et al. (2017). Modeling task fMRI data via deep convolutional autoencoder. *IEEE Trans. Med. Imaging* 37, 1551–1561. doi: 10.1109/TMI.2017.2715285

Jiang, X., Yan, J., Zhao, Y., Jiang, M., Chen, Y., Zhou, J., et al. (2023). Characterizing functional brain networks via spatio-temporal attention 4D convolutional neural networks (STA-4DCNNs). *Neural Netw.* 158, 99–110. doi: 10.1016/j.neunet.2022.11.004

Jiang, X., Zhang, T., Zhang, S., Kendrick, K. M., and Liu, T. (2021). Fundamental functional differences between gyri and sulci: Implications for brain function, cognition, and behavior. *Psychoradiology* 1, 23–41. doi: 10.1093/psyrad/kkab002

LaConte, S., Strother, S., Cherkassky, V., Anderson, J., and Hu, X. (2005). Support vector machines for temporal classification of block design fMRI data. *Neuroimage* 26, 317–329. doi: 10.1016/j.neuroimage.2005.01.048

Lee, Y.-B., Lee, J., Tak, S., Lee, K., Na, D. L., Seo, S. W., et al. (2016). Sparse SPM: Group sparse-dictionary learning in SPM framework for resting-state functional connectivity MRI analysis. *Neuroimage* 125, 1032–1045. doi: 10.1016/j.neuroimage.2015.10.081

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., et al. (2021). "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision* (Montreal, QC: IEEE), 10012–10022. doi: 10.1109/ICCV48922.2021.00986

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. (2022). "A convnet for the 2020s," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (New Orleans, LA: IEEE), 11976–11986. doi: 10.1109/CVPR52688.2022.01167

Lv, J., Jiang, X., Li, X., Zhu, D., Chen, H., Zhang, T., et al. (2015). Sparse representation of whole-brain fMRI signals for identification of functional networks. *Med. Image Anal.* 20, 112–134.

Lv, J., Jiang, X., Li, X., Zhu, D., Zhang, S., Zhao, S., et al. (2014). Holistic atlases of functional networks and interactions reveal reciprocal organizational architecture of cortical function. *IEEE Trans. Biomed. Eng.* 62, 1120–1131. doi: 10.1109/TBME.2014.2369495

McKeown, M. J. (2000). Detection of consistently task-related activations in fMRI data with hybrid independent component analysis. *Neuroimage* 11, 24–35. doi: 10.1006/nimg.1999.0518

Mourao-Miranda, J., Reynaud, E., McGlone, F., Calvert, G., and Brammer, M. (2006). The impact of temporal compression and space selection on SVM analysis of single-subject and multi-subject fMRI data. *Neuroimage* 33, 1055–1065. doi: 10.1016/j.neuroimage.2006.08.016

Park, H.-J., and Friston, K. (2013). Structural and functional brain networks: From connections to cognition. *Science* 342:1238411. doi: 10.1126/science.1238411

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830.

Power, J. D., Fair, D. A., Schlaggar, B. L., and Petersen, S. E. (2010). The development of human functional brain networks. *Neuron* 67, 735–748. doi: 10.1016/j.neuron.2010.08.017

Qiang, N., Dong, Q., Ge, F., Liang, H., Ge, B., Zhang, S., et al. (2020). Deep variational autoencoder for mapping functional brain networks. *IEEE Trans. Cogn. Dev. Syst.* 13, 841–852. doi: 10.1109/TCDS.2020.3025137

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-net: Convolutional networks for biomedical image segmentation," in *Proceedings of the 18th international conference, medical image computing and computer-assisted intervention–MICCAI* (Munich: Springer), 234–241. doi: 10.1007/978-3-319-24574-4_28

Schuster, M., and Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* 45, 2673–2681. doi: 10.1109/78.650093

Shen, H., Xu, H., Wang, L., Lei, Y., Yang, L., Zhang, P., et al. (2017). Making group inferences using sparse representation of resting-state functional mRI data with application to sleep deprivation. *Hum. Brain Mapp.* 38, 4671–4689. doi: 10.1002/hbm.23693

Sinha, K., Jia, R., Hupkes, D., Pineau, J., Williams, A., and Kiela, D. (2021). Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. *arXiv* arXiv: 2104.06644. doi: 10.18653/v1/2021.emnlp-main.230 [Preprint].

Smith, S. M., Hyvärinen, A., Varoquaux, G., Miller, K. L., and Beckmann, C. F. (2014). Group-PCA for very large fMRI datasets. *Neuroimage* 101, 738–749. doi: 10.1016/j.neuroimage.2014.07.051

Sporns, O., and Betzel, R. F. (2016). Modular brain networks. *Annu. Rev. Psychol.* 67, 613–640. doi: 10.1146/annurev-psych-122414-033634

Thirion, B., and Faugeras, O. (2003). Dynamical components analysis of fMRI data through kernel PCA. *Neuroimage* 20, 34–49. doi: 10.1016/S1053-8119(03)00316-1

Tong, Z., Song, Y., Wang, J., and Wang, L. (2022). Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *arXiv* arXiv: 2203.12602. [Preprint].

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E., Yacoub, E., Ugurbil, K., et al. (2013). The WU-Minn human connectome project: An overview. *Neuroimage* 80, 62–79. doi: 10.1016/j.neuroimage.2013.05.041

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *Adv. Neural Inform. Process. Syst.* 30.

Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., et al. (2022). "Simmim: A simple framework for masked image modeling," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (New Orleans, LA: IEEE), 9653–9663.

Yan, J., Chen, Y., Xiao, Z., Zhang, S., Jiang, M., Wang, T., et al. (2022). Modeling spatio-temporal patterns of holistic functional brain networks via multi-head guided attention graph neural networks (Multi-Head GAGNNs). *Med. Image Anal.* 80:02518. doi: 10.1016/j.media.2022.102518

Zhang, S., Li, X., Lv, J., Jiang, X., Guo, L., and Liu, T. (2016). Characterizing and differentiating task-based and resting state fMRI signals via two-stage

sparse representations. *Brain Imaging Behav.* 10, 21–32. doi: 10.1007/s11682-015-9359-7

Zhang, W., Lv, J., Li, X., Zhu, D., Jiang, X., Zhang, S., et al. (2018). Experimental comparisons of sparse dictionary learning and independent component analysis for brain network inference from fMRI data. *IEEE Trans. Biomed. Eng.* 66, 289–299. doi: 10.1109/TBME.2018.2831186

Zhao, Y., Li, X., Zhang, W., Zhao, S., Makkie, M., Zhang, M., et al. (2018). "Modeling 4d fMRI data via spatio-temporal convolutional neural networks (ST-CNN)," in *Proceedings of the 21st international conference, medical image computing and computer assisted intervention–MICCAI 2018* (Granada: Springer), 181–189.

Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., et al. (2021). ibot: Image bert pre-training with online tokenizer. *arXiv* arXiv: 2111.07832. [Preprint].