Spatio-temporal Pre-trained Foundation Model for Neural Decoding with Fine-grained Optimization

Ziyu Li[1], Zhiyuan Zhu[2], Yang Bai[3], Qing Li[1], and Xia Wu[1(✉)]

[1] Beijing Institute of Technology, Beijing 100081, China
wuxia@bit.edu.cn
[2] Chongqing University of Posts and Telecommunications, Chongqing 400065, China
[3] Beijing Century TAL Education Technology Co., Ltd., Beijing 102200, China

Abstract. Traditional neural decoding methods are heavily based on fully annotated brain data, which are both expensive to produce and scarce in availability. This limitation hinders the development of accurate and generalizable decoding models. Drawing inspiration from the success of foundational AI models in reducing dependency on annotated data in fields such as natural language processing, we introduce a novel foundation model that leverages the inherent spatiotemporal covariation of functional brain networks, which enables effective neural decoding with minimal annotation requirements. Our framework incorporates three key innovations: 1) A spatiotemporal importance-guided augmentation strategy is designed to capture the synergistic relationships between brain regions and their dynamic changes; 2) A progressive spatiotemporal-aware encoder is proposed to learn local-to-global brain interaction information; 3) A fine-grained consistency optimization technique is developed to enhance the representations of overall brain function. Evaluations of publicly available fMRI datasets demonstrate that our proposed framework not only achieves superior decoding performance, but also exhibits strong generalizability and reveals patterns of nervous activity. Our research advances brain representation learning and provides an innovative solution for universal neural decoding models.

Keywords: Neural decoding · Spatiotemporal · Self-supervised learning · fMRI.

1 Introduction

Neural decoding involves the systematic analysis and interpretation of neural activity patterns to infer brain states, cognitive processes, or behavioral intentions. This technique has significant applications in the elucidation of brain mechanisms, disease diagnosis, and the development of brain-inspired artificial intelligence [4]. Functional magnetic resonance imaging (fMRI) has garnered considerable attention in neural decoding due to its ability to capture cooperative relationships between regions of interest (ROIs) within the brain [2].

Recently, Graph Neural Networks (GNNs) have gained widespread application in fMRI-based neural decoding due to their robust capability to learn complex contextual representations and provide interpretability [6][16]. However, the superior performance of these models frequently depends on fully annotated fMRI datasets, which require expert knowledge for precise labeling. This dependency may undermine both the accuracy and generalizability of neural decoding models [11].

In contrast, graph foundation models leverage a substantial amount of unlabeled data through pre-training and fine-tuning techniques to adapt to complex graph tasks [9], thereby significantly reducing the cost of fMRI data annotation. During pre-training, general semantic information is extracted via GNNs-based self-supervised learning (GSSL), which facilitates downstream decoding tasks. Among them, contrastive-based GSSL has gained attention for its simplicity and efficiency [15], which predominantly focused on ROI connection analysis. However, this narrow-focused approach has a glaring limitation: it completely overlooks the crucial temporal dynamics within brain networks. Emerging research has provided compelling evidence that the temporal dependencies between ROIs play a pivotal role in neural activities. By neglecting this aspect, we are essentially missing out on a wealth of information that could potentially revolutionize our understanding of brain functions [12, **?**]. Numerous recent studies have shown that the incorporation of these temporal dependencies can lead to a substantial leap in self-supervised decoding performance [10]. Therefore, a novel graph foundation model for neural decoding is needed to fill this critical gap, by integrating the ignored temporal dynamics. The main challenges arise: i) How to design an efficient spatiotemporal representation extractor to capture more comprehensive and flexible brain representations, thereby improving the accuracy and reliability of neural decoding? and ii) What are the appropriate spatiotemporal augmented views for brain signals to provide a robust data foundation and enhance the model's capacity to interpret and analyze?

To address the aforementioned challenges, we propose an innovative SpatioTemporal Pre-Training Foundation model (STPTF) for neural decoding based on fMRI. STPTF employs spatiotemporal information of the brain, enabling pre-training with unlabeled fMRI data. With only a small amount of labeled fMRI data for fine-tuning, it can achieve substantial performance improvements in downstream decoding tasks. Specifically, STPTF introduces three key innovations: i) Strategy innovation through SpatioTemporal Importance Guided Augmentation (STIGA), STIGA comprises Spatial Centrality Augmentation (SCA) and Temporal Continuity Augmentation (TCA), which facilitate the model's deep analysis of brain representations from both spatial and temporal views; ii) Architectural innovation via Progressive SpatioTemporal-Aware encoder (PSTA), PSTA includes a simultaneously static and dynamic extractor, a continuous brain state calibration module, and a progressive temporal aggregation module, which help the model to thoroughly analyze global spatiotemporal features of brain activities; iii) Optimization innovation using Fine-Grained Consistency Optimization (FGCO), FGCO can effectively mitigate the suboptimal

representation problem caused by negative samples, which allows the model to fully and accurately interpret brain signals. We evaluate our proposed model, STPTF, on multiple medical imaging datasets, and the results demonstrate its state-of-the-art performance. To the best of our knowledge, this study represents one of the first attempts to integrate graph foundation models with spatiotemporal brain information, underscoring its significant potential in the field of neural decoding.

2 Method

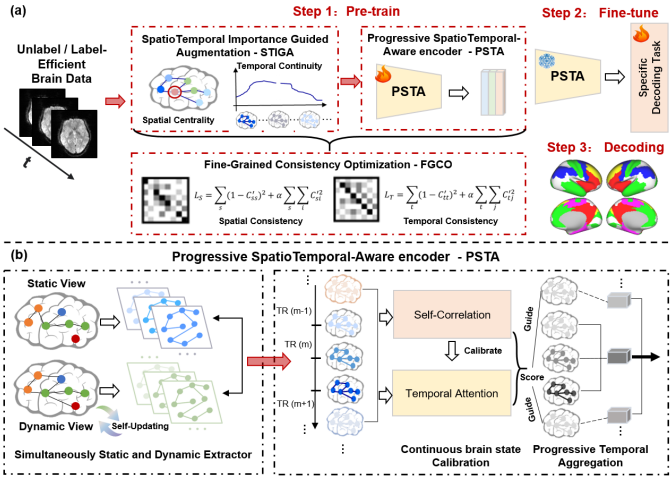

Fig. 1. (a) The overall framework of the proposed SpatioTemporal Pre-Training Foundation model (STPTF) for neural decoding. (b) Progressive SpatioTemporal-Aware encoder (PSTA).

2.1 Overall Framework

Implementing neural decoding based on our proposed STPTF is divided into three stages (Fig. 1): pre-training a general spatiotemporal encoder with unlabeled brain data, performing linear fine-tuning for various decoding tasks using the pre-trained encoder, and subsequently decoding different brain states based on the fine-tuned encoder. Specifically, STPTF integrates three core components to enhance fMRI-based neural decoding: i) Strategy innovation through STIGA; ii) Architectural innovation via PSTA (Fig. 1(b)); iii) Optimization innovation using FGCO.

2.2 SpatioTemporal Importance Guided Augmentation - STIGA

Spatial Centrality Augmentation - SCA Research shows that key ROIs, which serve as hubs in brain networks, frequently engage in interactions with other ROIs, thus facilitating neuronal communication and functional integration [5]. Based on these findings, we introduce SCA to generate an alternative view. In graph theory, nodes with high centrality are considered critical hubs [1]. We select degree centrality as a straightforward yet effective measure of node importance and adaptively assign masking probabilities according to the degree centrality of each ROI. Specifically, hubs are allocated higher masking probabilities, which aids in capturing the global semantic representation of brain networks.

Temporal Continuity Augmentation - TCA Considering the typically consistent ROI connections between adjacent timestamps and the concentration of semantic information, there could be redundant temporal information, indicating that changes in ROI connections exhibit continuity [13]. Based on this observation, we design the TCA for the other view. Specifically, we define the temporal change rate as the displacements between consecutive timestamps across the entire fMRI time series and adaptively assign mask probabilities according to the change rate at each timestamp. Timestamps with higher change rates are assigned greater mask probabilities, thereby compelling the model to learn the global temporal representation of the brain network.

2.3 Progressive SpatioTemporal-Aware encoder - PSTA

First, we extract the spatial representation of the original fMRI time series through a simultaneously static and dynamic extractor and then use GNN to capture the latent collaborative relationships among different ROIs. This approach allows us to simultaneously derive both static and dynamic brain representations, thereby reflecting both the integrity and the time-varying characteristics of the brain's functional network. The static and dynamic representations are complementary [8], and their joint learning provides a more comprehensive understanding of brain function. We incorporate automatic weight learning to capture the latent relationship between these two types of features, thus improving the adaptability and robustness of the model.

$$
\begin{aligned}
&Z_{Spatial} = \beta_1 Z_s + \beta_2 Z_d, \\
&\beta_c = \frac{exp(-\alpha_c)}{\sum_{k=1}^{2} exp(-\alpha_k)}, c \in 1, 2
\end{aligned} \tag{1}
$$

From a static perspective, we construct a static brain network by calculating partial correlation coefficients between pairs of ROIs after excluding potential confounding factors. We then employ GNNs to generate static representations of the brain denoted as Z_s. From a dynamic perspective, we adapt the adaptive

brain topology learning module [7], previously proposed, to learn brain networks that dynamically update with brain activities. By integrating GNN, the dynamic representations Z_d of the brain over time are automatically refined.

Assume that the original fMRI sequence is represented as X. For the learned spatial representations ($Z_{Spatial}$) of the brain, we further explore their progressive temporal representations. We introduce continuous brain state calibration to address the long-range error estimation challenge faced by typical attention mechanisms when establishing timestamp associations. Continuous brain state calibration adaptively determines the extent of the brain state transition information to be aggregated, without being constrained by the length of the fMRI sequence. Consequently, it facilitates the extraction of more refined long-term dependency information within the fMRI sequence and reveals contextual relationships of brain states within an appropriate temporal range.

$$\begin{aligned} &Z_{temporal} = Conv_2(Conv_V(Z_{Spatial}) \cdot CAM) + X, \\ &CAM = Softmax(Conv_1([CCM, TAW])), \\ &CCM = Softmax(Z_{Spatial}^T \cdot Z_{Spatial}), \\ &TAW = Softmax(Conv_Q(Z_{Spatial}) \cdot Conv_K(Z_{Spatial})^T) \end{aligned} \tag{2}$$

In order to improve the ability to model complex dynamic brain activities, we further design a progressive temporal aggregation module, which divides the temporal dimension of $Z_{temporal}$ into multiple subsegments. By integrating attention mechanism, progressive temporal aggregation computes the importance score for each timestamp within each sub-segment and derives the global representation of each sub-segment through weighted aggregation of these scores. Subsequently, progressive temporal aggregation calculates the importance score of each subsegment and progressively refines the global representation (Z_{st}) of the entire sequence, therefore effectively captures temporal dependencies from local to global levels.

2.4 Fine-Grained Consistency Optimization - FGCO

The spatially augmented view obtained through SCA is denoted as X_S, while the temporally augmented view obtained through TCA is denoted as X_T. By inputting X, X_S, and X_T into PSTA, the spatiotemporal features of each view are extracted. FGCO is designed to maximize the information correlation between the original view and its spatially/temporally augmented views by utilizing the cross-correlation matrix across different views. Specifically, this process is described below.

$$\begin{aligned} &L_{total} = L_S + L_T, \\ &L_S = \sum_s \left(1 - C_{ss}^{'}\right)^2 + \alpha \sum_s \sum_i C_{si}^{'2}, \\ &L_T = \sum_t \left(1 - C_{tt}^{'}\right)^2 + \alpha \sum_t \sum_j C_{tj}^{'2} \end{aligned} \tag{3}$$

This loss function encourages the augmented views to be closer to the original views while simultaneously minimizing redundancy within each view, thereby obtaining more informative and discriminative representations. The parameter α is utilized to balance the trade-off between consistency and redundancy. Through this optimization, a highly generalizable brain representation encoder named PSTA is obtained. Additionally, a linear layer is incorporated to fine-tune the entire network when applying it to downstream decoding tasks.

Table 1. Comparative Results (mean (std)). SM and SSM indicates graph supervised methods and self-supervised methods, respectively.

Dataset	Type	Method	Accuracy	AUC	Recall
Rest	SM	BrainGNN	66.60(3.63)	65.73(3.65)	**80.17(6.60)**
		STGCN	79.18(3.31)	78.55(2.68)	80.12(4.65)
	SSM	BrainGSL	68.10(1.19)	69.02(2.34)	67.56(4.83)
		GATE	65.93(2.26)	65.67(2.32)	68.63(2.25)
		STPTF	**83.37(1.77)**	**79.81(1.62)**	77.81(3.75)
Task	SM	BrainGNN	69.23(2.10)	82.22(0.99)	69.57(1.65)
		STGCN	68.59(3.63)	73.82(3.08)	55.14(5.28)
	SSM	BrainGSL	68.10(1.19)	69.02(2.34)	67.56(4.83)
		GATE	66.09(2.18)	65.82(2.20)	68.97(2.35)
		STPTF	**89.62(0.29)**	**90.62(0.35)**	**83.02(0.75)**

3 Experiments

3.1 Data Description and Implementation Details

The experimental data was from the publicly available fMRI dataset, Human Connectome Project (HCP). The resting state data from 1091 subjects and the task state data from 1007 subjects who participated fully in seven cognitive tasks are utilized to evaluate the effectiveness of STPTF. Despite originating from the same source, they represent distinct paradigms and cognitive states, providing a diverse evaluation. Specifically, the resting-state data examine gender differences in brain states, while the task-state data investigate variations in brain states across different cognitive tasks. To address computational complexity and information fitting challenges, we segment the entire brain into 22 major regions. These regions are derived by merging finer subdivisions from the existing brain atlas as [3].

We utilize a five-fold cross-validation to ensure the reliability of the decoding performance. Classification accuracy, area under the curve (AUC), and recall were used to evaluate decoding performance. In addition, a warm-up stage is incorporated into the training process to enhance stability. The learning rate begins at a small initial value, gradually increases to its maximum as training progresses, and subsequently follows a cosine decay schedule, leading to a gradual

decrease along a cosine curve. Compared with step decay, cosine decay offers a smoother adjustment, which can accelerate convergence and improve decoding performance.

Table 2. Ablation Results (mean (std)).

Dataset	Type	Method	Accuracy	AUC	Recall
Rest	STIGA	RSA + RTA	82.51(1.31)	79.50(1.29)	76.91(4.42)
		w/o / TCA	81.18(1.41)	78.29(1.14)	68.41(4.44)
		w/o / SCA	81.61(1.75)	79.17(2.02)	76.31(2.75)
	PSTA	GCN	58.62(2.02)	56.37(2.38)	39.46(4.90)
		LSTM	77.73(1.69)	75.61(0.15)	75.03(4.78)
		STGCN	77.14(3.81)	75.06(4.99)	73.91(5.54)
		STPTF	**83.37(1.77)**	**79.81(1.62)**	**77.81(3.75)**
Task	STIGA	RSA + RTA	89.15(0.44)	90.18(0.24)	82.39(0.45)
		w/o / TAM	89.17(3.57)	90.15(0.22)	82.29(0.32)
		w/o / SCM	89.25(0.50)	90.15(0.79)	82.85(0.60)
	PSTA	GCN	61.63(0.60)	72.73(0.34)	52.38(0.65)
		LSTM	88.93(0.28)	89.38(0.45)	81.34(0.28)
		STGCN	80.09(0.83)	80.22(0.69)	64.93(1.92)
		STPTF	**89.62(0.29)**	**90.62(0.35)**	**83.02(0.75)**

3.2 Comparative Results

To assess the advancement of the proposed model, we conduct a comparative analysis against several state-of-the-art graph supervised methods: including BrainGNN [6] and STGCN [3], which share the same goal as our work. They specialize in spatial and spatiotemporal brain representation mining, respectively; graph self-supervised methods based on brain spatial pattern, namely BrainGSL[14] and GATE [12]. BrainGSL is a representative example of the generative-based approach to the diagnosis of brain diseases. GATE is one of the pioneering contrastive-based methods for disease diagnosis. Table 1 illustrates the decoding performance of STPTF in two datasets. It is evident that most graph self-supervised methods exhibit a substantial performance gap compared to spatiotemporal supervised methods. Nevertheless, the proposed STPTF demonstrates comparable, and in some cases superior, decoding capabilities relative to state-of-the-art supervised spatio-temporal decoding methods. Moreover, STPTF significantly outperforms existing graph self-supervised decoding approaches, it may stem from the neglect of temporal dynamics and the limitations inherent in the backbone (GNNs), which restrict the effectiveness of existing graph self-supervised methods.

Table 3. The robustness results under different sample sizes (mean (std), %).

	Rest			Task		
%	Accuracy	AUC	Recall	Accuracy	AUC	Recall
20	78.45(0.84)	74.42(1.79)	71.71(6.24)	86.81(0.16)	88.18(0.32)	78.92(0.55)
40	81.41(1.04)	77.81(0.57)	69.95(2.08)	88.90(0.10)	90.48(0.09)	82.92(0.15)
60	81.27(0.37)	78.51(1.03)	72.70(2.07)	88.43(0.11)	90.12(0.14)	82.28(0.24)
80	83.37(1.77)	79.81(1.62)	77.81(3.75)	89.62(0.29)	90.62(0.35)	83.02(0.75)

3.3 Ablation Results

To further validate the effectiveness of our STPTF, we perform comprehensive ablation experiments on the proposed modules, including STIGA and PSTA. For STIGA, we replace the proposed spatial and temporal augmented strategies (SCA and TCA) with random spatial and temporal augmention (RSA and RTA). For PSTA, we replace it with only spatial encoders and traditional spatiotemporal encoders. Quantitative results (Table 2) show that STIGA and PSTA have the ability to effectively guide the model to focus on key information, map features to a more general representation space, and thus enhance the performance of downstream decoding tasks. It should be noted that, although the improvements brought about by STIGA may seem modest, they demonstrate consistency across various metrics and datasets, highlighting the robustness of STIGA. Furthermore, even minor improvements can bring it closer to practical application scenarios.

3.4 Robustness and Transferability

Robustness Considering that the ultimate goal of STPTF is to facilitate its application to different downstream decoding tasks and achieve efficient decoding even with a small amount of labeled fMRI data, we perform multiple experiments to fine-tune the sample size. The results show that even with a small sample size (20%), STPTF achieves better results than previous GSSL-based brain decoding methods and is comparable to supervised methods (Table 3).

Transferability Current GSSL decoding algorithms are intra-domain, meaning that pre-training and fine-tuning are based on the same dataset. To validate that the proposed STPTF has strong transferability and adapts well to different downstream decoding tasks, we further explore its performance in inter-domain scenarios. For example, we pre-train on task-related data and then fine-tune for resting-state data decoding, and vice versa. The results (Table 4) show that STPTF still achieves good performance when transferring between different datasets, similar to intra-domain results. This indicates that STPTF can precisely extract generalizable knowledge from pre-trained datasets.

Table 4. The transferability results on different fMRI sequence length, Task → Rest represents training on task-related data while fine-tuning on resting data, and vice versa (mean (std)).

Task→Rest			Rest→Task		
Accuracy	AUC	Recall	Accuracy	AUC	Recall
83.37(1.86)	80.01(1.45)	77.72(4.88)	88.18(0.49)	88.42(0.40)	79.40(0.63)

4 Conclusion

In this study, we introduce an innovative fine-grained spatio-temporal pre-training and decoding framework (STPTF) based on GSSL. This framework offers a novel technical approach to gain deeper insights into the mechanisms of brain state transitions. Specifically, STPTF can learn universal representations of brain states from unlabeled data by leveraging STIGA strategy, PSTA miner, and multi-view FGCO. Experimental results in multiple public datasets demonstrate that STPTF achieves significant improvements in unsupervised brain decoding, likely due to its incorporation of self-supervised temporal information. Furthermore, STPTF's effectiveness in downstream decoding tasks with limited data and across different datasets highlights its robustness and transferability. With broader data-sharing initiatives, STPTF has the potential to be applied to larger and more diverse decoding datasets, further validating its generalization capabilities.

Acknowledgments. This work was supported by the China Postdoctoral Science Foundation (Grant No. 2024M764143), the National Science Fund for Distinguished Young Scholars of China (Grant No. 62325601), the MoE Key Laboratory of Brain-inspired Intelligent Perception and Cognition, University of Science and Technology of China (Grant No. 2421001), and the Young Scientists Fund of the National Natural Science Foundation of China (Grant No. 62406049).

Disclosure of Interests. The authors have no competing interests in declaring that they are relevant to the content of this paper.

References

1. Borgatti, S.P., Everett, M.G.: A graph-theoretic perspective on centrality. Social Networks **28**(4), 466–484 (2006)
2. Cocchi, L., Yang, Z., Zalesky, A., Stelzer, J., Hearne, L.J., Gollo, L.L., Mattingley, J.B.: Neural decoding of visual stimuli varies with fluctuations in global network efficiency. Human Brain Mapping **38**(6), 3069–3080 (2017)
3. Gadgil, S., Zhao, Q., Pfefferbaum, A., Sullivan, E.V., Adeli, E., Pohl, K.M.: Spatio-temporal graph convolution for resting-state fMRI analysis. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VII 23. pp. 528–538. Springer (2020)

4. Greene, A.S., Horien, C., Barson, D., Scheinost, D., Constable, R.T.: Why is everyone talking about brain state? Trends in Neurosciences **46**(7), 508–524 (2023)
5. Van den Heuvel, M.P., Sporns, O.: Network hubs in the human brain. Trends in Cognitive Sciences **17**(12), 683–696 (2013)
6. 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. Medical Image Analysis **74**, 102233 (2021)
7. Li, Z., Li, Q., Zhu, Z., Hu, Z., Wu, X.: Multi-scale spatio-temporal fusion with adaptive brain topology learning for fMRI based neural decoding. IEEE Journal of Biomedical and Health Informatics **28**(1), 262–272 (2024)
8. Lin, X., Kong, W., Li, J., Shao, X., Jiang, C., Yu, R., Li, X., Hu, B.: Aberrant static and dynamic functional brain network in depression based on EEG source localization. IEEE/ACM Transactions on Computational Biology and Bioinformatics **20**(3), 1876–1889 (2022)
9. Liu, J., Yang, C., Lu, Z., Chen, J., Li, Y., Zhang, M., Bai, T., Fang, Y., Sun, L., Yu, P.S., Shi, C.: Towards graph foundation models: A survey and beyond. ArXiv **abs/2310.11829** (2023), https://api.semanticscholar.org/CorpusID:264288909
10. Luppi, A.I., Craig, M.M., Pappas, I., Finoia, P., Williams, G.B., Allanson, J., Pickard, J.D., Owen, A.M., Naci, L., Menon, D.K., Stamatakis, E.A.: Consciousness-specific dynamic interactions of brain integration and functional diversity. Nature Communications **10**(1), 4616 (2019)
11. Lutnick, B., Ginley, B., Govind, D., McGarry, S.D., LaViolette, P.S., Yacoub, R., Jain, S., Tomaszewski, J.E., Jen, K.Y., Sarder, P.: An integrated iterative annotation technique for easing neural network training in medical image analysis. Nature Machine Intelligence **1**(2), 112–119 (2019)
12. Peng, L., Wang, N., Xu, J., Zhu, X., Li, X.: GATE: Graph CCA for temporal self-supervised learning for label-efficient fMRI analysis. IEEE Transactions on Medical Imaging **42**(2), 391–402 (2022)
13. Wang, M., Huang, J., Liu, M., Zhang, D.: Modeling dynamic characteristics of brain functional connectivity networks using resting-state functional MRI. Medical Image Analysis **71**, 102063 (2021)
14. Wen, G., Cao, P., Liu, L., Yang, J., Zhang, X., Wang, F., Zaiane, O.R.: Graph self-supervised learning with application to brain networks analysis. IEEE Journal of Biomedical and Health Informatics **27**(8), 4154–4165 (2023)
15. Zhang, K., Wen, Q., Zhang, C., Cai, R., Jin, M., Liu, Y., Zhang, J., Liang, Y., Pang, G., Song, D., Pan, S.: Self-supervised learning for time series analysis: Taxonomy, progress, and prospects. arXiv preprint arXiv:2306.10125 (2023)
16. Zhang, Y., Tetrel, L., Thirion, B., Bellec, P.: Functional annotation of human cognitive states using deep graph convolution. NeuroImage **231**, 117847 (2021)