JGAT: a joint spatio-temporal graph attention model for brain decoding

Han Yi Chiu
Georgia Institute of Technology
Atlanta, GA 30332
r28875148@gmail.com

Liang Zhao
Emory University
Atlanta, GA 30322
liang.zhao@emory.edu

Anqi Wu
Georgia Institute of Technology
Atlanta, GA 30332
anqiwu@gatech.edu

Abstract

The decoding of brain neural networks has been an intriguing topic in neuroscience for a well-rounded understanding of different types of brain disorders and cognitive stimuli. Integrating different types of connectivity, e.g., Functional Connectivity (FC) and Structural Connectivity (SC), from multi-modal imaging techniques can take their complementary information into account and therefore have the potential to get better decoding capability. However, traditional approaches for integrating FC and SC overlook the dynamical variations, which stand a great chance to over-generalize the brain neural network. In this paper, we propose a Joint kernel Graph Attention Network (JGAT), which is a new multi-modal temporal graph attention network framework. It integrates the data from functional Magnetic Resonance Images (fMRI) and Diffusion Weighted Imaging (DWI) while preserving the dynamic information at the same time. We conduct brain-decoding tasks with our JGAT on four independent datasets: three of 7T fMRI datasets from the Human Connectome Project (HCP) and one from animal neural recordings. Furthermore, with Attention Scores (AS) and Frame Scores (FS) computed and learned from the model, we can locate several informative temporal segments and build meaningful dynamical pathways along the temporal domain for the HCP datasets. The URL to the code of JGAT model: https://github.com/BRAINML-GT/JGAT.

1 Introduction

Analysis of brain neural networks measured by neuroimaging techniques has helped in revealing potential structures and functions of human brains [1]. Specifically, these structures and functions of human brains can provide informative representations and patterns which play a pivotal role in identifying multiple brain neural diseases [2–8] and understanding human behavior and cognition. The functional and structural information measured by noninvasive neuroimaging technologies is often referred to as Functional Connectivity (FC) and Structural Connectivity (SC). Researchers have developed a variety of Machine Learning (ML) models to integrate FC and SC for decoding the brain networks of brain disorders and cognitive stimuli. Graph Neural Networks (GNNs) are the mainstream ML models commonly used to integrate FC and SC. There are existing graph-based models that have been built and dealt with many neurological problems successfully [9–15]. Nevertheless, a common issue of these works in neuroimaging is that they only integrate FC and SC for brain decoding and fail to capture dynamic changes and patterns.

In order to tackle the time information in neuroimaging data, recent literature implements several cutting-edge models for the purpose of both modeling dynamical systems and preserving the spatial relationship [16–18]. However, these approaches capture temporal embedding and spatial embedding independently, which leads to a limitation in modeling the conditions where a current brain region influences other brain regions in a few previous time steps, i.e., cross-region dynamics. Furthermore, they achieve the spatial-temporal embedding in a sequential way, e.g., deriving the spatial embedding

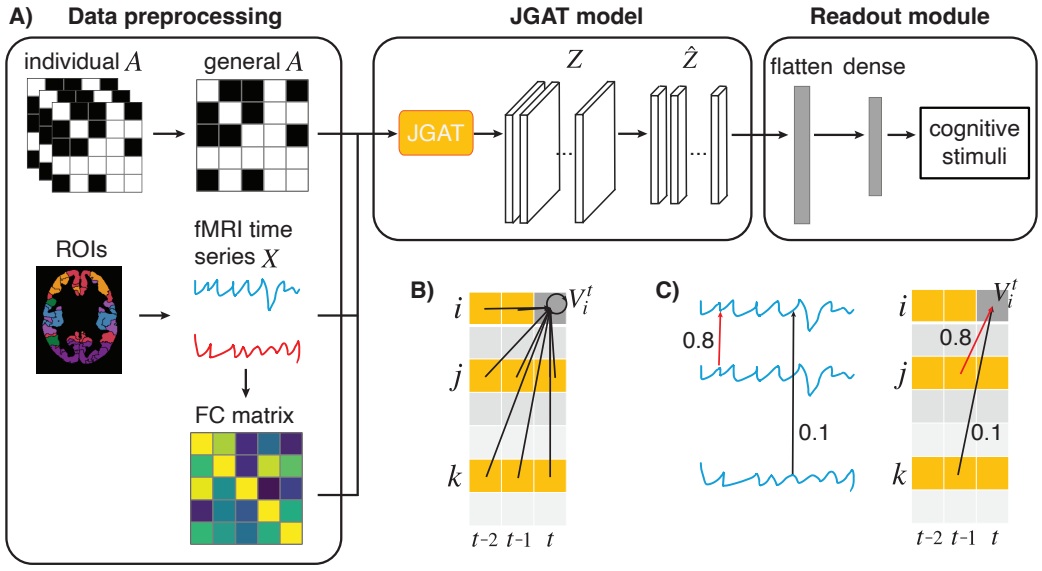


Figure 1: A) Overview of the JGAT architecture. B) Schematic of the joint graph. C) Effect of FC.

from the input data and then deriving the temporal embedding from the spatial one. This implies that only one of the embeddings relies on the original data inputs, potentially limiting the integration of both spatial and temporal information. We are interested in aggregating input features into both spatial and temporal embeddings jointly. The joint embedding helps build connections between brain regions from different locations and time frames.

For modeling the aforementioned conditions, we propose a Joint kernel Graph Attention Network (JGAT) to combine FC and SC for the integration of multi-modal imaging techniques. Besides integrating signals of the statistical dependency across brain regions and their physical locations, we also consider dynamical variations across several time frames. Moreover, by considering temporal and spatial domains at the same time, our approach allows to describe the dynamic cognitive pathways associated with various types of cognitive stimuli. In the evaluation section, we compare our model with both ML architectures and some advanced dynamical graph models. We will show that our JGAT model can provide meaningful interpretations for dynamical connectivity and construct pathways of informative segments across the time series of cognitive stimuli for several neuroimaging datasets.

2 Methodology

2.1 Architecture overview

Our goal is to perform classification tasks on several cognitive datasets. Therefore, we build a brain decoding architecture (Fig. 1A) taking neuroimaging data as the input and predicting labels corresponding to the cognitive stimuli. The inputs of the architecture consist of three streams that result from two resources. Diffusion Weighted Imaging (DWI) data is the first resource. We can obtain DWI data from individual subjects, each providing an SC or adjacency matrix $\tilde{A} \in \mathbb{R}^{N\times N}$. We then calculate a general adjacency $A \in \mathbb{R}^{N\times N}$ from $\tilde{A}$ as the first stream. The second resource is the functional Magnetic Resonance Images (fMRI) sequence $X \in \mathbb{R}^{N\times T}$, measured from N Region of Interests (ROIs) at T time frames. It contributes to the other two streams. The second stream is the fMRI sequence. The third stream is the FC matrix which is N by N, calculated from the fMRI data by the Pearson correlation coefficient between each pair of brain regions [19–22].

Afterward, three streams of input data are fed into the proposed JGAT model. The JGAT model consists of a novel JGAT layer and a dense neural net layer. The JGAT layer aggregates input features into joint spatial-temporal embedding, denoted as Z, via a graph attention mechanism. A dense layer is then applied to Z to reduce the computational complexity and form the downstream embedding $\hat{Z}$. Finally, the readout module for graph classification contains one flatten layer followed by two dense

layers, with the $Softmax$ function used in the last dense layer. We also apply two dropout layers in the readout module to prevent overfitting. We demonstrate the overview of JGAT architecture in Fig. 1A.

2.2 Joint Kernel Graph Attention Network

The core contribution of this paper is the development of the Joint kernel Graph Attention Network (JGAT) that can simultaneously incorporate both temporal and spatial information. For one fMRI sequence $X \in \mathbb{R}^{N \times T}$, there are N brain regions or ROIs with T time frames. To achieve a joint spatial-temporal graph, we can treat each entry of the data sequence as a node, resulting in a giant graph with NT nodes and a NT by NT adjacency matrix. However, such a method has too large a model structure and may not be necessary for time frames where there is no long-term effect, which is often the case in trial-based cognitive tasks. Consequently, instead of applying a giant matrix, we break down a giant graph into T smaller graphs, defined as joint kernels. Each kernel contains information in N brain regions across K time frames. For a better presentation of the model, we first denote a node in brain region i at time frame t as V_i^t. We then define a joint kernel graph of node V_i^t as G_i^t. The nodes in G_i^t include all brain regions that have anatomical connections to region i in a K-frame window (i.e., $\{t-K+1, t-K+2, ..., t-1, t\}$). $X_i^t \in \mathbb{R}^{1\times 1}$ is a scalar feature associated with V_i^t, representing the neural activity for region i at time t. The anatomical connections are indicated by the adjacency matrix A defined by DWI. Fig. 1B shows the schematic of a joint kernel G_i^t with $K = 3$. The yellow squares are neighbors of V_i^t in the joint kernel graph G_i^t with anatomical connections. All the yellow nodes within the K-frame time window communicate with V_i^t. We also assume self-communication, resulting in a self-loop edge. Therefore, for node V_i^t, not only can it receive messages from the neighborhood (e.g., j and k in Fig. 1B) in the current time, but it can also receive messages from its neighborhood from the previous time frames.

After defining the joint graph G_i^t, we now introduce the attention mechanism in the JGAT layer. The input feature for node V_i^t is X_i^t from the raw fMRI data. We aggregate information from its neighbors in G_i^t to get a new message $\tilde{X_i^t}$, computed as $\tilde{X_i^t} = \sum_{p \in G_i^t} \alpha_{ip}^t X_p W_a$, where α_{ip}^t refers to the Attention Scores (AS). It can be understood as the weight for the connection between the node V_i^t and its neighbor V_p in its joint domain graph G_i^t. Note that p indexes a node in G_i^t. But it actually implies both the index of the brain region and the time frame in the window. Equivalently, we can consider the vectorization of nodes in the matrix form of G_i^t (yellow squares in Fig. 1B). The new index in the vector form is p, which corresponds to a certain region index and time index in the matrix form. $W_a \in \mathbb{R}^{1\times d}$ denotes the learnable weight matrix of linear transformation for input feature X_p. We set $d = 10$ for all the experiments in this paper. The AS is defined as: $\alpha_{ip}^t = Softmax(c_{ip}^t)$, where c_{ip}^t is the AS before the activation function. α_{ip}^t can be expanded with the following attention mechanism:

$$\alpha_{ip}^t = \frac{exp(f((X_i^t W_a \parallel FC_{ip} \cdot X_p W_a) \cdot W_c))}{\sum_{p \in G_i^t} exp(f((X_i^t W_a \parallel FC_{ip} \cdot X_p W_a) \cdot W_c))}, \tag{1}$$

where $||$ denotes vector concatenation and $W_c \in \mathbb{R}^{2d\times 1}$ denotes a query in the attention mechanism. A key for this attention mechanism is the concatenation of X_i^t and X_p after a linear transformation by W_c. The query makes the key concentrates on one value that can reflect the importance of V_p with respect to node V_i^t. FC_{ip} in Eq. 1 is a Pearson correlation coefficient value that reflects the relationship between the whole time series of the region i and the region corresponding to index p. Fig. 1C displays the effects of FC. For node V_i^t in Fig. 1C, it should have stronger connectivity with V_j^{t-1} than with V_k^{t-1} since the higher functional similarity defined by the Pearson correlation coefficient. By applying FC_{ip} to regulate the importance of the neighborhood, we can expect the model to capture a more accurate embedding for node V_i^t by considering the similarity among its neighbors. f is the activation function which is $LeakyReLU$ in our implementation. The final output embedding Z_i^t for a single node V_i^t is the concatenation of node representations and its aggregated message multiplied by a corresponding Frame Scores (FS) β^t for frame t, defined as $Z_i^t = [X_i^t W_a \parallel \beta^t \cdot \tilde{X_i^t}]$, where β^t is computed from c_{ij}^t with normalization operations:

$$C_i^t = \frac{1}{|G_i^t|} \sum_{p \in G_i^t} c_{ip}^t, \quad \hat{\beta}^t = \frac{1}{N} \sum_{i=1}^{N} C_i^t, \quad \beta^t = Sigmoid(\frac{\hat{\beta}^t - \mu(\hat{\beta}^t)}{\sigma(\hat{\beta}^t)}). \tag{2}$$

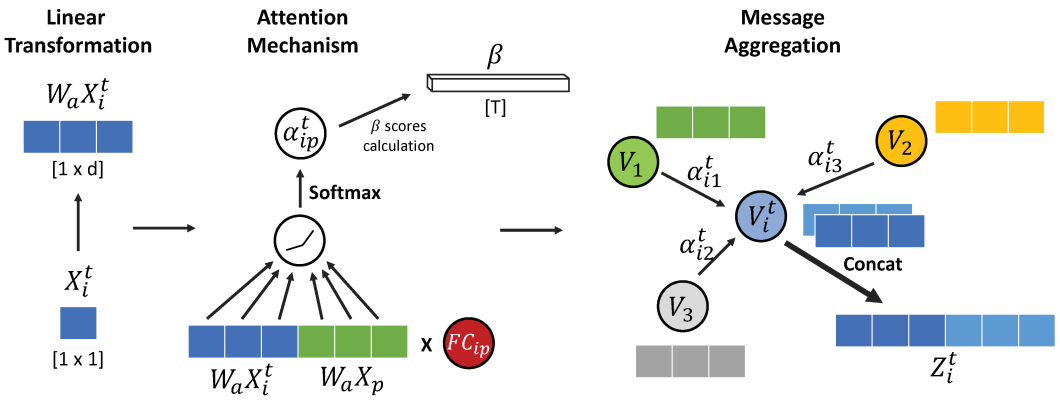


Figure 2: Overview of the JGAT layer.

C_i^t is the mean of all neighbors of node V_i^t, $|G_i^t|$ denotes the number of neighborhood of node V_i^t, and $\hat{\beta}^t$ is the mean of C_i^t over all brain regions in the same time frame. μ and σ in Eq. 2 denotes the mean and standard deviation of all FS. After normalization and the $Sigmoid$ function, we get a vector $\beta = [\beta^1, \beta^2, ..., \beta^T]^\top \in \mathbb{R}^{T\times 1}$ that reflects the importance of each frame with values ranging from 0 to 1. Fig. 2 demonstrates the overview of attention mechanism in the JGAT layer.

Finally, as mentioned above, we apply another dense layer to Z to get $\tilde{Z}$ which is further fed to the readout module to build the classification loss, which is the regular categorical cross-entropy loss.

3 Experiments and results

Datasets: In this work, we use four neuroscience datasets: three of them from WU-Minn Human Connectome Project (HCP) 1200 Subjects Data Release [23] which contains preprocessed 7T fMRI data and the fourth one is from animal neural recordings [24]. For HCP datasets, We select Working Memory (WM), Social, and Emotion to do binary classification tasks. Experimental designs of each dataset typically include an experimental group and a control group to accurately measure the brain activities and regions of interest. The experimental designs usually repeat with their interested stimuli and intertwined with the control group. Hence, by following HCP protocol, we can slice several segments of both the experimental and control groups from the fMRI time series of a single participant. The reason for selecting these three datasets is that they have uniform task blocks for both the experimental and control groups. The presence of uniform task blocks eliminates the need to crop or zero-pad to match the shape of the different cognitive stimuli (classes), thereby avoiding the risk of losing information.

For animal neural recordings, we apply our model to perform an 8-class classification task. The experimental design involved training two rhesus macaques, Chewie and Mihi, to perform a reaching task towards one of eight different directions following a cue. Data was collected from the primary motor cortex of their brains. We only present the JGAT model on neural recordings of mihi in this paper. The animal neural recordings help validate the classification performance of our JGAT model.

Experimental setup: For each of the HCP datasets, the experiment includes the fMRI and DWI data from 48 participants (subjects). After the slicing process to separate the experimental and control groups, WM has two stimuli, 2-back WM task and 0-back WM task, with input shape of M = 768, N = 200, T = 35; Social has two stimuli, Mental and Random, with input shape of M = 384, N = 200, T = 32; Emotion has two stimuli, Fear and Neutral, with input shape of M = 384, N = 200, T = 25. M indicates the number of instances/trials. As shown in Fig. 1A, the first stream of the model input takes advantage of DWI data. To seize group-level features, we use a threshold to filter out some rare edges in $\tilde{A}$ and therefore preserve edges that most people have. Since each subject has one individual $\tilde{A}$, the threshold is firstly set to be 40 to represent about 80% of participants. For animal neural recordings, there are 8 directions as labels with total inputs shape of M = 215, N = 172, T = 30. Since neural recordings do not have data similar to DWI, we use fully connected A as the adjacency matrix.

Table 1: Size of joint kernel K.

		$K=1$	$K=3$	$K=5$
WM	Train	96.21 ± 0.53	99.02 ± 0.30	96.95 ± 0.64
	Test	85.66 ± 0.69	86.50 ± 0.76	$\mathbf{86.59 \pm 0.70}$
Social	Train	99.22 ± 0.16	99.81 ± 0.08	99.70 ± 0.07
	Test	91.93 ± 0.92	$\mathbf{94.27 \pm 0.65}$	93.06 ± 0.69
Emotion	Train	98.13 ± 0.29	98.93 ± 0.37	98.71 ± 0.34
	Test	87.76 ± 0.96	$\mathbf{90.15 \pm 0.68}$	88.11 ± 0.80

Table 2: Ablation study on FC and FS.

		Without FC	Without FS	Without FC/FS	Full model
WM	Train	96.50 ± 0.52	96.18 ± 0.54	95.35 ± 0.82	99.02 ± 0.30
	Test	86.11 ± 0.71	86.33 ± 0.83	85.94 ± 0.85	$\mathbf{86.50 \pm 0.76}$
Social	Train	99.70 ± 0.10	99.75 ± 0.07	99.70 ± 0.09	99.81 ± 0.08
	Test	94.14 ± 0.68	93.84 ± 0.65	93.71 ± 0.64	$\mathbf{94.27 \pm 0.65}$
Emotion	Train	98.92 ± 0.30	99.09 ± 0.17	98.26 ± 0.62	98.93 ± 0.37
	Test	88.32 ± 0.84	88.15 ± 0.74	88.15 ± 0.98	$\mathbf{90.15 \pm 0.68}$

We train and test the algorithm on Tensorflow in the Python environment using Google Colab. Furthermore, considering the relatively small datasets, the whole experiments on HCP datasets use 8-Fold cross-validation, and the animal dataset uses 5-Fold cross-validation. In addition, the accuracy shown in the following tables takes an average of 48 samples for the HCP datasets and 30 samples for the animal dataset with standard error of mean (SEM). The HCP datasets are also arranged in the order of participants to prevent data leaking. The basic architecture of the JGAT model is implemented with one-head attention and one JGAT layer with the sliding window $K = 3$ and an initial learning rate of 0.001. We also provide results of multi-head attention with two JGAT layers.

Hyperparameter evaluation and ablation study: The first evaluation is the size of the sliding window K. Tab. 1 shows the accuracy of both the training set and testing set under different model settings of $K = 1, 3, and 5$. In general, all HCP datasets show better results when using $K = 3$ compared to $K = 1$. This observation supports the validity of the strategy of considering several previous time frames as neighbors. Moreover, in the Social and Emotion dataset, using $K = 3$ preserves more temporal information compared to $K = 5$. In the WM dataset, it exhibits a longer temporal dependency, as evidenced by the similar and decent performance when using both $K = 3$ and $K = 5$.

Secondly, aside from the regular attention mechanism, we also use FC to regulate features from the neighborhood and FS to indicate the importance of each time frame. The impact of FC and FS is presented in Tab. 2 for all HCP datasets with the accuracy performance of both the training set and testing set.

Thirdly, in the preliminary experimental setting, the threshold for filtering the general adjacency matrix is set to 40, resulting in a total number of approximately 5k edges in the joint kernel with $K = 3$. In the following evaluation (Tab. 3), we demonstrate the differences in accuracy with $K = 3$ under roughly 13k edges, 5k edges, 1.5k edges, and 0 edges by changing the threshold. Threshold = 24 produces roughly 13k edges, threshold = 40 produces roughly 5k edges, threshold = 48 produces roughly 1.5k edges, and Multilayer Perceptron (MLP) model represents the model setting of 0 edges. From Tab. 3, we can observe that both 1.5k and 5k settings perform well, but there is no significant difference between 1.5k and 5k edges. Considering that the main objective of this work is to uncover the dynamic connectivity between brain regions and provide interpretations from an edge perspective, we use 5k edges in the final model setting in order to preserve edge information as much as possible.

Comparison with baseline and advanced models: We compare our method with some basic ML models including MLP, Convolutional Neural Networks (CNN), Graph Attention Network (GAT) [25], and some advanced methods such as GCN-LSTM, Spatial Temporal Graph Convolutional Networks (ST-GCN) [18], SpatioTemporal Neural Data Transformer (STNDT) [17], Attention-Diffusion-Bilinear Neural Network (ADB-NN) [15] to evaluate the accuracy of classification results. The comparisons of the testing set are presented in Tab. 4.

MLP employs three fully connected layers to extract features from the whole fMRI time series and all regions. 2D CNN [26, 27] considers fMRI sequence as input and applies a filter size of 5×1 along

Table 3: Evaluation of number of edges.

		0 edge	1.5k edges	5k edges	13k edges
WM	Train	97.93 ± 0.32	99.06 ± 0.27	99.02 ± 0.30	97.22 ± 0.45
	Test	84.33 ± 0.74	86.28 ± 0.75	$\mathbf{86.50 \pm 0.76}$	85.91 ± 0.63
Social	Train	99.55 ± 0.14	99.63 ± 0.13	99.81 ± 0.08	99.48 ± 0.11
	Test	91.36 ± 0.59	$\mathbf{94.40 \pm 0.62}$	94.27 ± 0.65	93.14 ± 0.52
Emotion	Train	99.18 ± 0.22	99.43 ± 0.17	98.93 ± 0.37	98.85 ± 0.22
	Test	88.59 ± 0.67	$\mathbf{90.63 \pm 0.68}$	90.15 ± 0.68	88.80 ± 0.75

Table 4: Classification Accuracy Comparison: Baseline and Advanced ML Models.

	MLP	CNN	GAT	GCN-LSTM	ST-GCN	STNDT	ADB-NN	JGAT
WM	84.33 ± 0.74	84.05 ± 0.86	83.27 ± 0.55	85.53 ± 0.68	83.62 ± 0.93	86.37 ± 0.74	64.37 ± 1.07	$\mathbf{86.50 \pm 0.76}$
Social	91.36 ± 0.59	87.37 ± 1.30	88.06 ± 0.85	92.97 ± 0.61	86.63 ± 0.78	93.88 ± 0.69	72.01 ± 1.20	$\mathbf{94.27 \pm 0.65}$
Emotion	88.59 ± 0.67	80.21 ± 1.28	86.28 ± 0.48	$\mathbf{96.27 \pm 0.25}$	80.08 ± 0.95	89.71 ± 0.69	61.63 ± 0.87	90.15 ± 0.68
mihi	76.05 ± 0.86	64.34 ± 2.60	83.88 ± 1.10	78.68 ± 2.00	63.18 ± 2.43	84.11 ± 1.05	N/A	$\mathbf{90.85 \pm 0.86}$

Table 5: Comparison of three attention heads with two model layers.

	GAT	STNDT	JGAT
WM	83.77 ± 0.87	86.07 ± 0.95	$\mathbf{87.33 \pm 0.96}$
Social	91.06 ± 0.90	94.01 ± 0.88	$\mathbf{95.23 \pm 0.96}$
Emotion	87.59 ± 1.07	90.02 ± 1.00	$\mathbf{90.89 \pm 0.78}$
mihi	82.79 ± 0.86	84.47 ± 0.93	$\mathbf{87.89 \pm 0.48}$

the direction of time series with two CNN layers followed by 2 dense layers. GAT is implemented as a basic spatial graph model which treats fMRI time series of each region as vector features and aggregate neighborhood messages by A collected from DWI data. GCN-LSTM model is inspired by T-GCN model which combines Graph Convolutional Networks (GCN) and Recurrent Neural Network (RNN) models to do traffic prediction [16]. GCN-LSTM is treated as a dynamical graph model since GCN and Long short-term memory (LSTM) [28] can collect information from spatial and temporal directions, though respectively. LSTM is a light version of RNN, which has fewer parameters but can achieve similar performance as RNN.

For the advanced models, ST-GCN [18] is used for action recognition and prediction for the dynamics of human skeletons by combing CNN and GCN for several layers. STNDT [17] employs the design of Neural Data Transformer (NDT) [29] architecture and self-attention mechanism to learn spatial covariation and temporal progression from neural activity datasets. ADB-NN [15] makes use of FC, SC, and GAT to consider both direct and indirect connectivity for analyzing brain activities for patients with Frontal Lobe Epilepsy (FLE) and Temporal Lobe Epilepsy (TLE).

Some models share the GAT architecture such as JGAT, GAT, and STNDT. We only use one attention head with one model layer for these models in Tab. 4. To evaluate the effects of multi-head attention and a deeper model structure, we also provide the comparison of three attention heads with two model layers among these models in Tab. 5.

4 Interpretation of our JGAT model

In this section, we use AS and FS to interpret our method. AS are scores computed from learnable weights W_a and W_c following the attention mechanism in Eq.1. AS for each time frame can be reformed to be a sparse matrix with a dimension of $N \times 3N$, indicating the importance of each edge, given that K has been fixed to be 3. FS are parameters calculated from AS applying to each frame to regulate the relative importance of each frame.

The interpretation process is to formulate a general representation for each cognitive stimuli by mapping AS of 200 ROIs to 14 brain parcellations including both Left Hemisphere (LH) and Right Hemisphere (RH) according to Atlas Schaefer 7 network parcellation [30] (shown in the supplementary material). The standard procedure of the first stage follows three steps:

- Selecting edges in the $N \times 3N$ attention score matrix for each frame with scores above a threshold of $1.1 \times (1/|G_i^t|)$. Later, taking the mean of the filtered attention matrices for all instances

that belong to the same classes. We will get a group-level attention score matrix for each class.

- Mapping the $N \times 3N$ group-level attention matrix to a 14 by 14 matrix following 7 brain parcellations. After the mapping process, calculating the absolute value of the difference between the two classes to represent the contrast between the two classes.
- The output of the second step has dimensions of 14 by 14 by T. To obtain a general pattern, we calculate the average of AS across T and binarize it to get an N by N binary matrix.

The rationale for the above steps is to incorporate the frames with both high AS and contrast between the two classes. To summarize, the first step selects frames with high scores, the second step selects frames with high contrast and the final step extracts the most representative patterns across the whole sequence with both high scores and high contrast. Most importantly, the extracted representations of all cognitive stimuli can be validated by previous literature.

4.1 Working Memory (WM)

In the case of WM, Fig. 3A shows the general pattern of WM which is a 14×14 binary matrix. We can observe a highly activated brain network, Control(6), which constantly receives or sends messages to other brain networks including Dorsal Attention (DorsAttn)(3), Salient Ventral Attention (SalVentAttn)(4), and DorsAttn(10). Control(6) network contains a large area of Prefrontal cortex (PFC) that is related to WM which has been proved by some previous literature [31, 32] and the original HCP paper of task fMRI data [33]. In addition, Fig. 3B illustrates the diagram of FS with SEM, which reveals distinguishable and significant segments that reflect the trend of time frames. We extract subgraphs from the consecutive time frames of these segments to observe the dynamical connectivity of WM stimuli. These subgraphs form a dynamical pathway that describes the dynamical activity of WM, as shown in Fig. 3C. The square matrix shows the binary general pattern as shown in Fig. 3A. For each highlighted yellow square, we visualize the edges in the original 200-node graph. The multiple redness levels indicate the attention scores of the edges.

At the beginning of the WM task, although the FS shows that it is a distinguishable segment (red segment in Fig. 3B), it locates in the valley region, which indicates that the early stage of WM has less influence on the downstream brain-decoding task. WM-related neural activity is generated on the RH indicating dominating communications between DorsAttn(10) and SalVentAttn(11). In the middle stage, FS shows higher values (green segment in Fig. 3B), and the subgraphs of this stage diffuse from the RH to LH. Most importantly, the 2-back WM trial starts to dominate the neural activities and it reaches the peak around the 20th-25th frames, which form the most informative segment and explainable dynamical graphs. Information flows significantly from Control (6) to DorsAttn(3) and SalVentAttn(4). In the last stage of stimuli (blue segment in Fig. 3B), the WM task loses the retrieval of the 2-back trial and is dominated by the 0-back trial instead. The subgraphs during this period also diffuse back to the RH, which is a similar location but in a different direction than the beginning of the middle stage. This phenomenon could indicate that the 2-back WM is less related to the RH since it loses the retrieval during these time intervals. According to Fig. 3B and C, there is more information flowing from the Control network to the Limbic Network in the left hemisphere between the 15th and 25th frames, indicating that it might be the pathway for 2-back WM trials.

4.2 Social

In the case of Social, Fig. 4A shows the general pattern of Social. From Fig. 4A, two brain parcellations, Limbic(5), Default(7), are highlighted which have a relationship with Social stimuli proven by previous literature [33, 34]. To be more specific, both Limbic(5) and Default(7) consist partially of the temporal region of the brain which is indicated to be the main target of social cognition [33], especially the temporal parietal junction located at Default(7). Moreover, Fig. 4B shows the diagram of FS with SEM, which reveals the trend of both cognitive stimuli. We also extract subgraphs from the consecutive time frames of these segments to observe the dynamical connectivity of Social. These subgraphs form a dynamical pathway that describes the dynamical activity of Social, as shown in Fig. 4C. The square matrix shows the binary general pattern as shown in Fig. 4A. For each highlighted yellow square, we visualize the edges in the original 200-node graph. The multiple redness levels indicate the attention scores of the edges.

In the early stage of Social stimuli (red segment in Fig. 4B), due to the highest FS and the higher retrieval of the Mental trial, it should be the most recognizable duration for the model. The informative edges concentrate in the Default network and we can also observe the Default network in this stage

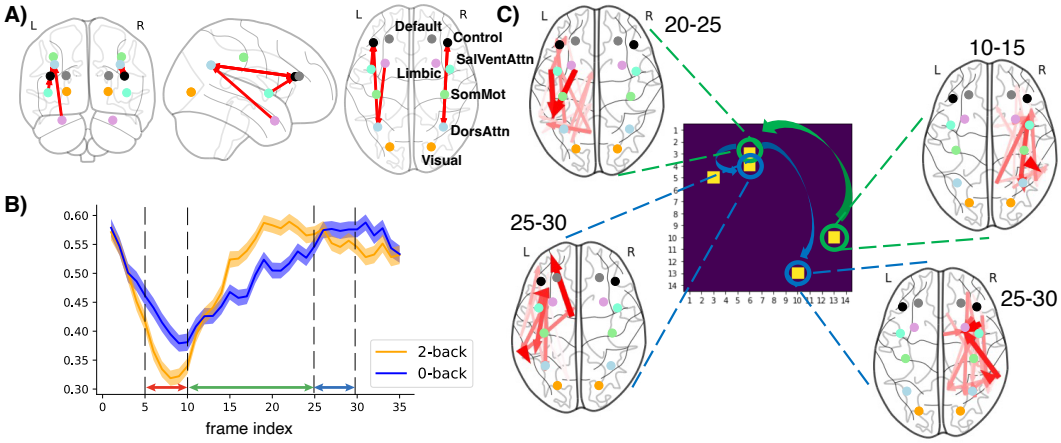


Figure 3: A) The general pattern of WM connectivity. B) FS of WM. C) Pathway of WM (Numbers indicate time frames).

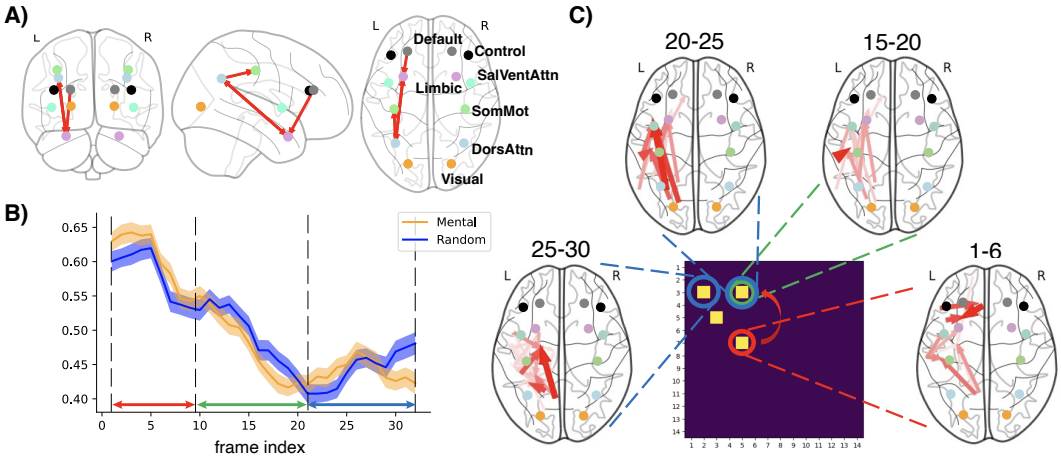


Figure 4: A) The general pattern of Social connectivity. B) FS of Social. C) Pathway of Social (Numbers indicate time frames).

from Fig. 4C. There is dominating communication between Default(7) and Limbic(5). During the middle stage (green segment in Fig. 4B), FS decreases dramatically, and the Random trial dominates as social signals decrease. This indicates that the middle stage is a less important stage for the decoding-task compared to the first stage. The edges in the subgraphs of this stage are predominantly light red, indicating low weights for these edges. Later, the Mental trial dominates for a short period again in the last stage (blue segment in Fig. 4B). However, since this short period is still around the valley of stimuli, this brief retrieval has minor importance for the decoding problem. In summary, from the early stage and the short period of the last stage in Fig. 4C, we can claim that the Default network on the LH is critical for the Mental trial, as it communicates constantly with many other networks during the stimuli, especially Limbic. The first stage might be the best segment to describe the pathway of the Mental trial.

4.3 Emotion

In the case of Emotion, Fig. 5A shows the general pattern of Emotion. From Fig. 5A, one clear brain parcellation, Limbic(5, 12), is emphasized in both hemispheres. The Limbic contains amygdala and hippocampus which are brain regions that relate to Emotion stimuli [33]. Additionally, Fig. 5B shows the diagram of FS with SEM, which reveals the trend of both cognitive stimuli. We also extract subgraphs from the consecutive time frames of these segments to observe the dynamical connectivity

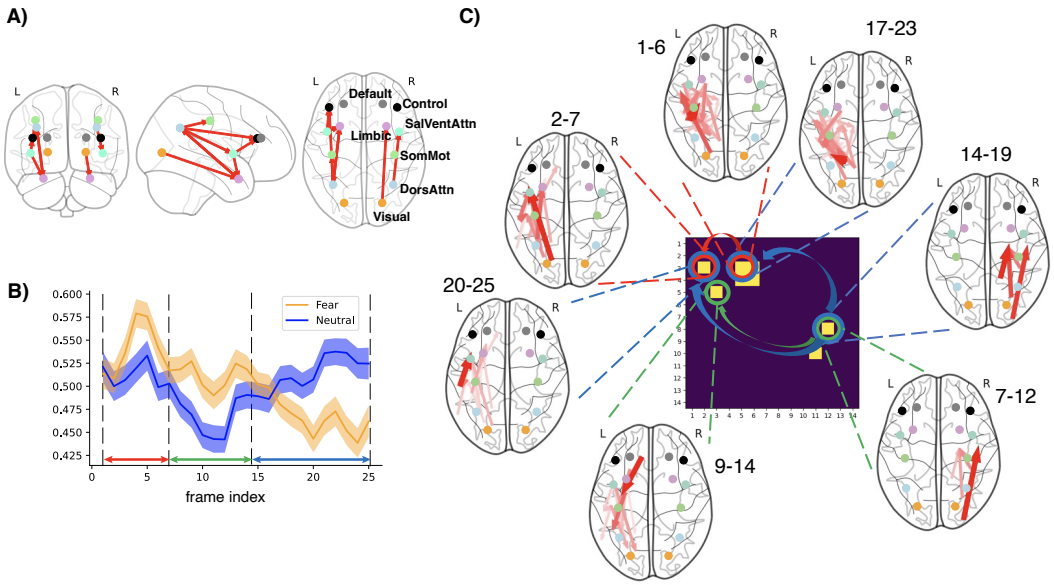


Figure 5: A) The general pattern of Emotion connectivity. B) FS of Emotion. C) Pathway of Emotion (Numbers indicate time frames).

of Emotion. These subgraphs form a dynamical pathway that describes the dynamical activity of Social, as shown in Fig. 5C. The square matrix shows the binary general pattern as shown in Fig. 5A. For each highlighted yellow square, we visualize the edges in the original 200-node graph. The multiple redness levels indicate the attention scores of the edges.

The trend of the Emotion FS is similar to that of the Social FS since their highly activated segments are at the beginning of their respective stimulus tasks. During the early stage of this stimuli (red segment in Fig. 5B), the Fear trial dominates the majority of signals and the signals clearly flow from the DorsAttn network to the Limbic network. We can also observe the subgraph emphasize the similar brain regions during this period in Fig. 5C. During the middle stage (green segment in Fig. 5B), the Fear trial still dominates the response, but both Fear and Neutral trials are relatively low compared with the first stage, which leads to non-representative edges for dynamical graphs. At the end of the Emotion task (blue segment in Fig. 5B), the main trial, Fear, drops to the valley, while the Neutral stimulus remains at the average level. The significant drops in FS could be interpreted as fatigue towards the Fear stimulus, but it needs more experiments to confirm this interpretation. In short, the first stage of the Emotion task can best delineate the pathway of the Fear trial, which flows from the DorsAttn network to the Limbic network. This is due to the first stage's highest FS and the domination of the Fear trial. Furthermore, the last stage is also recognizable in the brain-decoding process given the highest contrast between the Fear/Neutral trials. Since the response of Fear trials disappears at last, it is difficult to study Fear stimuli from this segment.

5 Conclusion

In this paper, we propose a Joint kernel Graph Attention Network (JGAT) to integrate Functional Connectivity (FC) and Structural Connectivity (SC) and take temporal variations into account simultaneously. In our model, we define joint graphs for the fMRI inputs to capture edge connections across multiple time frames. Additionally, we employ Attention Scores (AS) and Frame Scores (FS) to regulate the features of neighboring nodes within these graphs. This allows us to incorporate temporal information and optimize the representation of the data. In general, our method achieves higher accuracy when compared to both baseline ML models and some advanced ones. Finally, by analyzing AS and FS, we are able to provide meaningful interpretations for dynamical connectivity and construct pathways of informative segments across the time series of cognitive stimuli for the HCP datasets. We include a discussion of limitations and future works in the supplementary material.

References

[1] Ed Bullmore and Olaf Sporns, "Complex brain networks: graph theoretical analysis of structural and functional systems," *Nature Rev. Neurosci.*, vol. 10, no. 3, pp. 186–198, 2009.

[2] Marlies E. Vissers, Michael X Cohen, and Hilde M. Geurts, "Brain connectivity and high functioning autism: A promising path of research that needs refined models, methodological convergence, and stronger behavioral links," *Neuroscience and Biobehavioral Reviews*, vol. 36, no. 1, pp. 604–625, 2012.

[3] Jennifer Fitzsimmons, Marek Kubicki, and Martha E Shenton, "Review of functional and anatomical brain connectivity findings in schizophrenia," *Curr Opin Psychiatry*, vol. 26, no. 2, pp. 172–187, 2013.

[4] Alex Fornito, Andrew Zalesky, Christos Pantelis, and Edward T. Bullmore, "Schizophrenia, neuroimaging and connectomics," *NeuroImage*, vol. 62, no. 4, pp. 2296–2314, 2012.

[5] Grega Repovs, John G. Csernansky, and Deanna M. Barch, "Brain network connectivity in individuals with schizophrenia and their siblings," *Biological Psychiatry*, vol. 69, no. 10, pp. 967–973, 2011.

[6] Matthew T. Sutherland, Meredith J. McHugh, Vani Pariyadath, and Elliot A. Stein, "Resting state functional connectivity in addiction: Lessons learned and a road ahead," *NeuroImage*, vol. 62, no. 4, pp. 2281–2295, 2012.

[7] Leslie A. Hulvershorn, Kathryn Cullen, and Amit Anand, "Toward dysfunctional connectivity: a review of neuroimaging findings in pediatric major depressive disorder," *Brain Imaging and Behavior*, vol. 5, p. 307–328, 2011.

[8] Stephen M Strakowski, Caleb M Adler, Jorge Almeida, Lori L Altshuler, Hilary P Blumberg, Kiki D Chang, Melissa P DelBello, Sophia Frangou, Andrew McIntosh, Mary L Phillips, Jessika E Sussman, and Jennifer D Townsend, "The functional neuroanatomy of bipolar disorder: a consensus model," *Bipolar Disorders*, vol. 14, no. 4, pp. 313–325, 2012.

[9] Ulrike Basten, Kirsten Hilger, and Christian J. Fiebach, "Where smart brains are different: A quantitative meta-analysis of functional and structural brain imaging studies on intelligence," *Intelligence*, vol. 51, pp. 10–27, 2015.

[10] B. Blair Braden, Christopher J. Smith, Amiee Thompson, Tyler K. Glaspy, Emily Wood, Divya Vatsa, Angela E. Abbott, Samuel C. McGee, and Leslie C. Baxter, "Executive function and functional and structural brain differences in middle-age adults with autism spectrum disorder," *Autism Research*, vol. 10, no. 12, pp. 1945–1959, 2017.

[11] Jiashuang Huang, Qi Zhu, Mingliang Wang, Luping Zhou, Zhiqiang Zhang, and Daoqiang Zhang, "Coherent pattern in multi-layer brain networks: Application to epilepsy identification," *IEEE J Biomed Health Inform.*, vol. 24, no. 9, pp. 2609–2620, 2020.

[12] Y. Li, G. Mateos, and Z. Zhang, "Learning to model the relationship between brain structural and functional connectomes," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 8, pp. 830–843, 2022.

[13] Martin Dyrba, Michel Grothe, Thomas Kirste, and Stefan J Teipel, "Multimodal analysis of functional and structural disconnection in alzheimer's disease using multiple kernel svm," *Hum. Brain Mapping*, vol. 36, no. 6, pp. 2118–2131, 2015.

[14] Xiaoxiao Li, Yuan Zhou, Nicha Dvornek, Muhan Zhang, Siyuan Gao, Juntang Zhuang, Dustin Scheinost, Lawrence H. Staib, Pamela Ventola, and James S. Duncan, "Braingnn: Interpretable brain graph neural network for fmri analysis," *Medical Image Analysis*, vol. 74, p. 102233, 2021.

[15] Jiashuang Huang, Luping Zhou, Lei Wang, and Daoqiang Zhang, "Attention-diffusion-bilinear neural network for brain network analysis," *IEEE TRANSACTIONS ON MEDICAL IMAGING*, vol. 39, no. 7, pp. 2541–2552, 2020.

[16] Ling Zhao, Yujiao Song, Chao Zhang, Yu Liu, Pu Wang, Tao Lin, Min Deng, and Haifeng Li, "T-gcn: A temporal graph convolutionalnetwork for traffic prediction," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 9, pp. 3848–3858, 2020.

[17] T. Le and E. Shlizerman, "Stndt: Modeling neural population activity with a spatiotemporal transformer," *arXiv preprint arXiv:2206.04727*, 2022.

[18] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, 2018.

[19] J. S. Damoiseaux, S. Rombouts, F. Barkhof, P. Scheltens, C. J. Stam, S. M. Smith, and C. F. Beckmann, "Consistent resting-state networks across healthy subjects," *Proceedings of the national academy of sciences*, vol. 103, no. 37, pp. 13848–13853, 2006.

[20] M. D. Greicius, B. Krasnow, A. L. Reiss, and V. Menon, "Functional connectivity in the resting brain: a network analysis of the default mode hypothesis," *Proceedings of the national academy of sciences*, vol. 100, no. 1, pp. 253–258, 2003.

[21] W. R. Shirer, S. Ryali, E. Rykhlevskaia, V. Menon, and M. D. Greicius, "Decoding subject-driven cognitive states with whole-brain connectivity patterns," *Cerebral cortex*, vol. 22, no. 1, pp. 158–165, 2012.

[22] R. Salvador, J. Suckling, M. R. Coleman, J. D. Pickard, D. Menon, and E. Bullmore, "Neurophysiological architecture of functional magnetic resonance images of human brain," *Cerebral cortex*, vol. 15, no. 9, pp. 1332–1342, 2005.

[23] David C. Van Essen, Stephen M. Smith, Deanna M. Barch, Timothy E.J. Behrens, Essa Yacoub, and Kamil Ugurbil, "The wu-minn human connectome project: An overview," *NeuroImage*, vol. 80, pp. 62–79, 2013.

[24] Eva L. Dyer, Mohammad Gheshlaghi Azar, Matthew G. Perich, Hugo L. Fernandes, Stephanie Naufel, Lee E. Miller, and Konrad P. Körding, "A cryptography-based approach for movement decoding," *Nature Biomedical Engineering*, vol. 1, no. 12, p. 967–976, 2017.

[25] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *arXiv preprint arXiv:1710.10903*, 2017.

[26] Jeremy Kawahara, Colin J Brown, Steven P Miller, Brian G Booth, Vann Chau, Ruth E Grunau, Jill G Zwicker, and Ghassan Hamarneh, "Brainnetcnn: Convolutional neural networks for brain networks; towards predicting neurodevelopment," *Neuroimage*, vol. 146, pp. 1038–1049, 2017.

[27] Biao Jie, Mingxia Liu, Chunfeng Lian, Feng Shi, and Dinggang Shen, "Designing weighted correlation kernels in convolutional neural networks for functional connectivity based brain disease diagnosis," *Medical Image Analysis*, vol. 63, p. 101709, 2020.

[28] J. Dakka, P. Bashivan, M. Gheiratmand, I. Rish, S. Jha, and R. Greiner, "Learning neural markers of schizophrenia disorder using recurrent neural networks," *arXiv preprint arXiv:1712.00512*, 2017.

[29] J. Ye and C. Pandarinath, "Representation learning for neural population activity with neural data transformers," *arXiv preprint arXiv:2108.01210*, 2021.

[30] Alexander Schaefer, Ru Kong, Evan M Gordon, Timothy O Laumann, Xi-Nian Zuo, Avram J Holmes, Simon B Eickhoff, and B T Thomas Yeo, "Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity mri," *Cerebral Cortex*, vol. 28, no. 9, p. 3095–3114, 2018.

[31] Bonnie J. Nagel, Megan M. Herting, Emily C. Maxwell, Richard Bruno, and Damien Fair, "Hemispheric lateralization of verbal and spatial working memory during adolescence," *Brain Cogn.*, vol. 82, no. 1, pp. 58–68, 2013.

[32] Wen Jia Chai, Aini Ismafairus Abd Hamid, and Jafri Malin Abdullah, "Working memory from the psychological and neurosciences perspectives: A review," *Front. Psychol.*, vol. 9, 2018.

[33] Deanna M Barch, Gregory C Burgess, Michael P Harms, et al., "Function in the human connectome: Task-fmri and individual differences in behavior," *NeuroImage*, vol. 80, pp. 169–189, 2013.

[34] Riitta Hari and Miiamaaria V. Kujala, "Brain basis of human social interaction: From concepts to brain imaging," *Physiological Review*, vol. 89, no. 2, pp. 453–479, 2009.

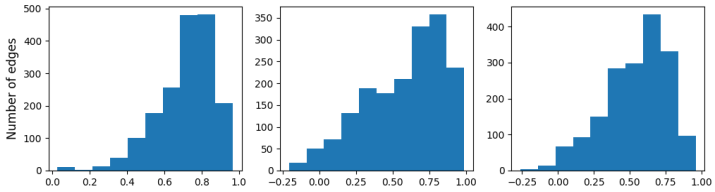
Supplementary Material

Table 6: Brain 7 Networks Parcellation

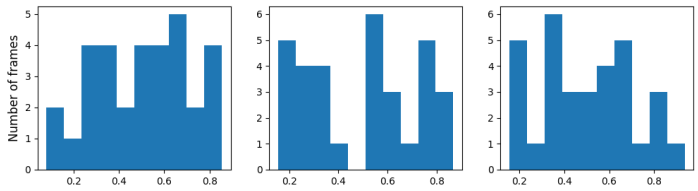
Left Hemisphere(LH)			Right Hemisphere(RH)		
Number	Network	Nodes range	Number	Network	Nodes range
1	Visual	1 – 14	8	Visual	101 - 115
2	Somatosensory/Motor (SomMot)	15 – 30	9	SomMot	116 - 134
3	DorsAttn	31 – 43	10	DorsAttn	135 - 147
4	SalVentAttn	44 – 54	11	SalVentAttn	148 - 158
5	Limbic	55 – 60	12	Limbic	159 - 164
6	Control	61 – 73	13	Control	165 - 181
7	Default	74 – 100	14	Default	182 - 200

Effect of FC and FS

From the modeling perspective, we involve FC and FS to regulate the dynamical features. We also do the ablation study on these two scores in Tab. 2. By observing the results, it seems that for the same datasets such as WM and Social, there does not exist a significant difference by applying these two scores. However, we can move a step back to visualize the histogram of two scores first in Fig. 6. The plots show the histograms of FC and FS from some random Social instances. These plots indicate that these scores do have certain distributions instead of barely uniform values. After employing two scores with certain distributions, the results do not drop and some of them even improve. Furthermore, the FS is useful in the interpretation part as it helps identify numerous meaningful edges and brain parcellations.



(a) Functional Connectivity (FC)



(b) Frame Scores (FS)

Figure 6: Histogram of FC and FS

Limitations and Future Works

There are several limitations related to the JGAT model. First of all, we can only provide group-level interpretation for this model at this stage. The individual explanation is a more complex topic, but it is certainly worth exploring and delving into due to its significance and intriguing nature. One of the reasons for the presence of noise in our interpretation is the significant influence of individual differences among all instances when averaging our results. Therefore, providing individual explanations has the potential to yield clearer dynamical graphs and results by accounting for the specific characteristics and variations present in each individual's data. Some thoughts on individual interpretation could involve utilizing individual adjacency matrices instead of a general one or adding additional loss to the control group or individual level. Secondly, running time is an issue for this model. Due to the large number of edges in our method, which is three times more than other models, and the inclusion of multiple scores to enhance the model, our structure appears to be

the most complex when compared to other models used in the comparison. In the future, it would be better to optimize the algorithm and codes. Lastly, for datasets that lack an adjacency matrix or similar information, such as the mihi dataset, the only option is to use a fully connected adjacency matrix as input. In such cases, the model does not have access to specific connectivity information and treats all brain regions as fully connected. Indeed, using a fully connected adjacency matrix can pose challenges when trying to generate a meaningful interpretation using the method we employed. In essence, datasets of this type may not be suitable for our current method, or it may be necessary to develop an alternative interpretation approach specifically tailored for fully connected adjacency matrices.