fTSPL: Enhancing Brain Analysis with fMRI-Text Synergistic Prompt Learning

Pengyu Wang[1], Huaqi Zhang[2], Zhibin He[3], Zhihao Peng[1], Yixuan Yuan[1(✉)]

[1] Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong SAR, China
yxyuan@ee.cuhk.edu.hk
[2] School of Computer Science and Technology, Beijing Jiaotong University, Beijing, China
[3] School of Automation, Northwestern Polytechnical University, Xi'an, China

Abstract. Using functional Magnetic Resonance Imaging (fMRI) to construct the functional connectivity is a well-established paradigm for deep learning-based brain analysis. Recently, benefiting from the remarkable effectiveness and generalization brought by large-scale multi-modal pre-training data, Vision-Language (V-L) models have achieved excellent performance in numerous medical tasks. However, applying the pre-trained V-L model to brain analysis presents two significant challenges: (1) The lack of paired fMRI-text data; (2) The construction of functional connectivity from multi-modal data. To tackle these challenges, we propose a fMRI-Text Synergistic Prompt Learning (fTSPL) pipeline, which utilizes the pre-trained V-L model to enhance brain analysis for the first time. In fTSPL, we first propose an Activation-driven Brain-region Text Generation (ABTG) scheme that can automatically generate instance-level texts describing each fMRI, and then leverage the V-L model to learn multi-modal fMRI and text representations. We also propose a Prompt-boosted Multi-modal Functional Connectivity Construction (PMFCC) scheme by establishing the correlations between fMRI-text representations and brain-region embeddings. This scheme serves as a plug-and-play preliminary that can connect with various Graph Neural Networks (GNNs) for brain analysis. Experiments on ABIDE and HCP datasets demonstrate that our pipeline outperforms state-of-the-art methods on brain classification and prediction tasks. The code is available at https://github.com/CUHK-AIM-Group/fTSPL.

Keywords: Prompt learning · Vision-language model · Multi-modal functional connectivity · Brain analysis.

1 Introduction

At the neuroscience fronts, brain functional Magnetic Resonance Imaging (fMRI) [16] is a key technology for revealing human behaviors and cognitions. Concretely,

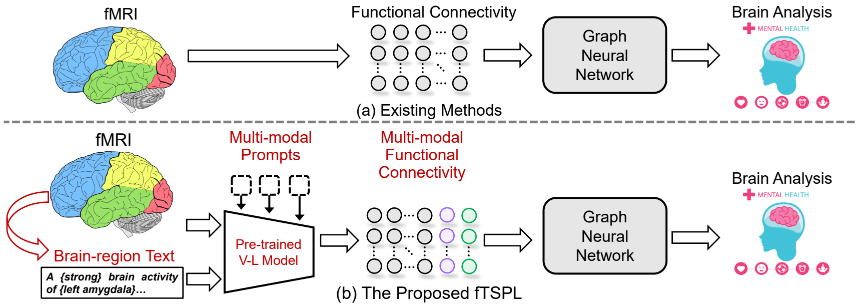


Fig. 1. Illustration of the proposed fTSPL pipeline. (a) Existing methods directly produce the functional connectivity and then use GNNs for brain analysis. (b) Our fTSPL utilizes the pre-trained V-L model with multi-modal prompts to construct the multi-modal functional connectivity, thereby enhancing GNN-based brain analysis.

fMRI produces functional connectivity [5,20] that describing the communication and collaboration patterns between brain regions in different behaviors and cognitions. Therefore, many studies [4,21] focus on analyzing functional connectivities for multiple brain disease classification and cognitive prediction tasks.

With the interdisciplinarity between artificial intelligence and neuroscience, deep learning methods are widely applied to assist in brain analysis [6,12,15,25,29]. As shown in Fig. 1(a), a well-established paradigm is constructing graph-structured functional connectivity, followed by connecting a Graph Neural Network (GNN) to classify brain diseases or predict brain cognitions. However, most existing GNNs are tailored for uni-modal fMRI data, which limits their effectiveness since recent works [13,27,30] have shown that incorporating text modality can provide additional supervision to improve performance. Currently, the pre-trained Vision-Language (V-L) models [10,18] have attracted extensive attention as they utilize a self-supervised manner to learn numerous generic and effective multi-modal representations from large-scale pre-training data. In the medical field, V-L models have also been explored in various tasks and have yielded promising results [24,26,28]. Inspired by this, we aim to introduce the pre-trained V-L model for constructing multi-modal functional connectivity, thus improving the performance of multiple brain analysis tasks. To be noted, we represent the first effort to leverage the V-L model for multi-modal brain analysis.

Nevertheless, there are two major challenges to applying the V-L model for multi-modal brain analysis: (1) Current fMRI data lacks the corresponding texts. Meaningful texts describing brain-region connectivities and activities could provide the extra text-modal supervision to learn more effective fMRI representations. Therefore, it is highly demanded for generating instance-level fMRI-text data; (2) Existing functional connectivity construction methods only consider uni-modal fMRI data. We aim to further explore the relations between high-level fMRI, text, and brain-region features to construct the multi-modal functional connectivity and improve the performance of brain analysis.

Aiming to address the above challenges, we propose a fMRI-Text Synergistic Prompt Learning (fTSPL) pipeline for multi-modal brain analysis, which comprises two main components: (1) Activation-driven Brain-region Text Generation (ABTG); (2) Prompt-boosted Multi-modal Functional Connectivity Construction (PMFCC). In ABTG, we screen the activated brain regions according to fMRI intensities, and quantify the activation degree of brain regions using functional connectivity. This scheme allows us to obtain the text description of each fMRI and enables the use of the V-L model. In PMFCC, we tune the pre-trained V-L model via multi-modal prompts to produce fMRI and text representations, and then construct the multi-modal functional connectivity by establishing the correlations between fMRI-text representations and brain-region embeddings. Experimental results demonstrate that the proposed pipeline achieves excellent performance on multiple brain analysis tasks. The main contributions are as follows: (1) We propose a novel prompt learning paradigm fTSPL. To the best of our knowledge, this is the first application of the V-L model for multi-modal brain analysis; (2) We propose ABTG to provide instance-level text descriptions for fMRI, which is suitable for fMRI with different brain atlas; (3) We propose PMFCC to construct the multi-modal functional connectivity, it is a plug-and-play preliminary for GNN-based brain analysis. Experiments on brain disease classification and cognitive prediction verified the effectiveness of PMFCC.

2 Method

In Fig. 2, we display the overall architecture of the proposed fTSPL pipeline. Firstly, given the pre-processed fMRI time-series $\{X_n\}_{n=1}^{N}$, we use the ABTG scheme to generate fMRI's text descriptions $\{Y_n\}_{n=1}^{N}$ according to the connectivity and activity of brain regions. The fMRI time-series and brain-region text $\{X_n, Y_n\}$ are denoted as the multi-modal input for BiomedCLIP [24] text and image encoders. Afterward, we design learnable multi-layer text and image prompts $\{P_k^T\}_{k=1}^{K}$ and $\{P_k^I\}_{k=1}^{K}$, where K is the prompt depth. The text and image encoders E_T and E_I are kept frozen, while only multi-modal prompts are optimized to produce fMRI and text representations x and y. Next, we propose the PMFCC scheme to enhance the original functional connectivity by supplementing the fMRI and text-modal connectome information. The constructed multi-modal functional connectivity F_m can connect different GNNs as adapters to improve multiple brain analysis tasks. During the training, the contrastive loss $\mathcal{L}_{con}$ and task loss $\mathcal{L}_{task}$ jointly achieve the optimization of our pipeline.

2.1 Activation-driven Brain-region Text Generation (ABTG)

In fMRI time-series, each brain region corresponds to an independent medical terminology, and the connectivity and activity of brain regions are crucial for neuroscientists to achieve brain analysis. Motivated by this, we propose the ABTG scheme, which counts the activated brain regions of each fMRI as well as the corresponding activation degrees to provide instance-level brain-region texts.

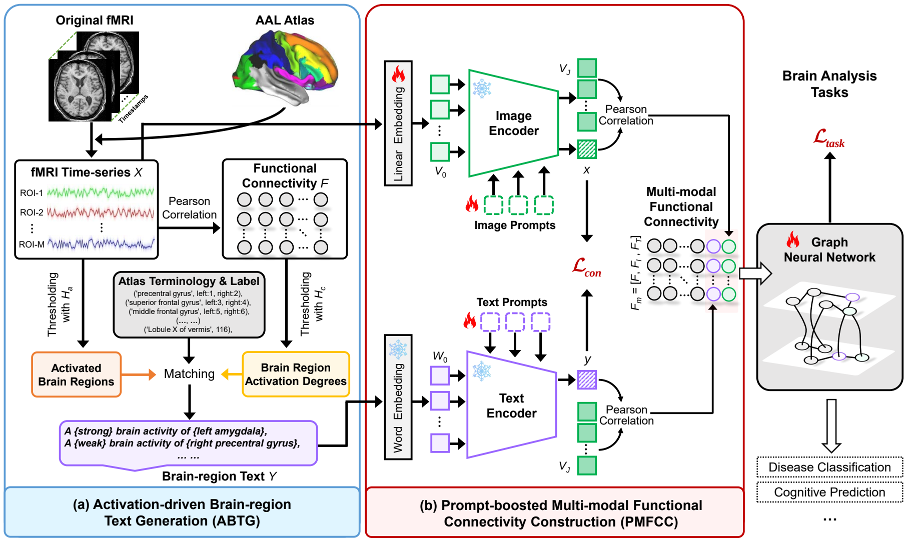

Original fMRI
AAL Atlas
Timestamps
fMRI Time-series X
ROI-1
ROI-2
ROI-M
Pearson Correlation
Functional Connectivity F
Thresholding with H_a
Atlas Terminology & Label
('precentral gyrus', left:1, right:2),
('superior frontal gyrus', left:3, right:4),
('middle frontal gyrus', left:5, right:6),
(..., ...)
('Lobule X of vermis', 116),
Thresholding with H_c
Activated Brain Regions
Matching
Brain Region Activation Degrees
A {strong} brain activity of {left amygdala},
A {weak} brain activity of {right precentral gyrus},
... ...
Brain-region Text Y
(a) Activation-driven Brain-region Text Generation (ABTG)
Linear Embedding
V_0
Image Encoder
V_J
Pearson Correlation
x
Image Prompts
$\mathcal{L}_{con}$
Text Prompts
Word Embedding
W_0
Text Encoder
y
Pearson Correlation
V_J
Multi-modal Functional Connectivity
$F_m = [F, F_I, F_T]$
(b) Prompt-boosted Multi-modal Functional Connectivity Construction (PMFCC)
Brain Analysis Tasks
$\mathcal{L}_{task}$
Graph Neural Network
Disease Classification
Cognitive Prediction
...

Fig. 2. Overview of the proposed fTSPL pipeline. fTSPL first generates the brain-region text for each fMRI time-series to form instance-level multi-modal data. Then, fTSPL tunes the pre-trained V-L model by optimizing multi-modal prompts to produce fMRI and text representations. Finally, fTSPL correlates the fMRI-text representations and brain-region embeddings to construct the multi-modal functional connectivity, and connects a learnable GNN adapter for enhancing brain analysis.

Firstly, we map original fMRI data into the grayordinate system [7] to obtain vertices on the reconstructed cortical surface, and then perform the within-subject and cross-subject registrations to establish the subject-level correspondence. Taking Automated Anatomical Labeling (AAL) [19] as an example, we apply the pre-defined AAL atlas to cortical surfaces, which can produce 45 regions on left and right cerebral hemisphere and 26 regions on the cerebellum. At each fMRI timestamp, the vertices in each brain region are averaged to produce the fMRI time-series $X \in \mathbb{R}^{M \times S}$, where M is the number of brain regions, and S is the number of timestamps. To obtain the activation state and degree of each region, we first introduce an activation threshold H_a, which aims to extract the activated brain region in X at each timestamp. H_a is positively correlated with the maximum activity of fMRI time-series, i.e., $H_a = \lambda \cdot \max(X)$. We define an activated brain region if its value exceeds H_a more than $S/20$ times:

$$\text{The } m\text{−th brain region is} \begin{cases} activated, \text{ if } \sum_{s=1}^{S} \mathbb{1}(X_{m,s} > H_a) \geq \frac{S}{20} \\ non-activated, \text{ if } \sum_{s=1}^{S} \mathbb{1}(X_{m,s} > H_a) < \frac{S}{20} \end{cases} . \quad (1)$$

Then, we define a correlation threshold H_c to evaluate the activation degree of brain regions. Concretely, Pearson correlations between brain regions are computed to construct a 116 × 116 functional connectivity. Here, we consider two regions to be correlated when their correlation value is greater than 0.5. On this basis, we further define the “strong”, “moderate”, or “weak” region as the one with M_1, M_2, or M_3 correlated regions, where $M_1 > 3H_c$, $H_c \leq M_2 \leq 3H_c$, and $M_3 < H_c$. Finally, we describe the above brain regions to generate the instance-level brain-region text. To be specific, the activated brain regions are matched with the corresponding brain atlas terminologies, and further combined with their activation degree words “strong”, “moderate”, and “weak”.

2.2 Prompt-boosted Multi-modal Functional Connectivity Construction (PMFCC)

To enhance the original functional connectivity, we combine the pre-trained V-L model and multi-modal prompt learning to construct the multi-modal functional connectivity, which theoretically contributes to improve the performance of multiple brain analysis tasks.

Multi-modal Text-Image Prompting: In this work, multi-modal prompt learning is achieved by synergizing multi-layer text prompts $\{P_k^T\}_{k=1}^K$ with image prompts $\{P_k^I\}_{k=1}^K$. Given a brain-region text Y_n, we first adopt a Tokenizer embedding layer to convert Y_n into the word embedding $W_0 \in \mathbb{R}^{B \times C^T}$, where B and C^T are the number and channels of word embeddings. Then, we take the current layer text prompt $P_0^T \in \mathbb{R}^{D \times C^T}$, where D is the prompt length, to combine with W_0 as inputs for the text encoder E_T. For the j-th Transformer layer, we define the prompting process as:

$$[W_j, \_] = E_{T,j}(W_{j-1}, P_j^T), \quad j \leq K, \tag{2}$$

when $j \leq K$, we will discard the prompt output of the current layer, and add a new text prompt in the next Transformer layer for learning.

Finally, we introduce a linear layer to project the last word token w_J^B from W_J into a common latent space, obtaining the text representation y:

$$y = \mathrm{TextProj}(w_J^B), \quad y \in \mathbb{R}^C. \tag{3}$$

As the fMRI time-series X_n has been pre-processed according to AAL atlas, we remove the image patching step and add a linear layer to convert $X_n \in \mathbb{R}^{M \times S}$ into the image embedding $V_0 \in \mathbb{R}^{M \times C^I}$, where C^I is the channels of image embeddings. Meanwhile, we add the class token c_0 and image prompt $P_0^I \in \mathbb{R}^{D \times C^I}$ into V_0 to feed the image encoder E_I:

$$[c_j, V_j, \_] = E_{I,j}(c_{j-1}, V_{j-1}, P_j^I), \quad j \leq K. \tag{4}$$

After image prompting, we extract the class token c_J and project it into a common latent space using a linear layer to obtain the fMRI representation x:

$$x = \mathrm{ImageProj}(c_J), \quad x \in \mathbb{R}^C. \tag{5}$$

Multi-modal Functional Connectivity: Given the fMRI representation $x \in \mathbb{R}^{C}$, text representation $y \in \mathbb{R}^{C}$, and brain-region embeddings $V_J \in \mathbb{R}^{M \times C^{I}}$. Since x, y, and H_J have different dimensions, we first use a linear layer to project V_J as $\bar{V}_J \in \mathbb{R}^{M \times C}$. Then, we compute Pearson correlations between x and $\bar{V}_J$ to obtain the image-modal connectome supplement F_I, which provides the correlations between the global fMRI feature and local brain-region features:

$$F_I = \text{Pearson}\{x, \bar{V}_J\}. \tag{6}$$

Similarly, we compute the text-modal connectome supplement F_T to provide the text-brain region correlations for functional connectivity:

$$F_T = \text{Pearson}\{y, \bar{V}_J\}. \tag{7}$$

Finally, we concatenate F_I and F_T with the original functional connectivity F at the node dimension to obtain the multi-modal functional connectivity F_m:

$$F_m = \text{Concat}[F, F_I, F_T]. \tag{8}$$

For existing GNN methods, the multi-modal functional connectivity F_m can replace F as their inputs. Due to only two extra nodes, F_m has almost no increasing computations. In this work, we use a standard GNN as an adapter to process F_m for brain analysis, which consists of two graph convolutional layers and a linear layer. Specifically, we first construct a graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}, \mathcal{A}\}$, where $\mathcal{V}$ denotes the graph nodes, is F_m. The edge $\mathcal{E}$ and adjacency matrix $\mathcal{A}$ encode the correlations between nodes. The predicted result is obtained by:

$$\mathcal{H}^{(1)} = \sigma(\tilde{\mathcal{D}}^{-\frac{1}{2}}\tilde{\mathcal{A}}\tilde{\mathcal{D}}^{\frac{1}{2}}\mathcal{V}\mathcal{W}^{(1)}), \;\; \mathcal{H}^{(2)} = \text{Proj}(\sigma(\tilde{\mathcal{D}}^{-\frac{1}{2}}\tilde{\mathcal{A}}\tilde{\mathcal{D}}^{\frac{1}{2}}\mathcal{H}^{(1)}\mathcal{W}^{(2)})), \tag{9}$$

where $\tilde{\mathcal{A}} = \mathcal{A} + \text{I}$, I is the identity matrix, $\tilde{\mathcal{D}}$ is the degree matrix, $\mathcal{W}$ is the graph convolution weights, and σ is the activation function.

2.3 Model Optimization

In this work, we adopt the contrastive loss $\mathcal{L}_{con}$ and task loss $\mathcal{L}_{task}$ to jointly optimize the proposed fTSPL pipeline. The former aligns fMRI and text representations x and y, and the latter can be the cross-entropy loss for disease classification, or the L2 loss for cognitive prediction. The total loss function is:

$$\mathcal{L} = \alpha \mathcal{L}_{con} + \beta \mathcal{L}_{task}, \tag{10}$$

where α and β are the loss weights. It is worth noting that only the multi-modal prompts, two linear layers, and GNN are optimized in the training.

3 Experiments

3.1 Experimental Setup

Datasets: We evaluate the effectiveness of the proposed fTSPL pipeline on the ABIDE [3] and HCP [23] datasets. The former corresponds to autism classification, and the latter involves cognitive prediction.

Table 1. Quantitative results of the proposed pipeline and state-of-the-art methods on ABIDE and HCP datasets. The best performance is highlighted in boldface.

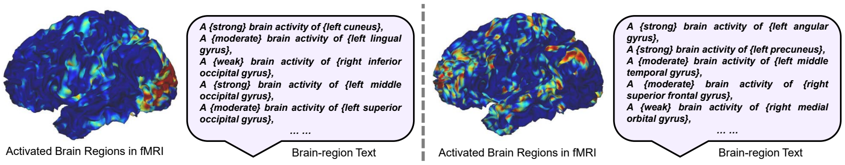
Classification Methods	Accuracy	AUROC	Sensitivity	Specificity
BrainGNN (MedIA21) [14]	62.7±3.7	59.6±2.5	56.8±20.7	70.2±19.3
BrainGB (TMI22) [2]	69.4±3.4	63.2±2.0	63.5±8.6	60.7±10.4
BNT (NeurIPS22) [11]	71.0±1.2	80.2±1.0	72.5±5.2	69.3±6.5
Com-BrainTF (MICCAI23) [1]	72.5±4.4	79.6±3.8	80.1±5.8	65.7±6.4
Ours	**75.4**±2.7	**82.5**±2.4	**81.9**±4.1	**74.1**±4.9
Prediction Methods	MAE	MSE	PCC	R^2
BrainGNN (MedIA21) [14]	0.170±0.004	0.047±0.145	0.195±0.006	0.040±0.004
BrainGB (TMI22) [2]	0.168±0.003	0.044±0.202	0.223±0.004	0.045±0.007
RegGNN (BIB22) [8]	0.164±0.005	0.040±0.124	0.280±0.002	0.057±0.005
Meta-RegGNN (PRIME22) [9]	0.161±0.004	0.038±0.168	0.304±0.003	0.066±0.004
Ours	**0.156**±0.003	**0.035**±0.157	**0.369**±0.004	**0.101**±0.003

ABIDE Dataset: This dataset contains 1035 subjects' fMRI data. According to AAL atlas, we produce fMRI time-series and functional connectivities, which have two classes: *autism* or *non-autism*. To improve reliability, we evenly split this dataset into five subsets for 5-fold cross-validation. Each fold uses 1 subset for test and the other 4 for training, each subset containing 207 subjects.

HCP Dataset: This dataset collects 870 subjects' fMRI data, we also perform AAL atlas to produce fMRI time-series and functional connectivities. Then, we choose a representative task that predicting the cognitive score of "ReadEng". We define 174 subjects as a subset to achieve 5-fold cross-validation.

Evaluations: For classification tasks, we adopt the Accuracy, AUROC, Sensitivity, and Specificity as evaluation metrics. The higher the values of these metrics, the better the performance. To further quantify prediction performance, we introduce the Mean Absolute Error (MAE), Mean Squared Error (MSE), Pearson Correlation Coefficient (PCC), and R-squared (R^2) as evaluation metrics. The lower MAE and MSE values, and the higher PCC and R^2 values indicate competitive results. After that, we compare the proposed fTSPL pipeline with some state-of-the-art brain analysis methods, including BrainGNN [14], BrainGB [2], BNT [11], Com-BrainTF [1], RegGNN [8], and Meta-RegGNN [9].

Implementation Details: Our pipeline is implemented by PyTorch 1.18.0 [17] and NVIDIA 4090 GPU. For classification tasks, we use the SGD optimizer [22] with 5×10^{-4} learning rate, 1×10^{-4} weight decay, 16 batch size, and 50 epochs. The loss weights are α=0.5 and β=0.5. For prediction tasks, we use the SGD optimizer with 1×10^{-3} learning rate, 1×10^{-4} weight decay, 32 batch size, and 50 epochs. The loss weights are α=0.2 and β=0.8. For multi-modal prompts, we set the depth K=4 and the length D=2. In addition, we set λ=0.6 in H_a.



A {strong} brain activity of {left cuneus},
A {moderate} brain activity of {left lingual gyrus},
A {weak} brain activity of {right inferior occipital gyrus},
A {strong} brain activity of {left middle occipital gyrus},
A {moderate} brain activity of {left superior occipital gyrus},
... ...

A {strong} brain activity of {left angular gyrus},
A {strong} brain activity of {left precuneus},
A {moderate} brain activity of {left middle temporal gyrus},
A {moderate} brain activity of {right superior frontal gyrus},
A {weak} brain activity of {right medial orbital gyrus},
... ...

Activated Brain Regions in fMRI
Brain-region Text
Activated Brain Regions in fMRI
Brain-region Text

Fig. 3. Illustration of the activated brain regions and corresponding brain-region texts.

Table 2. Ablation study and hyperparameters analysis on ABIDE and HCP datasets.

Variants	Accuracy	PCC
w/o Text-modal Connectome Supplement	73.9±2.8	0.348±0.005
w/o Image-modal Connectome Supplement	72.4±2.5	0.322±0.006
w/o Multi-modal Connectome Supplement	70.5±3.2	0.306±0.004
Prompt Depth = 2, Prompts Length = 2	73.6±3.1	0.345±0.005
Prompt Depth = 4, Prompts Length = 2	**75.4**±2.7	**0.369**±0.004
Prompt Depth = 8, Prompts Length = 2	74.2±3.0	0.364±0.004
Prompt Depth = 4, Prompts Length = 4	73.8±2.5	0.350±0.005
Prompt Depth = 4, Prompts Length = 8	72.9±2.9	0.332±0.003

3.2 Experimental Results

Comparison with State-of-the-arts Methods: The quantitative results are illustrated in Table 1. We find that the proposed fTSPL achieves 75.4% Accuracy, 82.5% AUROC, 81.9% Sensitivity, and 74.1% Specificity, as well as 0.156 MAE, 0.035 MSE, 0.369 PCC, and 0.101 R^2, which show that our pipeline outperforms existing GNN methods by a significant margin. Compared with advanced Transformer methods, our pipeline not only has lower training costs, but also achieves slightly better performance of 2.3% and 2.9% AUROC improvements.

Ablation Study: Next, we conduct ablation studies to verify the effectiveness of PMFCC. As shown in Table 2, we ablate text-modal, image-modal, and multi-modal connectome supplements in the multi-modal functional connectivity, respectively. We observe that removing the multi-modal connectome supplement reduces performance by 4.9% in Accuracy and 0.063 in PCC. Moreover, image-modal connectome supplement is more important than that of the text-modal, as it brings performance improvements of 3.4% Accuracy and 0.038 PCC. These results demonstrate that PMFCC is effective to improve brain analysis tasks.

Hyperparameters Analysis: In Table 2, we also analyze hyperparameters, including the depth and length of multi-modal prompts. It can be seen that the highest performance of 75.4% Accuracy and 0.369 PCC is achieved when the prompt depth and length are 4 and 2. Furthermore, insufficient prompt depth leads to a significant performance decline. In contrast, too large prompt length fails to improve performance, and even obtaining relatively poor results.

Text Generation: Fig. 3 visualizes the activated brain regions and corresponding brain-region texts of two patients' fMRI. These results illustrate that the proposed ABTG can generate accurate text descriptions for different fMRI.

4 Conclusion

In this paper, we propose a novel pipeline fTSPL for enhancing brain analysis. Concretely, fTSPL comprehensively considers the connectivity and activity of brain regions to generate instance-level texts to describe fMRI, and leverages the pre-trained V-L model and multi-modal prompts to construct the multi-modal functional connectivity. Experiments demonstrate that the proposed fTSPL pipeline achieves promising performance on multiple brain analysis tasks.

Acknowledgment. This work was supported by Hong Kong Research Grants Council (RGC) General Research Fund 14204321.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bannadabhavi, A., Lee, S., Deng, W., Ying, R., Li, X.: Community-aware transformer for autism prediction in fMRI connectome. In: Proc. Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 287–297 (2023)
2. Cui, H., Dai, W., Zhu, Y., Kan, X., Gu, A.A.C., Lukemire, J., Zhan, L., He, L., Guo, Y., Yang, C.: BrainGB: A benchmark for brain network analysis with graph neural networks. IEEE Trans. Med. Imag. **42**(2), 493–506 (2022)
3. Di Martino, A., Yan, C.G., Li, Q., Denio, E., Castellanos, F.X., Alaerts, K., Anderson, J.S., Assaf, M., Bookheimer, S.Y., Dapretto, M., et al.: The autism brain imaging data exchange: Towards a large-scale evaluation of the intrinsic brain architecture in autism. Mol. Psychiatr. **19**(6), 659–667 (2014)
4. Farahani, F.V., Karwowski, W., Lighthall, N.R.: Application of graph theory for identifying connectivity patterns in human brain networks: A systematic review. Front. Neurosci. **13**, 585 (2019)
5. Finn, E.S., Shen, X., Scheinost, D., Rosenberg, M.D., Huang, J., Chun, M.M., Papademetris, X., Constable, R.T.: Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. Nature Neurosci. **18**(11), 1664–1671 (2015)
6. Gao, J., Zhao, L., Zhong, T., Li, C., He, Z., Wei, Y., Zhang, S., Guo, L., Liu, T., Han, J., et al.: Prediction of cognitive scores by joint use of movie-watching fMRI connectivity and eye tracking via Attention-CensNet. In: Proc. Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 287–296 (2023)
7. Glasser, M.F., Sotiropoulos, S.N., Wilson, J.A., Coalson, T.S., Fischl, B., Andersson, J.L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J.R., et al.: The minimal preprocessing pipelines for the human connectome project. Neuroimage **80**, 105–124 (2013)

8. Hanik, M., Demirtaş, M.A., Gharsallaoui, M.A., Rekik, I.: Predicting cognitive scores with graph neural networks through sample selection learning. Brain Imag. Behav. **16**(3), 1123–1138 (2022)
9. Jegham, I., Rekik, I.: Meta-RegGNN: Predicting verbal and full-scale intelligence scores using graph neural networks and meta-learning. In: Proc. PRedictive Intelligence In MEdicine (PRIME). pp. 203–211 (2022)
10. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: Proc. International Conference on Machine Learning (ICML). pp. 4904–4916 (2021)
11. Kan, X., Dai, W., Cui, H., Zhang, Z., Guo, Y., Yang, C.: Brain network transformer. In: Proc. Neural Information Processing Systems (NeurIPS). vol. 35, pp. 25586–25599 (2022)
12. Kawahara, J., Brown, C.J., Miller, S.P., Booth, B.G., Chau, V., Grunau, R.E., Zwicker, J.G., Hamarneh, G.: BrainNetCNN: Convolutional neural networks for brain networks; towards predicting neurodevelopment. NeuroImage **146**, 1038–1049 (2017)
13. Khattak, M.U., Rasheed, H., Maaz, M., Khan, S., Khan, F.S.: MaPLe: Multi-modal prompt learning. In: Proc. IEEE/CVF Computer Vision and Pattern Recognition (CVPR). pp. 19113–19122 (2023)
14. Li, X., Zhou, Y., Dvornek, N., Zhang, M., Gao, S., Zhuang, J., Scheinost, D., Staib, L.H., Ventola, P., Duncan, J.S.: BrainGNN: Interpretable brain graph neural network for fMRI analysis. Med. Image Anal. **74**, 102233 (2021)
15. Liang, W., Zhang, K., Cao, P., Zhao, P., Liu, X., Yang, J., Zaiane, O.R.: Modeling alzheimers' disease progression from multi-task and self-supervised learning perspective with brain networks. In: Proc. Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 310–319 (2023)
16. Noback, C.R., Ruggiero, D.A., Strominger, N.L., Demarest, R.J.: The human nervous system: Structure and function. No. 744, Springer Science & Business Media (2005)
17. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. In: Proc. Neural Information Processing Systems (NeurIPS). vol. 32 (2019)
18. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: Proc. International Conference on Machine Learning (ICML). pp. 8748–8763 (2021)
19. Rolls, E.T., Huang, C.C., Lin, C.P., Feng, J., Joliot, M.: Automated anatomical labelling atlas 3. Neuroimage **206**, 116189 (2020)
20. Rosenberg, M.D., Finn, E.S., Scheinost, D., Papademetris, X., Shen, X., Constable, R.T., Chun, M.M.: A neuromarker of sustained attention from whole-brain functional connectivity. Nature Neurosci. **19**(1), 165–171 (2016)
21. Sporns, O.: Graph theory methods: Applications in brain networks. Dialogues Clin. Neurosci. **20**(2), 111–121 (2018)
22. Sutskever, I., Martens, J., Dahl, G., Hinton, G.: On the importance of initialization and momentum in deep learning. In: Proc. International Conference on Machine Learning (ICML). pp. 1139–1147 (2013)
23. Van Essen, D.C., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T.E., Bucholz, R., Chang, A., Chen, L., Corbetta, M., Curtiss, S.W., et al.: The human connectome project: A data acquisition perspective. Neuroimage **62**(4), 2222–2231 (2012)

24. Wang, F., Zhou, Y., Wang, S., Vardhanabhuti, V., Yu, L.: Multi-granularity cross-modal alignment for generalized medical visual representation learning. In: Proc. Neural Information Processing Systems (NeurIPS). pp. 33536–33549 (2022)
25. Wang, Q., Wu, M., Fang, Y., Wang, W., Qiao, L., Liu, M.: Modularity-constrained dynamic representation learning for interpretable brain disorder analysis with functional MRI. In: Proc. Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 46–56 (2023)
26. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: MedCLIP: Contrastive learning from unpaired medical images and text. In: Proc. Empirical Methods in Natural Language Processing (EMNLP). pp. 3876–3887 (2022)
27. Wasim, S.T., Naseer, M., Khan, S., Khan, F.S., Shah, M.: Vita-CLIP: Video and text adaptive CLIP via multimodal prompting. In: Proc. IEEE/CVF Computer Vision and Pattern Recognition (CVPR). pp. 23034–23044 (2023)
28. Zhang, S., Xu, Y., Usuyama, N., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., Wong, C., et al.: Large-scale domain-specific pretraining for biomedical vision-language processing. arXiv preprint arXiv:2303.00915 (2023)
29. Zhang, S., Chen, X., Shen, X., Ren, B., Yu, Z., Yang, H., Jiang, X., Shen, D., Zhou, Y., Zhang, X.Y.: A-GCL: Adversarial graph contrastive learning for fMRI analysis to diagnose neurodevelopmental disorders. Med. Image Anal. **90**, 102932 (2023)
30. Zhou, K., Yang, J., Loy, C., Liu, Z.: Conditional prompt learning for vision-language models. In: Proc. IEEE/CVF Computer Vision and Pattern Recognition (CVPR). pp. 16816–16825 (2022)