LOGO IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING, VOL. XX, NO. XX, XXXX 1

Explainable Multimodal Deep Dictionary Learning to Capture Developmental Differences from Three fMRI Paradigms

Lan Yang, Chen Qiao, Huiyu Zhou, Vince D. Calhoun, Julia M. Stephen, Tony W. Wilson and Yuping Wang

***Abstract*—Objective: Multimodal-based methods show great potential for neuroscience studies by integrating complementary information. There has been less multimodal work focussed on brain developmental changes. Methods: We propose an explainable multimodal deep dictionary learning method to uncover both the commonality and specificity of different modalities, which learns the shared dictionary and the modality-specific sparse representations based on the multimodal data and their encodings of a sparse deep autoencoder. Results: By regarding three fMRI paradigms collected during two tasks and resting state as modalities, we apply the proposed method on multimodal data to identify the brain developmental differences. We found that both children and young adults prefer to switch among states during two tasks while staying within a particular state during rest, but the difference is that children possess more diffuse functional connectivity patterns while young adults have more focused functional connectivity patterns. Conclusion and Significance: To uncover the commonality and specificity of three fMRI paradigms to developmental differences, multimodal data and their encodings are used to train the shared dictionary and the modality-specific sparse representations. Identifying brain network differences helps to understand how the neural circuits and brain networks form and develop with age.**

***Index Terms*—Explainablity, Multimodal dictionary learning, Dynamic functional connectivity, Brain development**

I. Introduction

This research was supported by the National Natural Science Foundation of China (No. 12090021 and No. 12271429), the National Key Research and Development Program of China (No. 2020AAA0106302), the Natural Science Basic Research Program of Shaanxi (No. 2022JM-005), and was partly supported by the National Institutes of Health (R01 MH104680, R01 GM109068, R01 MH121101, R01 MH116782, R01 MH118013 and P20-GM144641) and the HPC Platform, Xi'an Jiaotong University. (Corresponding author: Chen Qiao and Yuping Wang)

Lan Yang and Chen Qiao are with the School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an 710049 P.R. China (e-mail: qiaochen@mail.xjtu.edu.cn).

Huiyu Zhou is with the School of Computing and Mathematical Sciences, University of Leicester, United Kingdom.

Vince D. Calhoun is with the Tri-Institutional Center for Translational Research in Neuroimaging and Data Science (TReNDS), Georgia State University, Georgia Institute of Technology, Emory University, Atlanta, GA 30030.

Julia M. Stephen is with the Mind Research Network, Albuquerque, NM 87106.

Tony W. Wilson is with the Institute for Human Neuroscience, Boys Town National Research Hospital, Boys Town, NE 68010.

Yuping Wang is with the Department of Biomedical Engineering, Tulane University, New Orleans, LA 70118 USA (e-mail: wyp@tulane.edu).

NORMAL brain development is a complex process, from the establishment of basic cognitive functions in childhood to the gradual maturity of more complex self-regulatory functions throughout adolescence [1]–[3]. Functional magnetic resonance imaging (fMRI) can capture hemodynamic responses to neuronal activities by measuring the blood oxygenation level-dependent (BOLD) signal, based on which the changes in neural interaction and integration between functionally interconnected regions with development can be revealed [4]. Compared with BOLD signals, dynamic functional connectivity (dFC) measured by a sliding window approach can reflect time-varying dependencies between spatially separated brain regions. It helps to quantify the changes of correlation strength between functional activities of paired brain regions over time. Thus, there has been growing interest in identification of the recurring whole-brain functional connectivity patterns (i.e., states) based on dFC recently. These studies aim to divide the whole-brain dFC profiles into distinct states observed reliably across subjects throughout the fMRI scans [5]–[8]. It enables us to investigate the differences of states related to brain development, capture the transition mechanism among these states, and provide insights into neural brain dynamics from the perspective of functional connectivity [4], [5].

Compared with single modality methods for fMRI analysis, multimodal-based methods can take advantage of complementary information provided by different modalities. Studies have shown that integrating the multimodal prior or combining the complementary information from diverse modalities can promote model enhancement and diagnosis [9]–[11]. Many methods have been extended for multimodal data integration including multitask learning, linear regression, neural network, support vector machine, and dictionary learning [9]–[14]. Due to the ability to reduce dimensionality and identify the reoccurring patterns of dFC [4], multimodal dictionary learning methods have attracted considerable attention. For example, Li et al. [11] proposed a multimodal discriminative dictionary learning (mSCDDL) method based on a weighted combination strategy, and further applied it to fuse information from structural magnetic resonance imaging and fluorodeoxyglucose positron emission tomography for Alzheimer's disease classification. In [15], a ℓ_1-norm regularized dictionary learning approach was proposed to identify the epilepsy-related dFC states, where the time courses representative of epileptic activity extracted by electroencephalogram are incorporated into the fMRI for dFC state analysis. In [16], multimodal dictionary learning was applied to the diagnosis of schizophrenia, which embeds the correlation information of multimodal data into the learning model. Additionally, to achieve the nonlinearity or higher-level features of data, Li et al. [10] improved the mSCDDL with the multi-feature kernel trick to obtain the nonlinear representations of data. D'Souza et al. [17] proposed a framework for Autism Spectrum Disorder's diagnosis, which couples a structurally-regularized dynamic dictionary learning model (sr-DDL) with a deep network to predict behavioral scores, where the dFCs of fMRI were decomposed by sr-DDL while constraining the decomposition by the FCs of diffusion tensor imaging.

Of particular note, the aforementioned methods either fail to uncover both commonality and specificity of different modalities, or overlook the nonlinear higher-level features of data, or have difficulty in explainability (i.e., it fails to identify the reoccurring patterns of dFC, or brain regions and FCs related to development or disease). To address these issues, we propose an explainable multimodal deep dictionary learning (EMDDL) method, which connects the multimodal dictionary learning in the original space and the encoding space through a sparse deep autoencoder (sDAE). Within this framework, all modalities share the same dictionary to reveal the inherent commonality. To achieve the specificity of each modality, Fisher cost is used to constrain the sparse representations due to its ability to learn the modality-specific features by avoiding the overlap of neighboring pairs between different modalities. Moreover, the shared dictionary and the modality-specific sparse representations are learned based on the multimodal data and their encodings of the sDAE. In this way, multimodal dictionary learning can attain the nonlinear higher-level features while reconstructing the original data for identifying the reoccurring patterns or functional connectivity related to development. To maintain the complex relationships among subjects, a hypergraph Laplacian regularization is used, which helps to enhance the learning ability through prior knowledge.

We apply EMDDL to the multimodal data from Philadelphia Neurodevelopmental Cohort (PNC) to recognize the developmental differences between children and young adults, where the three fMRI paradigms collected during two tasks and resting state are regarded as modalities. We found that both children and young adults tend to switch frequently among states during two tasks and stay within a particular state during rest. The main difference is that children have more diffuse functional connectivity patterns while young adults possess more focused functional connectivity patterns under three fMRI paradigms. Besides, the differences in functional connectivity between children and young adults are mainly related to information processing, cognition, emotion, and working memory under three fMRI paradigms.

II. Preliminary Work

In this section, some preliminary work is presented including hypergraph learning to preserve the higher-order relationships among subjects and Fisher cost to extract modality-specific features.

A. Hypergraph Learning

Given that the traditional graph learning loses information inevitably by squeezing the complex relationships into pairwise ones, hypergraph has been widely applied to identify the high-order relationships among subjects [18], [19]. Generally, a hypergraph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{W})$ consists of three parts, namely, the vertex set $\mathcal{V} = \{\mathcal{V}_i | i = 1, 2, \cdots, N_v\}$, the hyperedge set $\mathcal{E} = \{\mathcal{E}_i | i = 1, 2, \cdots, N_e\}$ and the hyperedge weight $\mathcal{W} = \{\mathcal{W}_i | i = 1, 2, \cdots, N_e\}$. To represent the relationships between hyperedges and vertices, the incidence matrix $\mathcal{H} \in \mathbb{R}^{N_v \times N_e}$ of hypergraph $\mathcal{G}$ is defined as

$$\mathcal{H}(\mathcal{V}_i, \mathcal{E}_j) = \begin{cases} 1 & \mathcal{V}_i \in \mathcal{E}_j \\ 0 & otherwise \end{cases}$$

where the (i, j)-th entry of $\mathcal{H}$ denotes whether the i-th vertex belong to the j-th hyperedge. Based on the incidence matrix $\mathcal{H}$, the degree of the i-th vertex $d_{\mathcal{V}_i} = \sum_{\mathcal{E}_j \in \mathcal{E}} \mathcal{W}_j \mathcal{H}(\mathcal{V}_i, \mathcal{E}_j)$ and the degree of the i-th hyperedge $d_{\mathcal{E}_i} = \sum_{\mathcal{V}_j \in \mathcal{V}} \mathcal{H}(\mathcal{V}_j, \mathcal{E}_i)$ can be obtained. Then the diagonal matrices $\mathcal{D}_v \in \mathbb{R}^{N_v \times N_v}$ and $\mathcal{D}_e \in \mathbb{R}^{N_e \times N_e}$ are composed of the degree of all vertices and hyperedges respectively. Specifically, the i-th diagonal element of $\mathcal{D}_v$ and $\mathcal{D}_e$ are $d_{\mathcal{V}_i}$ and $d_{\mathcal{E}_i}$ respectively.

To construct a hypergraph, the k nearest neighbor strategy is usually applied because the geometric structure relationship among data can be approximately represented by the nearest neighbor graph [18], [20]. Specifically, for a chosen vertex, the distances between the chosen vertex and other vertices are calculated, and then the k nearest vertices are connected by a hyperedge. The weight of the i-th hyperedge is $\mathcal{W}_i = \frac{1}{k(k-1)} \sum_{\{\mathcal{V}_j, \mathcal{V}_l\} \in \mathcal{E}_i} \exp\left(-\frac{||\mathcal{V}_j - \mathcal{V}_l||_2^2}{\sigma_i}\right)$, where $\sigma_i = \frac{\sum_{\{\mathcal{V}_j, \mathcal{V}_l\} \in \mathcal{E}_i} ||\mathcal{V}_j - \mathcal{V}_l||_2^2}{k(k-1)}$. To obtain the diagonal matrix $\mathcal{W}_h \in \mathbb{R}^{N_e \times N_e}$, the hyperedge weight $\mathcal{W}_i$ is arrayed as the i-th diagonal element of $\mathcal{W}_h$. By analogizing the definition of a simple graph Laplacian matrix [21], hypergraph Laplacian matrix is defined as

$$L^h = \mathcal{D}_v - \mathcal{S} \tag{1}$$

where $\mathcal{S} = \mathcal{H} \mathcal{W}_h \mathcal{D}_e^{-1} \mathcal{H}^{\mathrm{T}}$ is the similarity matrix to define the similarity between each pair of vertices.

Compared with the traditional graph Laplacian regularization, hypergraph Laplacian regularization has the characteristics of preserving complex local geometric structure and incorporating the higher-order relationships among subjects, which are conducive to classification or clustering tasks in FC or dFC analysis [19].

B. Fisher cost

The Fisher discrimination criterion is to cluster the samples in the same modality and keep the samples in different modalities as far away from each other as possible, which helps to extract features corresponding to the specific modality [22]–[24]. Assume that the multimodal data $X = (x_1, x_2, \cdots, x_N) \in \mathbb{R}^{p \times N}$ contains M modalities with N_m samples belonging to the m-th modality $\mathcal{N}_m$ and $\sum_{m=1}^{M} N_m = N$, where p-dimensional vector x_n is the n-th sample of X. The within-modality scatter matrix S_w and the between-modality scatter matrix S_b of samples are defined as

$$S_w(X) = \sum_{m=1}^{M} \sum_{n \in \mathcal{N}_m} (x_n - \mu_m)(x_n - \mu_m)^{\mathrm{T}}$$

$$S_b(X) = \sum_{m=1}^{M} N_m(\mu_m - \mu)(\mu_m - \mu)^{\mathrm{T}}$$

where $\mu_m = \frac{1}{N_m}\sum_{n\in\mathcal{N}_m} x_n$ and $\mu = \frac{1}{N}\sum_{n=1}^{N} x_n$ are the modality mean and the overall mean respectively. Then, the Fisher cost is as follows

$$\mathcal{F}(X) = tr\big(S_w(X)\big) - tr\big(S_b(X)\big) + ||X||_F^2$$

in which the Frobenius norm $||\cdot||_F$ is to ensure the convexity of the cost function [24].

To get a more concise expression and facilitate calculation [22], the Fisher cost $\mathcal{F}(X)$ can be rewritten as

$$\mathcal{F}(X) = tr\big(XFX^{\mathrm{T}}\big) \tag{2}$$

where $F = 2I - 2F_1 + F_2 \in \mathbb{R}^{N\times N}$ with $I \in \mathbb{R}^{N\times N}$ being the identity matrix, $F_1 \in \mathbb{R}^{N\times N}$ being defined as

$$F_1(i,j) = \begin{cases} \frac{1}{N_m} & i,j \in \mathcal{N}_m \\ 0 & otherwise \end{cases}$$

and $F_2 \in \mathbb{R}^{N\times N}$ with each component of it being $1/N$.

III. Methodology

The details of EMDDL and the corresponding optimization algorithm are presented in this section, which can learn the shared dictionary and modality-specific sparse representations in both the original space and the encoding space.

A. Explainable Multimodal Deep Dictionary Learning

Multimodal dictionary learning methods can not only embed the high-dimensional features into low-dimensional space, but also boost learning performance with the combination of multiple modalities [12]. However, most of the existing methods either cannot simultaneously reveal the inherent commonality and specificity of different modalities, or overlook the nonlinear higher-level features of data, or have difficulty in explainability. To address these problems, we propose EMDDL which couples multimodal dictionary learning with sDAE. Specifically, by sharing the same dictionary through all modalities to capture the inherent commonality and constraining sparse representations with Fisher cost to obtain the specificity of each modality, the inherent commonality and the specificity of different modalities can be concurrently achieved in multimodal dictionary learning. Moreover, to achieve the nonlinear higher-level features of data and reconstruct the original data to identify the developmental differences in reoccurring patterns or FCs, both the shared dictionary and the modality-specific sparse representations are learned not only in the original space, but also in the encoding space of the sDAE at the same time. By alternating minimization algorithms, the sDAE, dictionary, and sparse representations can be sequentially obtained. The flowchart of EMDDL is shown in Fig. 1.

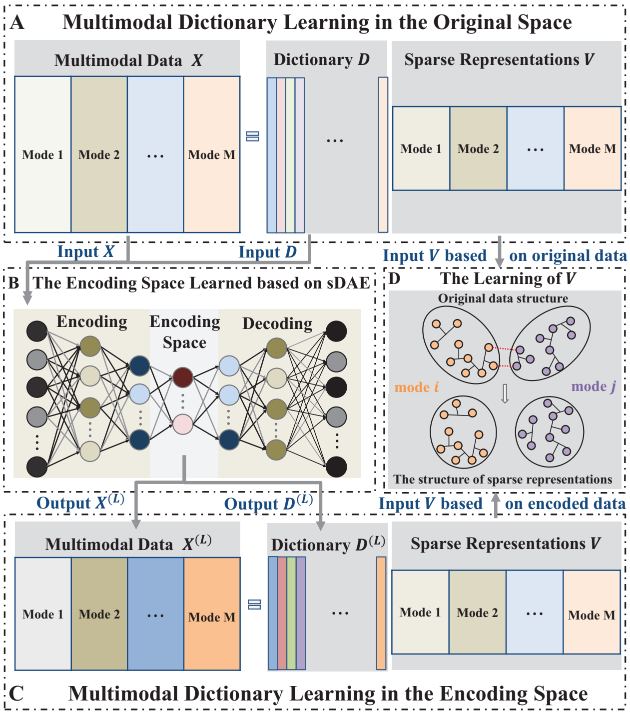

Fig. 1: The flowchart of EMDDL.

Suppose that there are M modalities with N_m samples belonging to the m-th modality $\mathcal{N}_m$ and the training data $X = (X_{(1)}, X_{(2)}, \cdots, X_{(M)}) \in \mathbb{R}^{p\times N}$ is composed of these M modalities, where $N = \sum_{m=1}^{M} N_m$ and the m-th modality is $X_{(m)} = (x_1^{(m)}, x_2^{(m)}, \cdots, x_{N_m}^{(m)}) \in \mathbb{R}^{p\times N_m}$. Besides, the sDAE contains $2L+1$ layers with $r^{(l)}$ neurons in the l-th layer and $r^{(2L-l)} = r^{(l)}$ holds for $l = 0, 1, \cdots, 2L$.

EMDDL contains two parts including J_{sDAE} and J_{MDL}, where J_{sDAE} is to efficiently learn the nonlinear higher-level features of data and J_{MDL} is to train multimodal dictionary learning in both the original space and the encoding space. The objective function of EMDDL is defined as

$$\begin{aligned} \min_{\{\tilde{W}^{(l)}\}_{l=1}^{2L}, D, V} \quad & J_{obj} = J_{sDAE} + J_{MDL} \\ s.t. \quad & ||d_k||_2^2 \le 1, \quad \forall \quad k = 1, 2, \cdots, K \end{aligned} \tag{3}$$

where J_{sDAE} and J_{MDL} are defined as

$$\begin{aligned} J_{sDAE} =& J_{recon} + \lambda_1 J_{KL} + \lambda_2 J_{W_F} \\ =& \frac{1}{2N}||X^{(2L)} - X||_F^2 + \lambda_1 \sum_{l=1}^{2L-1}\sum_{j=1}^{r^{(l)}} KL(\rho||\rho_j^{(l)}) \\ &+ \frac{\lambda_2}{2}\sum_{l=1}^{2L}||\tilde{W}^{(l)}||_F^2 \end{aligned} \tag{4}$$

$$\begin{aligned} J_{MDL} =& J_{MDL_O} + J_{MDL_E} + \lambda_3 J_{Fisher} + \lambda_4 J_{hyperL} \\ &+ \lambda_5 J_{V_F} + \lambda_6 J_{V_{\ell_1}} \\ =& \frac{1}{2N}||X - DV||_F^2 + \frac{1}{2N}||X^{(L)} - D^{(L)}V||_F^2 \\ &+ \frac{\lambda_3}{2}tr(VHV^T) + \frac{\lambda_4}{2}tr(VLV^T) + \frac{\lambda_5}{2}||V||_F^2 \\ &+ \lambda_6||V||_1 \end{aligned} \tag{5}$$

where $X^{(2L)} \in \mathbb{R}^{r^{(2L)} \times N}$ is the reconstruction of the input data X by sDAE, $D = (d_1, d_2, \cdots, d_K) \in \mathbb{R}^{p \times K}$ is the dictionary with K atoms in the original space, $X^{(L)} \in \mathbb{R}^{r^{(L)} \times N}$ and $D^{(L)} \in \mathbb{R}^{r^{(L)} \times K}$ are the encoding of X and D respectively (i.e., its outputs in the L-th layer), $V = (V_{(1)}, V_{(2)}, \cdots, V_{(M)}) \in \mathbb{R}^{K \times N}$ consists of each $V_{(m)} = (v_1^{(m)}, v_2^{(m)}, \cdots, v_{N_m}^{(m)}) \in \mathbb{R}^{K \times N_m}$ being the sparse representation corresponding to the m-th modality in both the original space and the encoding space. $\tilde{W}^{(l)} = (W^{(l)}, b^{(l)}) \triangleq \{\tilde{W}_{ij}^{(l)}\} \in \mathbb{R}^{r^{(l)} \times r^{(l-1)}+1}$ for $l = 1, 2, \cdots, 2L$, in which $W^{(l)} \in \mathbb{R}^{r^{(l)} \times r^{(l-1)}}$ and $b^{(l)} \in \mathbb{R}^{r^{(l)}}$ are the connection weight matrix and bias of sDAE between l-th layer and $(l-1)$-th layer respectively. As defined in Appendix I-A, Kullback-Leibler divergence $KL(\rho||\rho_j^{(l)})$ measures the difference between two Bernoulli distributions with mean ρ and $\rho_j^{(l)}$, where ρ is a sparsity hyperparameter and $\rho_j^{(l)}$ is the average activation of neuron j in the l-th layer of sDAE. Similar to the definition of F in (2), $H \in \mathbb{R}^{N \times N}$ is given by $H = I - 2H_1 + H_2$, where $I \in \mathbb{R}^{N \times N}$ is the identity matrix, $H_1 \in \mathbb{R}^{N \times N}$ is defined as

$$H_1(i,j) = \begin{cases} \frac{1}{N_m} & i,j \in \mathcal{N}_m \\ 0 & otherwise \end{cases}$$

and the each component of $H_2 \in \mathbb{R}^{N \times N}$ is $1/N$. $L \in \mathbb{R}^{N \times N}$ consists of hypergraph Laplacian matrix of all modalities, which is defined as

$$L = \begin{pmatrix} L_{(1)}^h & 0 & \cdots & 0 \\ 0 & L_{(2)}^h & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & L_{(M)}^h \end{pmatrix}$$

where $L_{(m)}^h \in \mathbb{R}^{N_m \times N_m}$, the hypergraph Laplacian matrix of the m-th modality, is defined by (1). $||V||_1 = \sum_{n=1}^{N} \sum_{k=1}^{K} |V_{nk}|$ with V_{nk} being the k-th element of the n-th column of the matrix V.

In (4), J_{recon} is to train the sDAE by minimizing the error between original data and its reconstruction. J_{KL} is to prevent overfitting of the sDAE by controlling the activation of neurons. Compared with L_1-norm and L_2-norm, Kullback-Leibler divergence has better sparsity ability, which helps to improve model performance, and the details can be seen in Appendix I-A. J_{W_F} is to prevent overfitting of the sDAE by controlling the weights. In (5), J_{MDL_O} is to learn the shared dictionary of all modalities and the modality-specific sparse representations based on the original data. Meanwhile, by encoding the original data and the shared dictionary through the sDAE, J_{MDL_E} is to achieve the multimodal dictionary learning in the encoding space for capturing the nonlinear higher-level features of data. Inspired by [25], we use the same sparse representations to synchronously characterize the local geometric relationships between data and dictionary in the original space as to characterize those between encoded data and encoded dictionary in the encoding space. In other words, our objective is to use the sparse representations to capture the intrinsic local geometric relationships between data and dictionary. It helps to learn the locality-sensitive dictionary, resulting in improved generalization ability in reconstrcution or classification. By clustering the samples within modalities and separating the samples between the modalities, J_{Fisher} helps to learn the modality-specific representations. J_{hyperL} is designed to retain the complex neighborhood relationships of samples hidden in each modality. J_{V_F} guarantees the convexity of Fisher cost and $J_{V_{\ell_1}}$ is to ensure the sparsity. The constraint on dictionary atoms is to prevent sparse representation from being too small due to the large dictionary. In addition, the positive parameters λ_1, λ_2, λ_3, λ_4, λ_5 and λ_6 are used to balance the network fitting, dictionary learning and the complexity of model.

B. Optimization

The alternating minimization algorithm is applied to solve the problem (3) to optimize the parameters $\{\tilde{W}^{(l)}\}_{l=1}^{2L}$, D and V, which contains three parts, i.e., the training of sDAE, the learning of the dictionary, and sparse representations learning.

Denote

$$\begin{cases} \tilde{h}_n^{(l)} = (h_n^{(l)}; 1) \\ z_n^{(l+1)} = \tilde{W}^{(l+1)} \tilde{h}_n^{(l)} & l = 0, 1, \cdots, 2L-1 \\ h_n^{(l+1)} = \varphi(z_n^{(l+1)}) \end{cases}$$

$$\begin{cases} \tilde{g}_k^{(l)} = (g_k^{(l)}; 1) \\ q_k^{(l+1)} = \tilde{W}^{(l+1)} \tilde{g}_k^{(l)} & l = 0, 1, \cdots, L-1 \\ g_k^{(l+1)} = \varphi(q_k^{(l+1)}) \end{cases}$$

where φ is a differentiable activation function which is the sigmoid function in this paper; $h_n^{(0)} = X_n$ and $g_k^{(0)} = d_k$, where X_n is the n-th column of the multimodal data X and d_k is the k-th atom of the dictionary D. $X^{(l)} = (h_1^{(l)}, h_2^{(l)}, \cdots, h_N^{(l)}) \in \mathbb{R}^{r^{(l)} \times N}, l = 1, 2, \cdots, 2L$ and $D^{(l)} = (g_1^{(l)}, g_2^{(l)}, \cdots, g_K^{(l)}) \in \mathbb{R}^{r^{(l)} \times K}, l = 1, 2, \cdots, L$ are the outputs in the l-th layer when the input data are X and D respectively.

1) The Training of sDAE: To optimize the parameters of sDAE $\{\tilde{W}^{(l)}\}_{l=1}^{2L}$ with fixed D and V, problem (3) can be rewritten as

$$\min_{\{\tilde{W}^{(l)}\}_{l=1}^{2L}} \frac{1}{2N} \Big(||X^{(2L)} - X||_F^2 + ||X^{(L)} - D^{(L)}V||_F^2 \Big) + \lambda_1 \sum_{l=1}^{2L-1} \sum_{j=1}^{r^{(l)}} KL(\rho||\rho_j^{(l)}) + \frac{\lambda_2}{2} \sum_{l=1}^{2L} ||\tilde{W}^{(l)}||_F^2 \quad (6)$$

To update the parameters of sDAE $\{\tilde{W}^{(l)}\}_{l=1}^{2L}$, the back-propagation algorithm with gradient descent method is applied. Then, the gradient of $\tilde{W}^{(l)}$ is given by

$$\nabla \tilde{W}^{(l)} = \frac{1}{N} \sum_{n=1}^{N} \Big(\Delta H_n^{(l)} \tilde{h}_n^{(l-1)^{\mathrm{T}}} + I(L-l) \Delta T_n^{(l)} + \lambda_1 I(2L-1-l) \Delta S_n^{(l)} \tilde{h}_n^{(l-1)^{\mathrm{T}}} \Big) + \lambda_2 \tilde{W}^{(l)} \quad (7)$$

in which $\Delta H_n^{(l)}$ is defined as

$$\Delta H_n^{(l)} = \begin{cases} (h_n^{(l)} - x_n) \odot \varphi'(z_n^{(l)}) & l = 2L \\ (W^{(l+1)^{\mathrm{T}}} \Delta H_n^{(l+1)}) \odot \varphi'(z_n^{(l)}) & l = 2L-1, \cdots, 2, 1 \end{cases}$$

where the operation $\odot$ denotes the element-wise multiplication. $I(\cdot)$ is an indicator function defined by

$$I(s) = \begin{cases} 1 & s \geq 0 \\ 0 & otherwise \end{cases}$$

$\Delta T_n^{(l)} = \Delta T_n^{(l)}(0)\tilde{h}_n^{(l-1)^{\mathrm{T}}} - \sum_{k=1}^{K} \Delta T_n^{(l)}(k)\tilde{g}_k^{(l-1)^{\mathrm{T}}}$ with $\Delta T_n^{(l)}(0)$ and $\Delta T_n^{(l)}(k)$ being defined as

$$\Delta T_n^{(l)}(0) = \begin{cases} \left(h_n^{(l)} - D^{(l)}V_n\right) \odot \varphi'(z_n^{(l)}) & l = L \\ \left(W^{(l+1)^{\mathrm{T}}} \Delta T_n^{(l+1)}(0)\right) \odot \varphi'(z_n^{(l)}) & l = L-1, \cdots, 2, 1 \end{cases}$$

$$\Delta T_n^{(l)}(k) = \begin{cases} V_{nk}\left(h_n^{(l)} - D^{(l)}V_n\right) \odot \varphi'(q_k^{(l)}) & l = L \\ & k = 1, 2, \cdots, K \\ \left(W^{(l+1)^{\mathrm{T}}} \Delta T_n^{(l+1)}(k)\right) \odot \varphi'(q_k^{(l)}) & l = L-1, \cdots, 2, 1 \\ & k = 1, 2, \cdots, K \end{cases}$$

in which V_n is the n-th column of V. $\Delta S_n^{(l)}(t)$ is defined as

$$\Delta S_n^{(l)}(t) = \begin{cases} R^{(l)} \odot \varphi'(z_n^{(l)}) & t = 2L - l \\ & l = 2L-1, \cdots, 2, 1 \\ \left(W^{(l+1)^{\mathrm{T}}} \Delta S_n^{(l+1)}(t)\right) \odot \varphi'(z_n^{(l)}) & t = 1, 2, \cdots, 2L-1-l \\ & l = 2L-2, \cdots, 2, 1 \end{cases}$$

where $R^{(l)}$ is a $r^{(l)}$-dimensional column vector with i-th element being $(\frac{-\rho}{\rho_i^{(l)}} + \frac{1-\rho}{1-\rho_i^{(l)}})$, and $\Delta S_n^{(l)} = \sum_{t=1}^{2L-l} \Delta S_n^{(l)}(t)$. The update formula for $\tilde{W}^{(l)}$ is

$$\tilde{W}^{(l)} = \tilde{W}^{(l)} - \eta_1 \nabla \tilde{W}^{(l)}$$

where η_1 is the learning rate.

2) The Learning of Dictionary: To update dictionary D with fixed $\{\tilde{W}^{(l)}\}_{l=1}^{2L}$ and V, problem (3) can be rewritten as

$$\begin{aligned} \min_{D} \quad & \frac{1}{2N}\left(||X - DV||_F^2 + ||X^{(L)} - D^{(L)}V||_F^2\right) \\ s.t. \quad & ||d_k||_2^2 \leq 1, \quad \forall \quad k = 1, 2, \cdots, K \end{aligned} \tag{8}$$

The gradient descent method is used to optimize the above problem and the gradient of D is given by

$$\nabla D = \frac{1}{N}\left((DV - X)V^{\mathrm{T}} + \Delta R\right) \tag{9}$$

in which the k-th column of ΔR is computed by $\sum_{n=1}^{N} \Delta R_n^{(1)}(k)$ and $\Delta R_n^{(l)}(k)$ is defined as

$$\Delta R_n^{(l)}(k) = \begin{cases} W^{(l)^{\mathrm{T}}}\left(V_{nk}\left(D^{(l)}V_n - h_n^{(l)}\right) \odot \varphi'(q_k^{(l)})\right) & l = L \\ & k = 1, 2, \cdots, K \\ W^{(l)^{\mathrm{T}}}\left(\Delta R_n^{(l+1)}(k) \odot \varphi'(q_k^{(l)})\right) & l = L-1, \cdots, 2, 1 \\ & k = 1, 2, \cdots, K \end{cases}$$

The update formula for D is

$$D = D - \eta_2 \nabla D$$

where η_2 is the learning rate. Considering the constraint on the dictionary, each column of the updated dictionary D is normalized to unit length by

$$d_k = \frac{1}{||d_k||_2} d_k \tag{10}$$

3) Sparse Representations Learning: With the fixed $\{\tilde{W}^{(l)}\}_{l=1}^{2L}$ and D, the sparse representations can be obtained by solving the following optimization problem

$$\min_{V} f(V) + g(V) \tag{11}$$

where $f(V)$ and $g(V)$ are

$$\begin{aligned} f(V) = & \frac{1}{2N}\left(||X - DV||_F^2 + ||X^{(L)} - D^{(L)}V||_F^2\right) \\ & + \frac{\lambda_3}{2} tr(VHV^T) + \frac{\lambda_4}{2} tr(VLV^T) + \frac{\lambda_5}{2}||V||_F^2 \\ g(V) = & \lambda_6 ||V||_1 \end{aligned}$$

To ensure the convexity of $f(V)$, $\lambda_5 > \lambda_3 \geq 0$ holds and the details can be seen in Appendix I-B. In problem (11), $f(V)$ is convex and differentiable, while $g(V)$ is convex but nondifferentiable. Thus, the Fast Iterative Shrinkage Thresholding Algorithm (FISTA) [26] is adopted to optimize V. The gradient of $f(V)$ with respect to V is

$$\nabla V = \frac{1}{N}\left(D^{\mathrm{T}}(DV - X) + D^{(L)^{\mathrm{T}}}(D^{(L)}V - X^{(L)})\right) + VS \tag{12}$$

in which $S = \lambda_3 H + \lambda_4 L + \lambda_5 I$. The Lipschitz constant of the gradient ∇V is given by (13) in Appendix I-C. Besides, the soft thresholding function in FISTA is defined as $ST_{\frac{\lambda_6}{L_f}}(\cdot) = sign(\cdot) max\{0, |\cdot| - \frac{\lambda_6}{L_f}\}$ with $|\cdot|$ representing absolute value function. The total optimization process is described in Algorithm 1.

IV. Results and Analysis

In this section, EMDDL is utilized to explore the dynamic functional connectivity changes of brain during two tasks and resting state.

A. Data Acquisition and Preprocessing

PNC is a large scale collaborative project between the Brain Behaviour Laboratory at the University of Pennsylvania and the Children's Hospital of Philadelphia, which contains data collected using three fMRI paradigms from nearly 900 youth aged from 8 to 22, i.e., two tasks including emotion identification (Emoid fMRI) and working memory (Nback fMRI), and resting-state (Rest fMRI) [27]. All fMRI scans were collected on a single 3T Siemens TIM Trio whole-body scanner using a single-shot, interleaved multi-slice, gradient-echo, echo planar imaging sequence. The Emoid fMRI, Nback fMRI and Rest fMRI scan durations were 10.5 minutes (210 TR), 11.6 minutes (231 TR) and 6.2 minutes (124 TR) respectively. During Emoid task, subjects were asked to identify 60

Algorithm 1: EMDDL

Input: Training data: $X = (X_{(1)}, X_{(2)}, \cdots, X_{(M)})$; The parameters of sDAE: $2L$, $\{r^{(l)}\}_{l=0}^{2L}$ and ρ; Regularization coefficients: λ_1, λ_2, λ_3, λ_4, λ_5 and λ_6; Size of the dictionary: K; Learning rate: η_1 and η_2; Initialize $\{\tilde{W}_{(0)}^{(l)}\}_{l=1}^{2L}$, $D_{(0)}$ and $V_{(0)}$ randomly and set $i = 0$

Output: $\{\tilde{W}_{(i)}^{(l)}\}_{l=1}^{2L}$, $D_{(i)}$ and $V_{(i)}$

while *not converged* **do**
  Update sDAE:
  for $l = 2L, 2L-1, \cdots, 1$ **do**
    Compute the gradient $\nabla \tilde{W}_{(i)}^{(l)}$ via (7)
    $\tilde{W}_{(i+1)}^{(l)} \leftarrow \tilde{W}_{(i)}^{(l)} - \eta_1 \nabla \tilde{W}_{(i)}^{(l)}$
  end
  Update dictionary:
  Compute the gradient $\nabla D_{(i)}$ via (9)
  $D_{(i+1)} \leftarrow D_{(i)} - \eta_2 \nabla D_{(i)}$
  Normalize the dictionary $D_{(i+1)}$ via (10)
  Update sparse representations:
  Set $j = 0$, $t_{(0)} = 1$, $V_{(i,j)} = V_{(i)}$ and $Z_{(i,j)} = V_{(i)}$
  Compute $L_{(i)}$ via (13)
  while *not converged* **do**
    Compute the gradient $\nabla V_{(i,j)}$ via (12)
    $Z_{(i,j+1)} \leftarrow ST_{\frac{\lambda_6}{L_{(i)}}}\Big(V_{(i,j)} - \frac{1}{L_{(i)}} \nabla V_{(i,j)}\Big)$
    $t_{(j+1)} \leftarrow \frac{1+\sqrt{1+4t_{(j)}^2}}{2}$
    $V_{(i,j+1)} \leftarrow Z_{(i,j+1)} + \frac{t_{(j)}-1}{t_{(j+1)}}(Z_{(i,j+1)} - Z_{(i,j)})$
    $j \leftarrow j + 1$
  end
  $V_{(i+1)} \leftarrow V_{(i,j)}$
  $i \leftarrow i + 1$
end

faces with neutral, happy, sad, angry, or fearful expressions. During Nback task to probe working memory, subjects were required to respond only when a presented fractal was the same as the one presented in the previous trial. During the resting-state scan, subjects were instructed to stay awake, keep eyes open, fixate on the displayed crosshair, and remain still. Of these, 123 children and 146 young adults completed all three paradigms. By using Statistical Parametric Mapping 12, motion correction, co-registration, spatial normalization to standard Montreal Neurological Institute space (spatial resolution of $3\times3\times3$ mm), and spatial smoothing with a 3 mm full width half maximum Gaussian kernel were implemented. Then, a regression procedure was used to remove the influence of motion and the functional time series were band-pass filtered using a 0.01 Hz to 0.1 Hz frequency range. According to the Power coordinates with a sphere radius parameter of 5 mm [28], 264 regions of interest (ROIs) containing 21384 voxels were extracted. The details of the 264 ROIs are shown in Table 1 of Supplementary material. Every subject file can be reduced to a $264 \times T$ matrix by averaging the time series of all voxels in the same brain region, where the time point T is 210, 231, and 124 for Emoid, Nback, and Rest fMRIs respectively.

We divided 264 ROIs into 13 functional networks to facilitate the understanding of functional connectivity relationships between the ROIs [28]. Among them, 12 functional networks including sensory/somatomotor network (SSN), cingulo-opercular task control network (COTCN), auditory network (AN), default mode network (DMN), memory retrieval network (MRN), visual network (VN), frontoparietal task control network (FPTCN), salience network (SN), subcortical network (SCN), ventral attention network (VAN), dorsal attention network (DAN), and cerebellar network (CN), are mainly associated with the perception of movement, memory, language, vision, cognition and other functions of the brain, while there are 28 ROIs unrelated to any of the above functional networks which belong to the uncertain network (UN).

To capture the dynamic characteristics of the brain, dFC is obtained by calculating the Pearson correlation between the time-courses of the BOLD signals of pair regions within a window [29]–[31]. The details of obtaining the multimodal data can be seen in Appendix I-D. By grid search, we choose window length w_l being 14, 17, and 33 for Emoid, Nback, and Rest fMRIs respectively, and scan length s_l is 1 for all three modalities. Thus, each subject provides a dFC matrix $M_{dFC} \in \mathbb{R}^{C_{264}^2 \times S_l}$ corresponding to Emoid, Nback, and Rest fMRIs, where $C_{264}^2 = 34716$. To reduce the complexity of computation, systematic sampling is used to select 20 sub-sequences from the dFC matrix corresponding to each modality of each subject [4]. Training data contains 80% of the subjects and the remaining subjects are test data.

B. Experimental Results

To evaluate the performance of the algorithm, the signal-to-noise ratio (SNR) [32] is used as evaluation index which is defined as

$$SNR = 10log_{10}\Big(\frac{||X||_F^2}{||X - DV||_F^2}\Big)$$

Given that the grid search method can simply make a complete search over a given hyperparameters space, easily be parallelized to find more stable optimal hyperparameter [33], [34], it is used to select appropriate hyperparameters. Specifically, one of the hyperparameters is selected by the grid search method when other hyperparameters are fixed. By repeating the above process, all hyperparameters are optimized, and the results are shown in Figure 1 of the Supplementary material. There are 7 layers of sDAE with 34716, 10000, 6000, 1000, 6000, 10000, 34716 units respectively. The number of atoms K is 300, the sparsity parameter ρ is 0.1, the regularization coefficients λ_1, λ_2, λ_3, λ_4, λ_5, and λ_6 are 0.0001, 0.0005, 0.0003, 0.0001, 0.0005, and 0.001 respectively, the k nearest neighbor of hypergraph is 9. Because problem (6) and (8) are nonconvex, RMSProp algorithm is used to update $\{\tilde{W}^{(l)}\}_{l=1}^{2L}$ and D due to the better generalization ability and it is less prone to overfitting [35]. For RMSProp algorithm, the learning rates η_1 and η_2 are 0.00005 and 0.00008 respectively, and the square gradient decay rates ξ_1 and ξ_2 are both 0.9.

Based on the optimal hyperparameters, we apply EMDDL to the training data to obtain the dictionary and sparse representations. The learning curves of loss functions and the SNR evaluation on both training data and testing data are shown in Fig. 2. The results testify that the sparse representations can characterize the local geometric relationships between data and dictionary in the two spaces. The SNR of EMDDL, multimodal dictionary learning (MDL) [36] and sparse deep dictionary learning (SDDL) [4] on testing data are shown in Table I. It shows that the multimodal-based methods have better reconstruction ability compared with the single modality methods, and the generalization ability in reconstruction of EMDDL are better than the other two methods. It testifies that integrating the multimodal prior or combining the complementary information from diverse modalities can promote model enhancement.

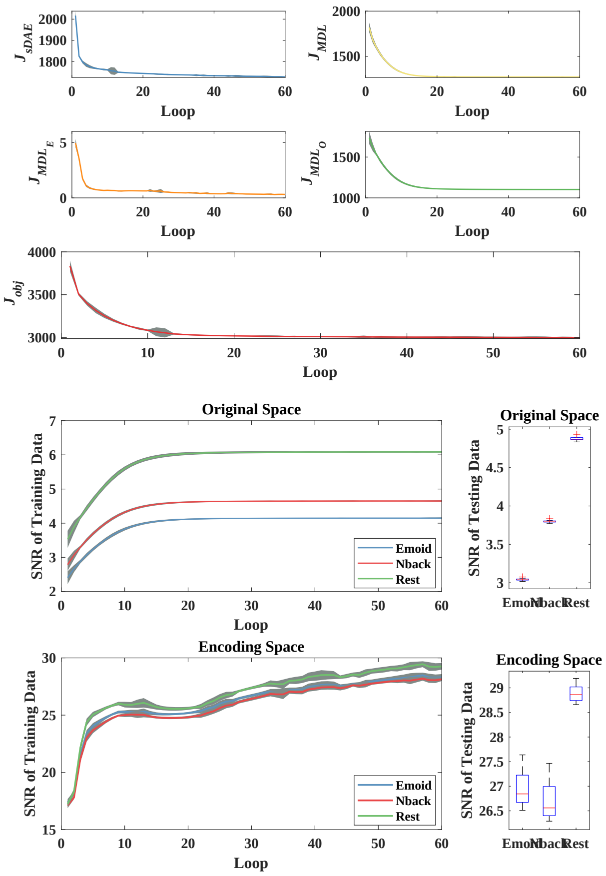

Fig. 2: The learning curves of loss functions and the SNR evaluation on both training data and testing data of EMDDL. The curve is formed by the average of 10 repetitions, and the gray shadow is formed by the standard deviation of 10 repetitions.

C. States Analysis of Multimodal Data

To find the differences in reoccurring patterns of dFC (i.e., states) between children and young adults, k-means clustering method with the cityblock distance metric is used to obtain the reoccurring patterns of each group in each modality [37].

TABLE I: The SNR on testing data of various methods.

	Multimodal based methods		Single modality based methods
Paradigm \ Method / SNR	EMDDL	MDL	SDDL
Emoid	3.2577	2.6416	0.9338
Nback	3.8650	3.2029	1.2219
Rest	4.8046	4.1108	1.0622

Specifically, sparse representations of each group in each modality are clustered, and then states can be obtained by multiplying the dictionary and the cluster centroid. We use the elbow criterion defined as within-cluster sums of distances to estimate the optimal number of dFC states, and the optimal number of dFC states for Emoid, Nback, and Rest fMRIs are 5, 5, and 4 respectively. To test whether the clustering results are consistent in multiple subgroups, we use the kappa coefficient as the indicator [38], and the details can be seen in Appendix I-E. The results indicate that the clustering results obtained from two different subgroups are substantial agreement or perfect agreement in a large probability. For Emoid task, the proportions of each state for children are 14.07%, 22.28%, 17.52%, 25.53%, and 20.61% respectively, while the proportions of each state for young adults are 9.08%, 21.06%, 14.55%, 27.29%, and 28.01% respectively. For Nback task, the proportions in these states for children are 11.14%, 19.19%, 24.23%, 22.97%, and 22.48% respectively, while the proportions in these states for young adults are 7.5%, 14.28%, 22.53%, 25.58%, and 30.1% respectively. For Rest fMRI, the proportions of these states for children are 31.14%, 29.35%, 34.15%, and 5.37% respectively, while the proportions of these states for young adults are 21.64%, 13.87%, 27.43%, and 37.05% respectively.

To further investigate the time occupied divergence of each state, dwell time (DT) and fraction of time (FT) are estimated from the state transition vector [7]. DT represents how long an individual spends in a given state on average, and FT is to describe the total time spent in a given state. For a subject i, DT and FT of k-th state are defined by

$$DT^{state(k)} = mean(TR_{end} - TR_{start})$$

$$FT^{state(k)} = \frac{sum(state\_vector_{(i)} == k)}{Total\ number\ of\ window}$$

where TR_{start} and TR_{end} are computed by

$$TR_{start} = count(difference(state\_vector_{(i)}, k) == 1)$$

$$TR_{end} = count(difference(state\_vector_{(i)}, k) == -1)$$

in which "1" and "-1" mean that the specific window of i-th subject belongs to a certain state k or not; $state\_vector_{(i)}$ is the states of the i-th subject in all window. Moreover, the reoccurring patterns and time occupied divergence of Emoid, Nback, and Rest fMRIs for children and young adults are shown in Figures 2-4 of the Supplementary material. We visualize the top 100 significant functional connectivity related to age in each state (i.e., the functional connectivity corresponding to the 100 smallest FDR-corrected p-values of two-sample t-test performed across subject's mean dFC by state) under Emoid, Nback, and Rest, which are shown in Figures 5-7 of the Supplementary material.

To study the changes in reoccurring patterns over time under two tasks and resting state, we define the transition probabilities P_{ij} from time t to time $t+1$ as follows

$$P_{ij} = \frac{sum\{I_{(S_t^n=i, S_{t+1}^n=j)} == 1\}_{n=1}^N}{sum\{I_{(S_t^n=i)} == 1\}_{n=1}^N} \qquad i,j = 1,2,\cdots,s$$

where $S^n = (S_1^n, S_2^n, \cdots, S_T^n) \in \mathbb{R}^{1\times T}$ is the state vector for n-th subject, and $S_t^n = i$ for $i = 1,2,\cdots,s$ (s is 5, 5, and 4 for Emoid, Nback, and Rest respectively) represents that the n-th subject is in state i at time t. $I_{(\cdot)}$ is an indicative function, which is 1 when the condition is true, otherwise it is 0. The probability of each state at the initial time is defined as

$$P_i = \frac{sum\{I_{(S_1^n=i)} == 1\}_{n=1}^N}{N} \qquad i = 1,2,\cdots,s$$

Specifically, for a given state i_t at time t, we can calculate the transition probabilities $P_{i_t j}$ for $j = 1,2,\cdots,s$ from time t to time $t+1$. Then we record the maximum transition probability and the corresponding state at time $t+1$, and denote them as $P_{i_t i_{t+1}}$ and i_{t+1} respectively. By repeating the above steps, we can obtain the state transition curve with maximum state transition probability, which is shown in Figures 8-10 A of the Supplementary material. To further explore how the strength of functional connectivity changes over time, we count the proportion of enhancement and decrease of functional connections within or between functional networks during state transition, which is shown in Figures 8-10 B of the Supplementary material. To contrast the functional connectivity matrices between two adjacent states at the state transition point, we visualized the differences of the functional connectivity matrices between two adjacent states at the state transition point for each group under Emoid task and Nback task, which are shown in Figure 11 of the Supplementary material.

V. Discussion

1) The Common Developmental Differences of Three fMRI Paradigms: Figures 2-4 A of the supplementary material show the reoccurring patterns of three paradigms for both children and young adults. For the child group, we found that states 1, 2, and 3 in the resting state are similar to the Emoid states 2, 3, and 4 (Pearson correlation coefficient is 0.9687, 0.9631, and 0.9631 respectively) and the Nback states 5, 2, and 3 (Pearson correlation coefficient is 0.9744, 0.9856, and 0.9602 respectively). The analogous conclusions also can be found in the young adult group, where all reoccurring patterns in resting state are similar to Emoid states 2, 3, 4, and 5 (Pearson correlation coefficient is 0.9379, 0.9561, 0.9191, and 0.9588 respectively) and Nback states 3, 2, 5, and 4 (Pearson correlation coefficient is 0.9552, 0.9612, 0.9566, and 0.9696 respectively). It indicates that the reoccurring patterns of three paradigms are similar for a subject. The same conclusion also can be found in previous research [39], which reveals that no matter in resting state or task, the basic structure of the brain functional network remains relatively consistent. The finding testifies that the brain has a shared functional architecture during resting and many directed tasks, and the shared functional architecture of the brain can only modulate the connectivity pattern in response to task demands. In other words, the overlapping functional connectivity patterns between Rest fMRI and two task fMRIs suggest a shared functional architecture underlying and even shaping brain function, and a potential explanation of overlap is that the functional connectivity during resting constrains the activation of brain regions in response to task demands [40].

Although the brain shares the basic functional architecture during task and resting state, the basic functional organization between children and young adults are different. The number of within or between functional networks that children exist high-strength functional connections is 43 in state 1 and 2 in state 3 under Emoid task, 55 in state 1 and 10 in state 2 under Nback task, and 13 in state 2 under resting state. The number of within or between functional networks that young adults exist high-strength functional connections is 9 in state 1 under Emoid task, 20 in state 1 under Nback task, and 2 in state 2 under resting state. For all three fMRI paradigms, we found that children have many high-strength functional connections distributed widely among 13 functional networks, young adults have high-strength functional connections only within and between some functional networks. It is consistent with the previous studies that children show more diffuse functional connectivity patterns while young adults show more focused functional connectivity patterns, and the changes in functional connectivity patterns between children and young adults explain how brain function changes from an undifferentiated system to a specialized system as one grows up [3], [4]. The brain organization of distinct and stronger within-network communication can promote precise modulation efficiently because it can transfer more information in a short time [41]. Thus, compared with children with more diffuse functional connectivity patterns, the brain organization of young adults with more focused functional connectivity patterns can transmit information more efficiently and facilitate precise modulation during resting and two tasks.

Additionally, the functional connectivity among DMN, SCN, MRN, CN, AN, FPTCN, and SN is decreased in most reoccurring patterns for Emoid, Nback, and Rest fMRIs during development. DMN, so-called task-negative network, is broadly inactivated across tasks, which are closely related to numerous key brain functions such as integration of autobiographic information, self-monitoring, and social cognition [28]. It is reported that the functional activity in DMN never stops but regulates during the resting state [42]. SCN participates in memory, attention, perception, and consciousness, and dominates the motivation and emotion state independent control of cortical functions [43]. MRN is reported to be engaged during autobiographical memory retrieval that involves strategic search processes guided by self-knowledge and current goals, memory recovery associated with a rich sense of re-experience, monitoring, and other control processes [44]. CN is not just considered as the domain of motor control that receives information from widespread regions to affect the generation and control of movement, but also is thought to be involved in cognition and visuospatial reasoning [45]. AN innervated by autonomic nerves, involves activities related to sound information including collection, conduction, and processing [46], [47]. FPTCN involving working memory maintenance, predictive perceptual coding, and cognitive task, is thought to play an important part in mediating the allocation of attentional resources to compete for auditory information under varying degrees of perceptual demand [48]. SN is thought to regulate attention and behavior adaptively through the physical characteristics and the relevant information of the task, and also is considered to be a key interface for cognitive, homeostatic, motivational, and affective systems [49]. Both resting and task fMRIs suggest that the functions of the brain in processing information, working memory, and cognition are not mature in children compared with young adults [3].

The functional connectivity between SSN, COTCN, DAN, and some other functional networks is increased in some reoccurring patterns for resting and task fMRIs during development. SSN participates in the process of emotional feeling and cognitive activities [50]. COTCN is the key to coordinate information transmission and involves many complex cognitive tasks [51]. DAN controls external and attention-demanding cognitive functions [52]. Three fMRI paradigms indicate that brain functions related to emotional feelings, cognition, and information transmission are still growing with age.

2) The Developmental Differences of Each fMRI Paradigm: Figures 2-4 B of the supplementary material show the time occupied divergence of children and young adults during task and resting state. Both children and young adults have lower DT and FT in each state for two tasks while having higher DT and FT in each state during rest. It indicates that subjects including children and young adults tend to switch frequently among states in tasks and prefer to stay in a particular state while resting. It reveals that the spontaneous functional activity is stable during resting state, and then the functional activity corresponding to task demands changes quickly when the participant is required to perform a task [53].

For Emoid fMRI, both children and young adults stay in states 2, 3, and 4 for about the same time, but children stay longer in state 1 while young adults stay longer in state 5. Under the Emoid task, whether the initial state is 2, 3, 4, or 5, the children group will eventually switch to state 4 at time 9, and then they will switch back and forth between state 2 and state 4. When the initial state is 1, children group will stay in state 1 for the most time and then switch to state 3. No matter which the initial state is, the young adult group will eventually switch to state 5 at time 5 and stay at state 5 for a long while, and then they will switch to state 4 at time 18. The result of the Emoid task indicates that children have more frequent state transitions between state 2 and state 4, and the strength of functional connections within or between functional networks changes over time. Compared with children, the strength of functional connectivity within or between functional networks decreases at the early stage for young adults, and then they prefer to stay in state 5.

For Nback fMRI, both children and young adults stay in states 2, 3, and 4 for about the same time, but children stay longer in state 1 while young adults stay longer in state 5. Under the Nback task, no matter which the initial state is, the children group will eventually switch to state 4 at time 9, and then they will stay at state 4 until they switch to state 3 at time 18. The young adult group switch between state 4 and state 5 after time 7 in any initial state. The result of the Nback task indicates that the strength of functional connectivity for children changes over time during the frequent state transition at an early time, and then they will stay at state 4 for a while and finally switch to state 3. Unlike children, young adults prefer to stay for a while after switching to state 4 or state 5, and the strength of functional connectivity within or between functional networks decreases first, then increases, and then decreases during state transition between state 4 and state 5.

For Rest fMRI, both children and young adults stay in state 3 for about the same time. Children stay longer in state 1 and state 2, whereas young adults prefer to stay in state 4. Under the resting state, both children and young adults prefer to stay in a specified state with no change in the strength of functional connectivity within or between functional networks. We found that children prefer to switch among states with diffuse functional connectivity patterns during the two tasks and stay in states with diffuse functional connectivity patterns during rest. On the other hand, young adults switch among states with focused functional connectivity patterns in two task fMRIs and stay in states with focused functional connectivity patterns during rest.

For Emoid fMRI, along with the enhanced functional connectivity among SSN, COTCN, and DAN with age in states 4 and 5, the functional connectivity in the rest states declines to various degrees. For Nback fMRI, there exists enhanced functional connections within and between 13 functional networks in state 3 during development. Also, the functional connectivity decreases in the rest states with age. For Rest fMRI, in states 1, 2, and 4, there are not only lessened functional connections which mainly exist among SCN, MRN, CN, DMN, AN, FPTCN, and SN, but also exist strengthen functional connections which are mainly among SSN, COTCN, and DAN. In state 3 of Rest fMRI, the functional connections within and between 13 functional networks enhance during development. We found that compared with children, the functional connectivity of young adults increases or reduces with time for resting fMRI while generally decreasing for the two tasks. It indicates that the changes of functional connectivity with age are more complex in resting, and the brain functions related to emotion and working memory are more mature and efficient during development [4], [41].

VI. Conclusion

In this paper, we present an explainable multimodal deep dictionary learning method to capture the developmental differences between children and young adults from three fMRI paradigms. Specifically, the shared dictionary and the modality-specific sparse representations are learned based on the multimodal data and their encodings of the sDAE to simultaneously reveal the commonality and specificity of different paradigms. By applying the proposed method to the three fMRI paradigms from PNC, we found that children share a diffuse functional connectivity pattern while young adults share a focused functional connectivity pattern during both resting and two tasks. Three fMRI paradigms reveal that compared with children, young adults possess more mature and efficient functional networks for processing information. Children and young adults rarely transit from one state to other states during resting and prefer to switch among states over time during a task.

References

[1] M. Edde *et al.*, “Functional brain connectivity changes across the human life span: From fetal development to old age,” *Journal of Neuroscience Research*, vol. 99, no. 1, pp. 236–262, 2021.
[2] W. Hu *et al.*, “Deep collaborative learning with application to the study of multimodal brain development,” *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 12, pp. 3346–3359, 2019.
[3] D. D. Jolles *et al.*, “A comprehensive study of whole-brain functional connectivity in children and young adults,” *Cerebral Cortex*, vol. 21, no. 2, pp. 385–391, 06 2010.
[4] C. Qiao *et al.*, “Sparse deep dictionary learning identifies differences of time-varying functional connectivity in brain neuro-developmental study,” *Neural Networks*, vol. 135, pp. 91–104, 2021.
[5] B. Cai *et al.*, “Capturing dynamic connectivity from resting state fmri using time-varying graphical lasso,” *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 7, pp. 1852–1862, 2019.
[6] Y. Sun *et al.*, “Brain state-dependent dynamic functional connectivity patterns in attention-deficit/hyperactivity disorder,” *Journal of Psychiatric Research*, vol. 138, pp. 569–575, 2021.
[7] B. Cai *et al.*, “Estimation of dynamic sparse connectivity patterns from resting state fmri,” *IEEE Transactions on Medical Imaging*, vol. 37, no. 5, pp. 1224–1234, 2018.
[8] A. S. Choe *et al.*, “Comparing test-retest reliability of dynamic functional connectivity methods,” *NeuroImage*, vol. 158, pp. 155–175, 2017.
[9] M. Rahim *et al.*, “Integrating multimodal priors in predictive models for the functional characterization of alzheimer’s disease,” in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. Frangi, Eds. Cham: Springer International Publishing, 2015, pp. 207–214.
[10] Q. Li *et al.*, “Classification of alzheimer’s disease, mild cognitive impairment, and cognitively unimpaired individuals using multi-feature kernel discriminant dictionary learning,” *Frontiers in Computational Neuroscience*, vol. 11, pp. 1–14, 2018.
[11] Q. Li *et al.*, “Multi-modal discriminative dictionary learning for alzheimer’s disease and mild cognitive impairment,” *Computer Methods and Programs in Biomedicine*, vol. 150, pp. 1–8, 2017.
[12] L. Xiao *et al.*, “A manifold regularized multi-task learning model for iq prediction from two fmri paradigms,” *IEEE Transactions on Biomedical Engineering*, vol. 67, no. 3, pp. 796–806, 2020.
[13] F. Liu *et al.*, “Inter-modality relationship constrained multi-modality multi-task feature selection for alzheimer’s disease and mild cognitive impairment identification,” *NeuroImage*, vol. 84, pp. 466–475, 2014.
[14] D. Hu *et al.*, “Disentangled-multimodal adversarial autoencoder: Application to infant age prediction with incomplete multimodal neuroimages,” *IEEE Transactions on Medical Imaging*, vol. 39, no. 12, pp. 4137–4149, 2020.
[15] R. Abreu *et al.*, “Identification of epileptic brain states by dynamic functional connectivity analysis of simultaneous EEG-fMRI: a dictionary learning approach,” *Scientific Reports*, vol. 9, no. 1, p. 638, 2019.
[16] H. Li *et al.*, “Multi-modality low-rank learning fused first-order and second-order information for computer-aided diagnosis of schizophrenia,” in *Intelligence Science and Big Data Engineering. Big Data and Machine Learning*, Z. Cui, J. Pan, S. Zhang, L. Xiao, and J. Yang, Eds. Cham: Springer International Publishing, 2019, pp. 356–368.
[17] N. D’Souza *et al.*, “Deep sr-ddl: Deep structurally regularized dynamic dictionary learning to integrate multimodal and dynamic functional connectomics data for multidimensional clinical characterizations,” *NeuroImage*, vol. 241, p. 118388, 2021.
[18] W. Shao *et al.*, “Hypergraph based multi-task feature selection for multimodal classification of alzheimer’s disease,” *Computerized Medical Imaging and Graphics*, vol. 80, p. 101663, 2020.
[19] D. Di *et al.*, “Hypergraph learning for identification of covid-19 with ct imaging,” *Medical Image Analysis*, vol. 68, p. 101910, 2021.
[20] C. Wang *et al.*, “High-level attributes modeling for indoor scenes classification,” *Neurocomputing*, vol. 121, pp. 337–343, 2013, advances in Artificial Neural Networks and Machine Learning.
[21] B. Schlkopf *et al.*, *Learning with Hypergraphs: Clustering, Classification, and Embedding*. MIT Press, 2007, pp. 1601–1608.
[22] R. Jin *et al.*, “Dictionary learning-based fmri data analysis for capturing common and individual neural activation maps,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 6, pp. 1265–1279, 2020.
[23] Z. Zhang *et al.*, “Trace ratio optimization-based semi-supervised nonlinear dimensionality reduction for marginal manifold visualization,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 5, pp. 1148–1161, 2013.
[24] M. Yang *et al.*, “Fisher discrimination dictionary learning for sparse representation,” in *2011 International Conference on Computer Vision*, 2011, pp. 543–550.
[25] Y. Zhou and K. E. Barner, “Locality constrained dictionary learning for nonlinear dimensionality reduction,” *IEEE Signal Processing Letters*, vol. 20, no. 4, pp. 335–338, 2013.
[26] A. Beck and M. Teboulle, “A fast iterative shrinkage-thresholding algorithm for linear inverse problems,” *SIAM J. Img. Sci.*, vol. 2, no. 1, pp. 183–202, Mar. 2009.
[27] T. D. Satterthwaite *et al.*, “Neuroimaging of the philadelphia neurodevelopmental cohort,” *NeuroImage*, vol. 86, pp. 544–553, 2014.
[28] J. D. Power *et al.*, “Functional network organization of the human brain,” *Neuron*, vol. 72, no. 4, pp. 665–678, 2011.
[29] S. Shakil *et al.*, “Evaluation of sliding window correlation performance for characterizing dynamic functional connectivity and brain states,” *NeuroImage*, vol. 133, pp. 111–128, 2016.
[30] Ü. Sakolu *et al.*, “A method for evaluating dynamic functional network connectivity and task-modulation: application to schizophrenia,” *Magnetic Resonance Materials in Physics, Biology and Medicine*, vol. 23, no. 5, pp. 351–366, 2010.
[31] V. Calhoun *et al.*, “The chronnectome: Time-varying connectivity networks as the next frontier in fmri data discovery,” *Neuron*, vol. 84, no. 2, pp. 262–274, 2014.
[32] K. Engan *et al.*, “Family of iterative ls-based dictionary learning algorithms, ils-dla, for sparse signal representation,” *Digital Signal Processing*, vol. 17, no. 1, pp. 32–49, 2007.
[33] M. Abas *et al.*, “Agarwood oil quality classification using support vector classifier and grid search cross validation hyperparameter tuning,” *International Journal of Emerging Trends in Engineering Research*, vol. 8, no. 6, pp. 2551–2556, 2020.
[34] S. Saud *et al.*, “Performance improvement of empirical models for estimation of global solar radiation in india: A k-fold cross-validation approach,” *Sustainable Energy Technologies and Assessments*, vol. 40, p. 100768, 2020.
[35] D. Xu *et al.*, “Convergence of the rmsprop deep learning method with penalty for nonconvex optimization,” *Neural Networks*, vol. 139, pp. 17–23, 2021.
[36] S. Bahrampour *et al.*, “Multimodal task-driven dictionary learning for image classification,” *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 24–38, 2016.
[37] T. Kanungo *et al.*, “An efficient k-means clustering algorithm: analysis and implementation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 881–892, 2002.
[38] J. Cohen, “A coefficient of agreement for nominal scales,” *Educational & Psychological Measurement*, vol. 20, pp. 37–46, 1960.
[39] M. Cole *et al.*, “Intrinsic and task-evoked network architectures of the human brain,” *Neuron*, vol. 83, no. 1, pp. 238–251, 2014.
[40] C. Hughes *et al.*, “Aging relates to a disproportionately weaker functional architecture of brain networks during rest and task states,” *NeuroImage*, vol. 209, p. 116521, 2020.
[41] E. Bullmore and O. Sporns, “The economy of brain network organization,” *Nature Reviews Neuroscience*, vol. 13, no. 5, pp. 336–349, 2012.
[42] R. N. Spreng *et al.*, “The common neural basis of autobiographical memory, prospection, navigation, theory of mind, and the default mode: A quantitative meta-analysis,” *J Cogn Neurosci*, vol. 21, no. 3, pp. 489–510, 2009.
[43] J. Kang *et al.*, “Energy landscape analysis of the subcortical brain network unravels system properties beneath resting state dynamics,” *NeuroImage*, vol. 149, pp. 153–164, 2017.
[44] P. L. St. Jacques *et al.*, “Dynamic neural networks supporting memory retrieval,” *NeuroImage*, vol. 57, no. 2, pp. 608–616, 2011.
[45] A. C. Bostan *et al.*, “Cerebellar networks with the cerebral cortex and basal ganglia,” *Trends in Cognitive Sciences*, vol. 17, no. 5, pp. 241–254, 2013.

[46] S. M. Smith *et al.*, "Correspondence of the brain's functional architecture during activation and rest," *Proceedings of the National Academy of Sciences*, vol. 106, no. 31, pp. 13 040–13 045, 2009.
[47] A. M. Leaver *et al.*, "Dysregulation of limbic and auditory networks in tinnitus," *Neuron*, vol. 69, no. 1, pp. 33–43, 2011.
[48] S. Pillay *et al.*, "Perceptual demand and distraction interactions mediated by task-control networks," *NeuroImage*, vol. 138, pp. 141–146, 2016.
[49] V. Menon, "Salience network," in *Brain Mapping*, A. W. Toga, Ed. Waltham: Academic Press, 2015, pp. 597–611.
[50] A. Londei *et al.*, "Sensory-motor brain network connectivity for speech comprehension," *Human Brain Mapping*, vol. 31, no. 4, pp. 567–580, 2010.
[51] J. M. Sheffield *et al.*, "Fronto-parietal and cingulo-opercular network integrity and cognition in health and schizophrenia," *Neuropsychologia*, vol. 73, pp. 82–93, 2015.
[52] W. Gao *et al.*, "The synchronization within and interaction between the default and dorsal attention networks in early infancy," *Cerebral Cortex*, vol. 23, no. 3, pp. 594–603, 02 2012.
[53] R. Jiang *et al.*, "Task-induced brain connectivity promotes the detection of individual differences in brain-behavior relationships," *NeuroImage*, vol. 207, p. 116370, 2020.
[54] K. R. Merikoski Jorma K, "Inequalities for spreads of matrix sums and products." *Applied Mathematics E-Notes [electronic only]*, vol. 4, pp. 150–159, 2004.

Appendix I

A. The Comparison of Sparsity of Activations Among L$_1$-norm, L$_2$-norm, and Kullback-Leibler Divergence

Let ρ be a small positive constant between 0 and 1, and $\rho_j^{(l)} = \frac{1}{N}\sum_{n=1}^{N} h_{nj}^{(l)}$ with $h_{nj}^{(l)}$ being the j-th element of $h_n^{(l)}$ is the average activation of neural j in the l-th layer, then the Kullback-Leibler divergence is defined as

$$KL(\rho||\rho_j^{(l)}) = \rho \log \frac{\rho}{\rho_j^{(l)}} + (1-\rho)\log\frac{1-\rho}{1-\rho_j^{(l)}}$$

The penalty functions based on L_1-norm and L_2-norm are defined as

$$f_{L_1} = \sum_{l=1}^{2L-1}\sum_{j=1}^{r^{(l)}} \left|\rho_j^{(l)}\right|$$

$$f_{L_2} = \sum_{l=1}^{2L-1}\sum_{j=1}^{r^{(l)}} \left(\rho_j^{(l)}\right)^2$$

The sparsity and SNR with L_1-norm, L_2-norm and Kullback-Leibler divergence have been compared and the results are shown in Figure 12 of the Supplementary material. The results show that the sparsity of Kullback-Leibler divergence is better than that of L_1-norm in most hidden layers, and the sparsity of L_2-norm is the worst among the above three penalty functions. The SNR evaluation of EMDDL on the multimodal data in both the original space and the encoding space show that, EMDDL based on Kullback-Leibler divergence has better reconstruction ability compared with L_1-norm and L_2-norm.

B. The Proof of the Convexity of f(V)

The convexity of $f(V)$ depends on whether its Hessian matrix $\nabla^2 f(V)$ is positive definite or not. Thus, as long as $\nabla^2 f(V)$ is positive definite, the convexity of $f(V)$ can be guaranteed. The Hessian matrix $\nabla^2 f(V)$ is

$$\nabla^2 f(V) = \frac{1}{N}\left(D^{\mathrm{T}}D + {D^{(L)}}^{\mathrm{T}} D^{(L)}\right) + S$$

According to the Weyl's inequality [54], we have

$$\begin{aligned}
\lambda_{min}(\nabla^2 f(V)) =& \lambda_{min}\left(\frac{1}{N}(D^{\mathrm{T}}D + {D^{(L)}}^{\mathrm{T}} D^{(L)}) + S\right) \\
\geq& \lambda_{min}\left(\frac{1}{N}D^{\mathrm{T}}D\right) + \lambda_{min}\left(\frac{1}{N}{D^{(L)}}^{\mathrm{T}} D^{(L)}\right) \\
&+ \lambda_{min}(\lambda_3 H_2) + \lambda_{min}(-2\lambda_3 H_1) \\
&+ \lambda_{min}(\lambda_4 L) + \lambda_{min}\Big((\lambda_3+\lambda_5)I\Big) \\
=& (\lambda_3+\lambda_5) - 2\lambda_3 \\
=& \lambda_5 - \lambda_3
\end{aligned}$$

where $\lambda_{min}(\cdot)$ denotes the smallest eigenvalue of a matrix. To ensure the positive definite of the Hessian matrix $\nabla^2 f(V)$, $\lambda_{min}(\nabla^2 f(V))$ should be greater than 0. Thus, $f(V)$ is convex when $\lambda_3 < \lambda_5$ holds.

C. The Lipschitz Constant of the Gradient ∇V

For every $V^1, V^2 \in \mathbb{R}^{K\times N}$, we have

$$\begin{aligned}
||\nabla V^1 - \nabla V^2||_2 =& ||\frac{1}{N}\left(D^{\mathrm{T}}D + {D^{(L)}}^{\mathrm{T}} D^{(L)}\right)\left(V^1 - V^2\right) \\
&+ \left(V^1 - V^2\right) S||_2 \\
\leq& ||\frac{1}{N}\left(D^{\mathrm{T}}D + {D^{(L)}}^{\mathrm{T}} D^{(L)}\right)||_2 \\
& ||V^1 - V^2||_2 + ||V^1 - V^2||_2 ||S||_2 \\
\leq& \left(\frac{1}{N}(||D^{\mathrm{T}}D||_2 + ||{D^{(L)}}^{\mathrm{T}} D^{(L)}||_2) + ||S||_2\right) \\
& ||V^1 - V^2||_2 \\
=& \Bigg(\frac{\lambda_{max}(D^{\mathrm{T}}D) + \lambda_{max}({D^{(L)}}^{\mathrm{T}} D^{(L)})}{N} \\
&+ \sqrt{\lambda_{max}(S^{\mathrm{T}}S)}\Bigg)||V^1 - V^2||_2
\end{aligned}$$

Thus, the Lipschitz constant of the gradient ∇V is

$$\begin{aligned}
L_f =& \frac{1}{N}\left(\lambda_{max}(D^{\mathrm{T}}D) + \lambda_{max}({D^{(L)}}^{\mathrm{T}} D^{(L)})\right) \\
&+ \sqrt{\lambda_{max}(S^{\mathrm{T}}S)}
\end{aligned} \tag{13}$$

where $\lambda_{max}(\cdot)$ denotes the largest eigenvalue of a matrix.

D. The Details of Obtaining Multimodal Data

There are 264 BOLD signals with T_m time points for the m-th modality of the n-th subject. $fc_{nk}^{(m)}(i,j)$, the functional connectivity between the i-th ROI and the j-th ROI within the k-th window for the m-th modality of the n-th subject, is calculated based on the Pearson correlation coefficient, which is defined as follows

$$fc_{nk}^{(m)}(i,j) = \frac{\sum_{t=1}^{w_l}(B_{nk}^{(m)}(i,t) - \bar{B}_{nk}^{(m)}(i))(B_{nk}^{(m)}(j,t) - \bar{B}_{nk}^{(m)}(j))}{\sqrt{\sum_{t=1}^{w_l}(B_{nk}^{(m)}(i,t) - \bar{B}_{nk}^{(m)}(i))^2}\sqrt{\sum_{t=1}^{w_l}(B_{nk}^{(m)}(j,t) - \bar{B}_{nk}^{(m)}(j))^2}}$$

where $B_{nk}^{(m)}(i,t)$ is the t-th BOLD signal value of the i-th ROI within the k-th window for the m-th modality of the n-th subject. $\bar{B}_{nk}^{(m)}(i) = \frac{1}{w_l}\sum_{t=1}^{w_l} B_{nk}^{(m)}(i,t)$ is the sample mean of the BOLD signals of the i-th ROI within the k-th window for the m-th modality of the n-th subject. By calculating the functional connectivity between any two ROIs within the k-th window for the m-th modality of the n-th subject, $C_{264}^{2}=34716$ functional connections can be obtained within the k-th window for the m-th modality of the n-th subject. For a BOLD signals with T time points, we can obtain $S_l=\frac{T-w_l}{s_l}+1$ windows with window length w_l and scan length s_l. Thus, a dynamic functional connection matrix $fc_n^{(m)}\in\mathbb{R}^{34716\times S_l}$ can be obtained for the m-th modality of the n-th subject. Let $X_{(m)}=(fc_1^{(m)},fc_2^{(m)},\cdots,fc_{N_s}^{(m)})\in\mathbb{R}^{p\times N_m}$ be the data of the m-th modality, and multimodal data $X=(X_{(1)},X_{(2)},\cdots,X_{(M)})\in\mathbb{R}^{p\times N}$ is composed of M modalities. In which, $p=C_{264}^{2}=34716$, $N_m=S_l\times N_s$ with N_s being the number of subjects, and $N=\sum_{m=1}^{M}N_m$. The flowchart of calculating $fc_n^{(m)}$ is shown in Figure 13 of the Supplementary material.

E. The Consistency of Clustering Results

Firstly, we can obtain two subgroups $X^{(1)}$ and $X^{(2)}$ by sampling 80% of items from the sparse representations without replacement. Then, we can obtain two clustering results $M^{(1)}$ and $M^{(2)}$ based on the two subgroups using the k-means clustering method with the cityblock distance metric, where $M(i,j)=1$ if item i and item j belong to the same cluster, otherwise it is 0. If both item i and item j are present in the subgroups $X^{(1)}$ and $X^{(2)}$, the corresponding element in $M^{(1)}$ and $M^{(2)}$ is retained. By vectorizing the upper triangle of M, we can obtain two vectors $h^{(1)}$ and $h^{(2)}$ which are used to obtain the confusion matrix. And the confusion matrix is shown in Table II.

TABLE II: The confusion matrix of clustering results based on $h^{(1)}$ and $h^{(2)}$.

	Items i and j belong to the same cluster based on $h^{(1)}$	Items i and j belong to the different clusters based on $h^{(1)}$
Items i and j belong to the same cluster based on $h^{(2)}$	N_1	N_2
Items i and j belong to the different clusters based on $h^{(2)}$	N_3	N_4

Then, the kappa coefficient is defined as

$$kappa=\frac{p_o-p_e}{1-p_e}$$

where $p_o=\frac{N_1+N_4}{N}$ is the proportion of units that the judges agreed and $p_e=\frac{(N_1+N_3)(N_1+N_2)+(N_2+N_4)(N_3+N_4)}{N^2}$ is the proportion of units for which agreement is expected by chance, and $N=N_1+N_2+N_3+N_4$. In which N_1 and N_4 represent the number of consistent clustering results based on two different subgroups, and N_2 and N_3 represent the number of inconsistent clustering results based on two different subgroups. To test the significance of the kappa coefficient (i.e., the null hypothesis $H_0: kappa=0$ and the alternative hypothesis $H_1: kappa>0$), the significance p-value can be performed by evaluating the normal curve deviate which is defined as

$$z=\frac{kappa}{\sqrt{\frac{p_o(1-p_o)}{N(1-p_e)^2}}}\tag{14}$$

The permutation test is also used to test the significance of the kappa coefficient. Specifically, for the giving clustering results $h^{(1)}$ and $h^{(2)}$, the corresponding statistic z can be obtained based on (14), and we denote it as z_0. Then, we randomly change the clustering results of one subgroup and obtain the corresponding statistic z based on (14), and we denote it as z_{perm}^{i}. Namely, we generate a random integer $N_i\leq N$ and also generate N_i different integers $\{I_1,I_2,\cdots,I_{N_i}\}$ which are less than or equal to N. And then the elements corresponding to the index $\{I_1,I_2,\cdots,I_{N_i}\}$ in $h^{(1)}$ are reversed (i.e., the reversed value is 1 if the original element is 0, and the reversed value is 0 if the original element is 1), and we denote it as $\widetilde{h}^{(1)}$. The corresponding statistic z can be obtained based on $\widetilde{h}^{(1)}$ and $h^{(2)}$ according to (14), and we denote it as z_{perm}^{i}. Finally, $z_{perm}=\{z_{perm}^{1},z_{perm}^{2},\cdots,z_{perm}^{N_{perm}}\}$ can be obtained by repeating the process N_{perm} times. And $z_{perm}=\{z_{perm}^{1},z_{perm}^{2},\cdots,z_{perm}^{N_{perm}}\}$ is used to estimate the distribution of statistic z, and then the p-value can be calculated by

$$p=\frac{count(z_{perm}\geq z_0)}{N_{perm}+1}$$

where N_{perm} is 1000 by considering the tradeoff between the time complexity and the estimation accuracy of distribution of statistic.

To ensure the reliability of the results, 1000 repeated experiments are implemented for each group of each modality. For each group of each modality, 1000 $kappa$ values, and 1000 corresponding values of statistic z, and 1000 corresponding p-values based on the normal curve deviate, and 1000 corresponding p-values based on the permutation test can be obtained. The results are shown in Figure 14 of the Supplementary material. For each group of each modality, all the obtained 1000 p-values based on the normal curve deviate and all the obtained 1000 p-values based on the permutation test are nearly 0, which are much less than 0.05, indicating that all the observed agreement is not accidental in 1000 repeated experiments. Most $kappa$ values are larger than 0.6, which indicates that the clustering results obtained from two different subgroups are substantial agreement or perfect agreement in a large probability. Besides, it shows that the clustering analysis results from one subgroup are basically consistent with the result from another subgroup.