BRAINTAL: A TARGET-ADAPTIVE BRAIN TOKEN LEARNER FOR CROSS-SITE BRAIN NETWORK ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-site brain network analysis provides an important tool for characterizing functional connectivity patterns across different medical centers and has attracted increasing attention in brain disease diagnosis. However, existing methods often struggle to adapt to target-site distribution shifts under source-free and label-free settings. Moreover, they usually overlook task-specific semantic differences within the same brain network, which limits their predictive ability. To address these limitations, we propose BrainTAL, a target-adaptive brain token learning framework for cross-site brain network analysis. Specifically, we first construct compact brain network instructions, where salient functional connections are introduced as neurobiological priors for language-guided modeling. Second, we design task-conditioned brain tokens, enabling the model to adaptively capture task-relevant functional connectivity patterns from the same brain network across different neuroimaging tasks. Finally, we propose a brain token adaptive calibration strategy, which performs lightweight adaptation of the source-trained model using unlabeled target-site data, thereby alleviating cross-site distribution shifts and improving prediction robustness on the target site. Experimental results demonstrate that BrainTAL consistently improves disease recognition performance in source-free cross-site scenarios and effectively captures task-relevant functional connectivity patterns, validating its effectiveness and interpretability.

1 INTRODUCTION

Accurate identification and early screening of brain diseases are important for clinical intervention and prognostic assessment (Gu et al., 2026; Yang et al., 2023). Functional brain networks derived from resting-state functional magnetic resonance imaging (rs-fMRI) characterize interregional connectivity and support the identification of disease-related abnormalities (Yin et al., 2026). However, differences in scanners, acquisition protocols, and subject populations introduce site shifts that hinder the generalization of source-trained models across imaging centers (Guan & Liu, 2021; Pei et al., 2025). Privacy and data governance constraints further restrict raw source-data sharing, while target-site data often remain unlabeled (Zhang & Xu, 2026; Gijsen et al., 2026).

Cross-site methods mitigate distribution shifts through data harmonization that corrects site-specific feature biases (Lv et al., 2017), domain generalization that learns site-invariant representations for unseen centers (Zhi et al., 2024; Zhou et al., 2022), and domain adaptation that aligns source and target distributions using target-site data (Song et al., 2024; Li et al., 2026). At deployment, access to raw source-site data can be restricted by privacy, ethical, and institutional requirements (Kanakasabapathy et al., 2021; Wang et al., 2023). This motivates adaptation from a source-trained model using unlabeled target data when raw source data are unavailable.

Source-free domain adaptation addresses this setting by updating a source-trained model with unlabeled target data (Kundu et al., 2020; Wu et al., 2023; Lee et al., 2023; Li et al., 2025c). Existing source-free brain network methods typically use graph-level representations with pseudo-labeling or parameter-efficient adaptation to mitigate site shifts. Their focus on a single objective, however, limits the extraction of task-specific information from the same network. Diagnosis, sex prediction, and age estimation can emphasize different connectivity patterns. This motivates task-conditioned

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

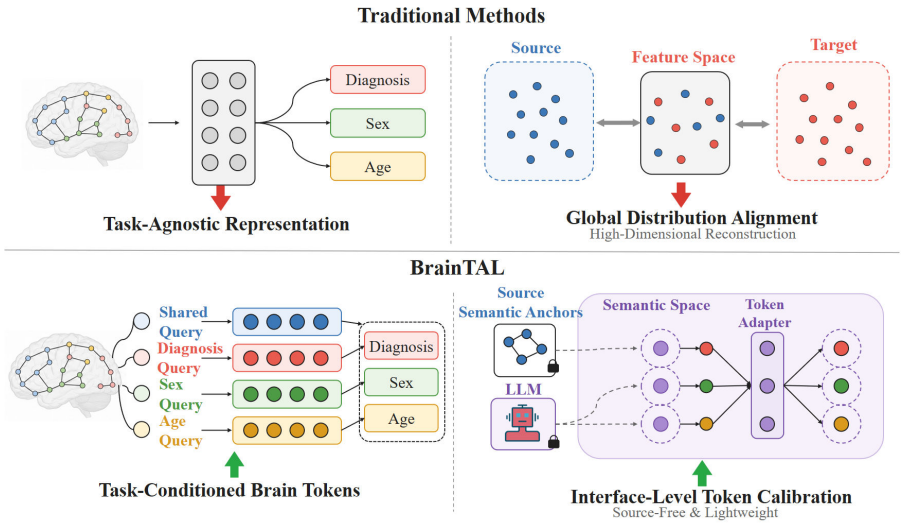

Figure 1: Comparison of traditional cross-site brain network paradigms and BrainTAL.

graph readouts, in which task semantics guide the aggregation of ROI features before they enter the language model.

Language-guided brain network models connect functional connectivity, task descriptions, and prediction targets through multi-task instruction tuning (Yang et al., 2024; Li et al., 2025a). FCN-LLM (Hu et al., 2026) projects multi-scale graph representations into the LLM semantic space for multi-task prediction, while BLEG (Dong et al., 2026) uses graph-derived descriptions and lightweight instruction tuning. These approaches mainly establish brain–semantic alignment through supervised source-site training, with generalization typically assessed through zero-shot evaluation on unseen data. Site shifts may nevertheless affect the source-learned alignment. The remaining challenge is to calibrate task-conditioned brain representations using unlabeled target data while keeping the source-trained graph encoder and language model fixed.

To address these challenges, we propose BrainTAL, a target-adaptive brain token learning framework for cross-site brain network analysis (Figure 1). BrainTAL uses compact descriptions of salient functional connections as textual priors and learns task-conditioned brain tokens through source-site multi-task training. Its central design couples task-dependent graph readouts with source-free calibration at the brain–language interface, allowing target-site adaptation while keeping the source-trained graph encoder, task queries, and language model fixed. Our contributions are:

- We develop a task-conditioned graph-to-token mapping. Shared queries and task-specific offsets determine how ROI features are aggregated, enabling the same brain network to produce task-dependent representations before language-model processing.
- We perform source-free target adaptation at the brain-token interface, updating only a residual adapter through confidence-guided momentum consistency. The adapter is calibrated using diagnosis predictions and shared across tasks at inference.
- Experiments on ABIDE and ADHD demonstrate improved cross-site diagnosis, sex classification, and age prediction. Ablations and qualitative visualizations support the task-conditioned design, with additional evaluation using Qwen2.5-3B and Qwen2.5-7B.

2 Related Work

2.1 Cross-Site Brain Network Analysis

Cross-site brain network analysis aims to improve the generalization ability of diagnostic models across imaging centers, but functional connectivity patterns are often affected by differences in scanners, acquisition protocols, and subject populations Su et al. (2021); Guan et al. (2021). In recent

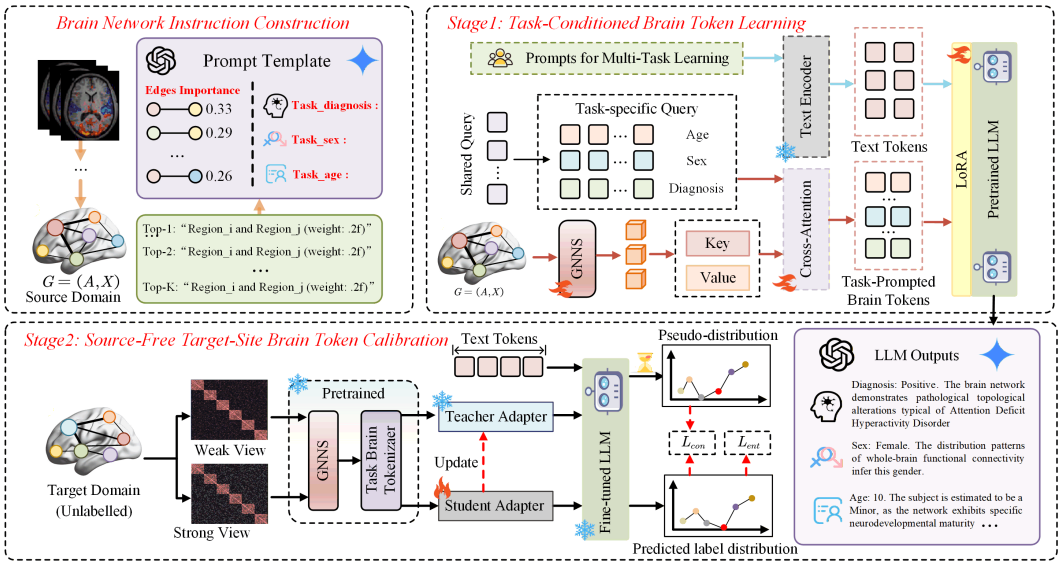

Figure 2: The proposed overall framework of BrainTAL.

years, source-free domain adaptation has been introduced into neuroimaging analysis to reduce the dependence on source-site data Li et al. (2025b). For example, SFHarmony Dinsdale et al. (2023) performs site harmonization through Gaussian mixture modeling and Bhattacharyya distance minimization. SPDIM Li et al. (2025b) learns distribution parameters under symmetric positive definite manifold constraints to alleviate site shifts in functional connectivity matrices. SCDA Fang et al. (2025) and SSTL Hu et al. (2023) further exploit target-domain information through multi-view feature enhancement and semi-supervised transfer learning, respectively. Despite these advances, existing methods primarily perform graph-level representation harmonization for a single prediction objective, without explicitly modeling task-specific semantics within the same brain network.

2.2 Language-Guided Brain Network Modeling

Large language models have demonstrated strong capabilities in semantic understanding and reasoning, promoting the application of language-guided modeling in brain network analysis Hui et al. (2024); Zhang et al. (2026). Recent studies have begun to align functional brain networks with the language semantic space Liu et al. (2025). For example, BrainEC-LLM Xiong et al. (2026) introduces a multi-scale decomposition mixture mechanism to stably estimate effective brain connectivity and enhance brain network modeling. IEBGL Duan et al. (2026) combines LITR with LLM reasoning to model brain connectivity relationships in a fine-grained manner, improving model interpretability. Other studies further map functional connectivity networks into the semantic space of language models through multi-task instruction tuning or graph-language representation alignment, thereby enhancing the model's understanding of neuroimaging tasks. Although existing methods demonstrate the potential of LLMs for brain network analysis, they largely focus on source-domain semantic enhancement or supervised alignment, overlooking adaptation to unlabeled target sites and explicit modeling of task-relevant functional connectivity. BrainTAL addresses these limitations through task-conditioned brain tokens and adaptive brain-token calibration, thereby improving cross-site robustness.

3 Method

Figure 2 shows two stages. Stage I uses labeled source-site graphs to train diagnosis, sex, and age prediction. Stage II receives the Stage-I checkpoint and unlabeled graphs from a held-out target site; no source graph or target label enters this optimization stage. Diagnosis predictions define the target adaptation objective. The resulting adapter is shared by all three tasks at inference, so sex and age receive no task-specific target optimization.

3.1 BRAIN INSTRUCTION CONSTRUCTION

For each subject, we construct a functional connectivity network based on the AAL90 brain atlas. Specifically, given the preprocessed resting-state fMRI time series $S_i \in R^{T\times R}$, where T denotes the number of time points and $R = 90$ denotes the number of brain regions, we calculate the Pearson correlation coefficients between the BOLD signals of different brain regions to obtain the functional connectivity matrix $A_i \in R^{R\times R}$. In this paper, we regard A_i as the adjacency matrix of the brain network and use the connectivity pattern between each brain region and all other regions as ROI-level node features, resulting in the brain graph $G_i = (A_i, X_i)$.

BrainTAL uses two views of each subject's functional connectivity network. The graph encoder receives the full network, while the language-model instruction contains a compact description of selected observed connections. We select ROI pairs from the upper triangle of the connectivity matrix according to absolute Pearson correlation:

$$\varepsilon_i^K = \underset{1\le p<q\le R}{TopK} \left(|A_i\,(p,q)|\right), \tag{1}$$

where, ε_i^K denotes the selected-pair index set, with $K = 3$. Each pair is mapped to its AAL90 anatomical names, and its original signed Pearson correlation is retained in the instruction. The absolute value is used only for connection selection. These data-derived descriptions constitute the brain-network priors supplied to the language model; the correlation values are subject-level input measurements rather than model-estimated, disease-specific edge-importance scores.

For task construction, we organize the same brain network into multi-task instruction samples, covering three types of neuroimaging prediction tasks: disease diagnosis, sex prediction, and age estimation. For each brain network G_i, we construct the corresponding task instruction P_i^m according to the task type $m \in diagnosis, sex, age$. To enable the model to not only learn the mapping between brain networks and labels during training, but also understand how to organize and interpret the same brain network information under different task semantics, each instruction is constructed with three components: task description, brain network prior, and target answer format:

$$I_i^m = (G_i, P_i^m, a_i^m), \tag{2}$$

where a_i^m denotes the target answer corresponding to task m. In this way, we uniformly construct the source-site data into a multi-task instruction training set, which is used as the input for LLM instruction tuning.

3.2 TASK-CONDITIONED BRAIN TOKEN LEARNING

Different neuroimaging tasks emphasize distinct semantics within the same functional brain network: pathology-related discriminative connectivity for disease diagnosis, stable population differences for sex prediction, and continuous developmental and aging patterns for age estimation. We therefore propose a task-conditioned brain tokenization module that combines shared and task-specific queries to extract task-relevant structural information from the same brain network.

Given a brain network G_i, we first use a brain graph encoder f to extract ROI-level brain network representations:

$$H_i = f\,(G_i)\,,\ H_i \in R^{R\times d}, \tag{3}$$

where, R denotes the number of ROIs and d is the ROI representation dimension. The encoder f consists of two graph convolutional layers. Each row in H_i corresponds to the structural representation of one ROI, which characterizes the contextual information of the brain region in the functional connectivity network. Subsequently, we introduce a task-conditioned query module to extract a small number of task-conditioned brain tokens from ROI-level brain representations, enabling the model to selectively aggregate relevant brain region information under different task semantics.

Specifically, we construct shared queries Q_s and task-specific offsets $\triangle Q_m$. The shared queries are used to model neural functional patterns that are consistent across tasks, while the task-specific offsets encode the semantic differences of task m. The query representation corresponding to task m is defined as:

$$Q_m = Q_s + \triangle Q_m, \tag{4}$$

where Q_s provides the shared query basis across tasks, and $\triangle Q_m$ enables the queries to adjust their attention to brain region information according to the current task. With this design, the same brain network can generate different task-conditioned representations under different tasks.

Subsequently, the task query Q_m interacts with the ROI-level brain representation H_i through cross-attention. In this process, the task query serves as the query, while the ROI-level brain representation serves as the key and value, aggregating task-relevant information from the brain network:

$$Z_i^m = CrossAttn\left(Q_m, H_i, H_i\right), \tag{5}$$

Here, $Z_i^m \in \mathbb{R}^{L\times d}$ contains L brain tokens. Changing the task changes Q_m before the graph representation reaches the language model, so the same ROI features H_i can yield different token sequences. The shared queries provide a common readout basis, whereas task-specific offsets alter which ROI features are aggregated. Without the offsets, all tasks use the same graph queries and task differences must be handled downstream by the language model. The tokens are projected to the language-model embedding space and inserted into the instruction:

$$U_i^m = \left[E\left(\hat{P}_i^m\right), \hat{Z}_i^m, E\left(\check{P}_i^m\right)\right], \tag{6}$$

where $\hat{Z}_i^m = Z_i^m W_z$, W_z denotes the projection matrix from brain tokens to the LLM embedding space, E denotes the text embedding layer of the language model, $\hat{P}_i^m$ and $\check{P}_i^m$ denote the text segments before and after $\hat{Z}_i^m$, respectively.

During the source-site training stage, BrainTAL uses graph-instruction-answer triplets as supervision signals to learn the correspondence between task-conditioned brain tokens and different task outputs. In this process, only the brain-network-side parameters and lightweight language adaptation parameters are updated, while the backbone of the language model remains frozen. Given a source-domain instruction sample, the training objective is defined as:

$$L_{S1} = -\sum \log p_\theta\left(a_i^m | G_i, P_i^m\right), \tag{7}$$

where p_θ denotes conditional probability. During Stage I, the LLM backbone remains frozen, while LoRA modules inserted into its attention projections are trained jointly with the brain encoder and task-conditioned tokenization module.

3.3 Target-Site Brain Token Adaptation

A site shift changes the ROI features and hence the brain tokens delivered to the source-trained language model. We keep the graph encoder, task queries, and language-model parameters fixed, and update only a bottleneck residual adapter on the target-site brain tokens. This confines target optimization to the interface used by the frozen source-trained predictor.

3.3.1 Brain-Token Interface Calibration

Given the i-th target brain network G_i^{tar} and the instruction P_i^d associated with the disease task, the frozen brain encoder E_ψ and task-conditioned brain tokenization module T_ϕ first generate the target-domain disease-specific brain tokens:

$$Z_i^{\mathrm{tar},d} = T_\phi\left(E_\psi\left(G_i^{\mathrm{tar}}\right), Q_d\right), \tag{8}$$

To compensate for the shift of target-domain brain tokens relative to the source-domain semantic interface, we introduce a bottleneck residual adapter before feeding the brain tokens into the language model:

$$\tilde{Z}_i^{\mathrm{tar},d} = Z_i^{\mathrm{tar},d} + W_{up}\sigma\left(W_{down} LN\left(Z_i^{\mathrm{tar},d}\right)\right), \tag{9}$$

where W_{down} and W_{up} denote the down-projection and up-projection matrices, respectively, $\sigma\left(\cdot\right)$ is a nonlinear activation function, and $LN\left(\cdot\right)$ denotes layer normalization.

3.3.2 CONFIDENCE-GUIDED MOMENTUM CONSISTENCY CALIBRATION

To obtain reliable calibration signals from unlabeled target-site data, we construct weakly and strongly augmented views of the same brain network and feed them into the teacher and student branches, respectively. The resulting disease-diagnosis distributions are formulated as:

$$p_{i,w}^{t} = softmax\left(F_t\left(\Lambda_w\left(G_i^{\mathrm{tar}}\right), P_i^{d}\right)\right), \tag{10}$$

$$p_{i,s}^{s} = softmax\left(F_s\left(\Lambda_s\left(G_i^{\mathrm{tar}}\right), P_i^{d}\right)\right), \tag{11}$$

where Λ_w and Λ_s denote weak noise perturbation and strong edge-dropout augmentation, respectively, and F_t and F_s represent the teacher and student branches. The weakly augmented view is fed into the teacher model, whereas the strongly augmented view is fed into the student model. No gradients are backpropagated through the teacher branch. Instead, its parameters are updated using an exponential moving average of the student parameters:

$$\vartheta_t^{(k+1)} = u\,\vartheta_t^{(k)} + (1-u)\,\theta_s^{(k+1)}, \tag{12}$$

where, u is the EMA momentum and k indexes adaptation steps. The teacher is updated after the corresponding student update. Furthermore, to suppress error accumulation caused by low-confidence pseudo-predictions, we filter target samples according to the diagnostic confidence of the teacher branch on the weakly augmented views. The confidence-gating variable for the i-th target sample is defined as:

$$w_i^{d} = I\left[\max_c p_{i,w}^{t,c} \geq \tau\right], \tag{13}$$

where $p_{i,w}^{t,c}$ denotes the probability assigned by the teacher model that the i-th target sample belongs to class c, $I[\cdot]$ is the indicator function, and τ is the confidence threshold. Only high-confidence samples contribute to the teacher–student consistency optimization:

$$L_{\mathrm{con}} = \frac{\sum_{i=1}^{N_{\mathrm{tar}}} w_i^{d} D_{\mathrm{KL}}\left(\mathrm{sg}[p_{i,w}^{t}] \| p_{i,s}^{s}\right)}{\sum_{i=1}^{N_{\mathrm{tar}}} w_i^{d} + \epsilon}, \tag{14}$$

where, N_{tar} denotes the number of target samples included in the loss, sg denotes stop-gradient, and $\epsilon > 0$ stabilizes normalization. The binary mask w_i^{d} is computed from detached teacher predictions and treated as constant during student optimization. Neither the mask nor the normalization term receives gradients; the teacher parameters are updated only by the EMA rule.

To further reduce the uncertainty of target-domain diagnostic predictions, the overall optimization objective in the second stage is defined as:

$$L_{S2} = L_{con} + \lambda L_{ent}, \tag{15}$$

where L_{ent} denotes the mean predictive entropy of the student's diagnostic probability distribution, and λ is its weighting coefficient. Notably, target-site adaptation is driven solely by the disease-diagnosis task. Sex classification and age prediction provide no direct target-domain optimization signals. Instead, during inference, they use the same brain-token adapter calibrated through the diagnosis task.

4 EXPERIMENTS AND RESULTS

4.1 DATASETS

We evaluate the proposed method on two public multi-center resting-state fMRI brain disease datasets, ABIDE Di Martino et al. (2014) and ADHD Shi et al. (2025). The ABIDE dataset is used for autism spectrum disorder analysis, with raw data collected from 17 international imaging centers. The ADHD dataset is used for attention deficit hyperactivity disorder analysis, with raw data collected from 8 international imaging centers. For each subject, we construct a functional brain network based on the AAL90 atlas as the model input. To ensure the stability of cross-center evaluation and reduce the influence of sites with extremely small sample sizes, we select the five sites with the largest sample sizes from each dataset for experiments. Specifically, for ABIDE, we select five sites, including UM, USM, UCLA, LEUVEN, and NYU, resulting in 470 subjects, including 218 ASD patients and 252 healthy controls. For ADHD, we select five sites, including KKI, OHSU, NYU, NeuroImage, and Peking, resulting in 782 subjects, including 353 ADHD patients and 429 healthy controls. Based on three tasks, we construct three task instructions for each subject. Therefore, ABIDE and ADHD contain 1,410 and 2,346 multi-task instruction samples, respectively.

Table 1: Performance comparison on ABIDE. Results are reported as mean ± standard deviation across five target sites. Classification metrics and PCC are in %, while MAE and RMSE are in years. Best results are in bold.

Methods	Diagnosis Classification				Sex Classification			Age Prediction		
	ACC	AUC	F1	MCC	ACC	F1	MCC	MAE	RMSE	PCC
GCN	57.09±2.20	61.34±2.06	56.52±3.07	22.18±4.76	79.38±10.40	55.22±5.95	10.26±9.58	16.30±4.49	17.09±4.82	12.50±8.41
GAT	57.16±4.04	59.94±4.79	53.28±5.95	20.78±5.79	79.76±10.69	51.80±3.92	6.95±6.42	16.37±3.66	17.18±3.95	10.21±8.20
GraphTrans	59.02±2.04	62.01±4.56	59.52±2.64	20.01±3.09	81.14±9.26	52.25±4.86	7.57±8.75	15.45±4.15	16.22±4.46	21.39±5.58
BrainGNN	57.74±5.21	58.45±2.68	55.22±6.85	14.85±6.80	83.77±9.55	90.75±5.50	9.04±7.65	13.95±6.52	17.53±8.50	11.15±6.86
BrainNPT	62.28±1.82	63.28±3.60	57.52±14.39	21.83±4.01	83.62±6.65	90.95±3.86	14.06±11.30	13.15±5.76	17.35±7.11	26.19±6.54
BLEG	62.55±3.50	62.39±2.81	61.34±6.78	22.21±6.23	83.42±9.21	90.25±4.96	16.28±8.33	7.86±3.48	8.32±3.49	27.94±9.56
BrainTAL-S1	63.93±3.91	63.56±3.38	63.70±7.11	27.72±5.58	85.40±8.46	91.82±4.71	17.11±10.70	5.79±2.37	7.54±2.69	28.71±10.53
DANN	66.23±2.37	65.46±2.45	59.28±11.59	26.44±6.69	86.07±8.79	92.23±4.97	14.28±11.55	11.47±6.97	15.06±8.84	29.87±6.36
SHOT	62.68±2.51	62.53±3.44	59.99±3.60	22.43±6.72	80.61±11.18	88.33±7.28	8.69±8.14	6.57±2.05	8.03±2.73	23.83±5.94
SFHarmony	61.60±5.20	65.63±4.09	58.38±9.61	26.42±11.32	85.03±10.01	91.50±5.79	9.63±10.33	4.98±2.26	6.52±2.82	32.29±3.63
GraphATA	64.29±1.59	65.08±4.75	62.65±5.93	13.71±8.09	81.73±17.13	89.95±15.64	12.63±3.55	5.33±5.54	7.49±5.38	11.20±7.24
BrainTAL-3B	66.65±2.54	66.29±3.16	67.96±5.73	29.20±7.93	87.60±8.05	93.15±4.57	20.20±13.26	4.88±2.20	6.40±2.82	30.98±8.50
BrainTAL-7B	**67.62±2.41**	**67.45±2.89**	**69.12±5.20**	**31.35±7.45**	**88.45±7.62**	**93.88±4.15**	**22.15±12.40**	**4.62±2.08**	**6.11±2.65**	**32.85±8.12**

Table 2: Performance comparison on ADHD. Results are reported as mean ± standard deviation across five target sites. Classification metrics and PCC are in %, while MAE and RMSE are in years. Best results are in bold.

Methods	Diagnosis Classification				Sex Classification			Age Prediction		
	ACC	AUC	F1	MCC	ACC	F1	MCC	MAE	RMSE	PCC
GCN	52.12±3.22	52.99±3.07	51.20±2.87	8.55±7.76	55.68±9.28	51.52±7.24	7.64±6.01	11.80±3.25	12.08±3.11	2.37±12.10
GAT	52.53±3.37	54.19±2.38	49.66±3.99	4.94±4.45	56.20±7.46	52.29±6.05	5.42±8.44	11.80±3.34	12.03±3.40	5.06±11.71
GraphTrans	52.01±1.01	51.73±1.05	49.70±2.06	4.18±5.32	58.01±5.66	54.48±4.40	9.93±6.60	11.60±3.38	11.79±3.43	2.21±11.96
BrainGNN	56.04±4.31	55.53±2.32	44.09±9.42	11.09±3.30	59.06±5.21	69.93±6.72	5.91±7.69	15.78±10.69	18.17±11.29	4.19±5.67
BrainNPT	54.99±2.78	51.79±3.31	44.73±4.53	9.27±8.57	61.17±5.31	74.15±3.66	19.12±5.55	15.26±9.69	18.48±10.43	3.70±9.57
BLEG	56.96±3.75	55.51±2.22	52.34±2.12	12.76±3.65	61.32±6.21	74.95±3.96	19.21±4.33	4.78±2.18	5.24±2.43	7.94±4.56
BrainTAL-S1	58.81±4.69	55.95±1.89	54.26±1.34	13.60±4.08	61.57±5.66	73.54±3.47	14.32±6.39	3.61±1.87	4.36±1.82	9.49±3.98
DANN	58.04±4.79	54.86±4.00	32.42±8.30	6.75±0.88	61.91±5.92	72.60±5.86	13.39±7.47	15.39±12.65	17.78±13.54	4.10±15.99
SHOT	55.06±3.55	52.83±8.02	48.61±8.62	8.76±7.82	56.46±4.16	64.08±4.54	11.37±9.04	4.97±2.65	5.48±2.61	6.08±15.02
SFHarmony	55.53±2.18	56.66±3.73	50.00±6.65	10.36±3.55	56.21±5.63	66.09±6.13	6.56±5.76	3.64±1.96	4.40±2.01	3.50±15.29
GraphATA	59.64±5.33	55.63±2.60	54.55±13.21	14.43±4.82	60.68±3.56	69.47±9.76	11.27±8.28	4.85±3.35	5.01±3.42	9.08±6.25
BrainTAL-3B	62.81±4.33	58.60±2.60	56.12±4.45	18.86±4.49	62.61±5.46	75.14±4.00	19.38±4.66	3.48±1.94	4.18±1.99	13.86±4.06
BrainTAL-7B	**63.85±4.12**	**59.82±2.45**	**57.34±4.18**	**20.75±4.20**	**63.58±5.15**	**76.05±3.78**	**21.05±4.32**	**3.26±1.82**	**3.94±1.85**	**15.42±3.85**

4.2 Experimental Settings

All experiments are implemented in PyTorch v1.12.1 on an NVIDIA GeForce RTX 3090 Ti GPU. We primarily use Qwen2.5-3B as the LLM backbone and additionally evaluate Qwen2.5-7B to examine backbone robustness. Both variants use a maximum text length of 256 and $L = 4$ additional brain tokens. LoRA is applied to the attention projections with rank 8, while the Stage-II token adapter uses a bottleneck dimension of 64. We adopt a leave-one-site-out protocol, treating one site as the unlabeled target domain and the remaining sites as source domains. All methods use identical data splits and ten repeated runs for each target site. Tables 1 and 2 report the mean and standard deviation across the five site-level means, where each site-level mean is averaged over the ten runs. BrainTAL is trained in two stages. First, source-site multi-task instruction tuning is performed for 50 epochs with a batch size of 16. The learning rates for the brain encoder and language adaptation parameters are 1e-4 and 2e-5, respectively. Source subjects are split 9:1 into training and validation sets; the best source-validation checkpoint initializes Stage II. Second, source-free target adaptation updates only the lightweight brain token adapter for 10 epochs with a batch size of 4 and a learning rate of 5e-6. The final Stage-II checkpoint is used for evaluation, without accessing target labels for model selection. AdamW is used in both stages, with weight decay 1e-4. The EMA momentum is 0.995 and the confidence threshold is 0.95. All Stage-II hyperparameters were selected without using target labels.

4.3 Comparison with State-of-the-Art Methods

To comprehensively evaluate BrainTAL, we compare it with several categories of representative methods. GCN Kipf & Welling (2016), GAT Veličković et al. (2018), and BrainGNN Li et al. (2021) serve as graph neural network baselines. GraphTrans Wu et al. (2021) and BrainNPT Hu et al. (2024) evaluate global-relation modeling. BLEG Dong et al. (2026) represents LLM-enhanced brain network modeling, and DANN Ganin et al. (2016), SHOT Liang et al. (2020), SFHarmony

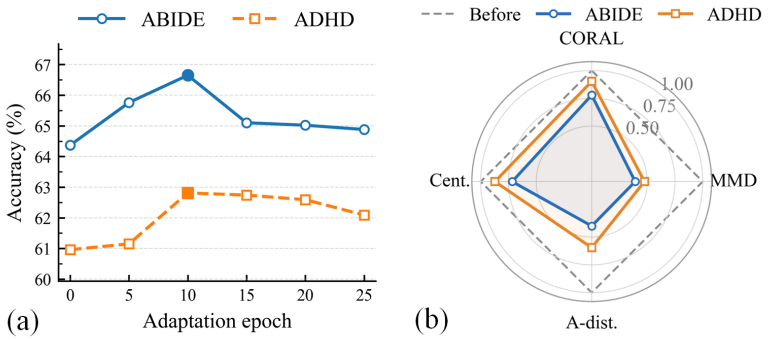

Figure 3: Target-site adaptation and source-target feature distance analysis. (a) Target-site accuracy over adaptation epochs. (b) source-target feature distances before and after adaptation.

Dinsdale et al. (2023), and GraphATA Zhang & He (2025) represent cross-domain adaptation or site-harmonization methods. Non-adaptation baselines and BrainTAL-S1 follow the source-only setting, whereas adaptation baselines and BrainTAL use unlabeled target data according to their respective protocols.

Tables 1 and 2 report leave-one-site-out results on diagnosis, sex, and age prediction. Among the source-only LLM systems, BrainTAL-S1 achieves higher mean diagnosis accuracy than BLEG on both ABIDE and ADHD under the cross-site protocol. Within BrainTAL, target adaptation with the same backbone improves diagnosis accuracy over the source-trained model on both datasets. Sex accuracy and age MAE also improve over BrainTAL-S1, although Stage II uses diagnosis predictions only. These auxiliary-task results measure transfer of the shared calibrated adapter.

4.4 Ablation Study

Table 3 separates the effects of token construction and target calibration. Removing task-conditioned tokenization reduces diagnosis accuracy by 4.93 percentage points on ABIDE and 6.68 points on ADHD. Removing task-specific offsets reduces it by 4.16 and 0.85 points, respectively, with a less pronounced effect on ADHD. Removing the brain-network prior text reduces accuracy by 3.16 and 4.85 points. This variant removes ROI-pair names and signed connection values, so it tests the complete description rather than edge weights alone. Table 3(b) shows contributions of tokenization to source-trained sex and age prediction. For Stage II, the gains over BrainTAL-S1 and the w/o DV and w/o CM variants support the combined target-calibration objective.

4.5 Interpretability and Visualization

To further evaluate BrainTAL, we analyze its cross-site adaptability and task-specific interpretability.

4.5.1 Target-Site Adaptation Analysis.

Figure 3(a) further investigates the effect of prolonged target adaptation by extending Stage II from the default 10 epochs to 25 epochs. Performance improves during the first 10 epochs but gradually declines with further adaptation, suggesting that excessive adaptation may amplify pseudo-label noise and lead to over-adaptation. Figure 3(b) further shows an overall reduction in MMD, CORAL, feature-centroid distance, and A-distance after adaptation, where lower values indicate smaller domain discrepancies. These results suggest that brain token calibration improves target-site performance while promoting source–target feature alignment. Source features are used only for this post hoc analysis and are not accessed during Stage-II target optimization or checkpoint selection.

4.5.2 Task-Related Brain Patterns.

To examine task-related interpretability, we visualize both key brain regions and functional connections across tasks using the AAL atlas Tzourio-Mazoyer et al. (2002). As shown in Figure 4,

Table 3: Ablation studies on ABIDE and ADHD: (a) disease classification and (b) Stage-I sex classification and age prediction. Classification metrics are in %, and MAE is in years.

(a) Disease Classification

Methods	ABIDE				ADHD			
	ACC	AUC	F1	MCC	ACC	AUC	F1	MCC
w/o LLM	60.95	61.56	58.81	23.87	55.77	53.78	38.60	9.65
w/o MT	61.83	61.79	66.82	23.31	56.57	52.06	52.14	12.87
w/o BPT	63.49	63.19	60.55	26.03	57.96	56.64	53.82	12.54
w/o TCBT	61.72	61.29	58.59	24.39	56.13	48.26	47.77	10.22
w/o TSO	62.49	62.64	61.80	26.55	61.96	56.58	52.27	15.68
w/o DV	64.21	64.78	61.73	26.32	60.27	56.25	55.31	16.94
w/o CM	66.32	65.58	65.27	28.68	61.74	57.64	55.82	17.01
BrainTAL-S1	63.93	63.56	63.70	27.72	58.81	55.95	54.26	13.60
BrainTAL	**66.65**	**66.29**	**67.96**	**29.20**	**62.81**	**58.60**	**56.12**	**18.86**

(b) Sex Classification and Age Prediction

Methods	ABIDE		ADHD	
	Sex ACC	Age MAE	Sex ACC	Age MAE
w/o LLM	83.02	7.74	59.25	5.33
w/o BPT	84.23	6.75	60.36	4.71
w/o TCBT	83.88	6.13	59.78	4.75
w/o TSO	84.76	6.86	60.51	3.83
BrainTAL-S1	**85.40**	**5.79**	**61.57**	**3.61**

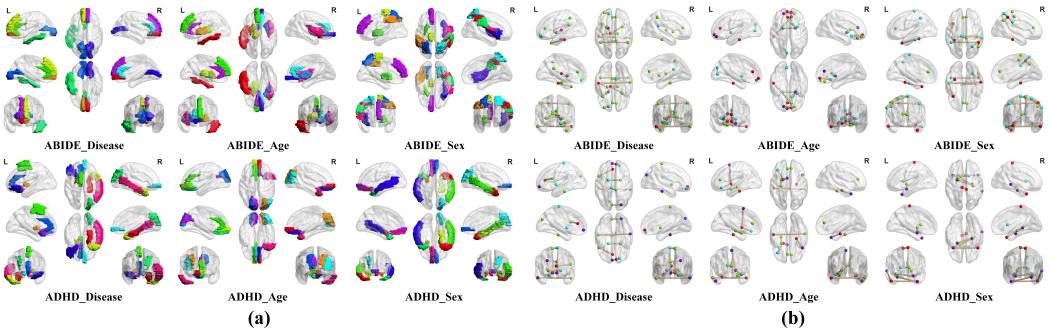

Figure 4: Visualization of task-related brain patterns. (a) Key brain regions across tasks. (b) Key functional connections across tasks.

BrainTAL identifies distinct task-specific spatial patterns. For ASD diagnosis on ABIDE, the highlighted regions mainly involve the medial orbitofrontal cortex, medial superior frontal gyrus, anterior cingulate cortex, and inferior temporal gyrus, consistent with previous findings Zhu et al. (2022); Perovnik et al. (2023). For ADHD diagnosis, the model emphasizes fronto-basal ganglia and temporal regions associated with executive, attentional, and motor functions. Age estimation and sex classification exhibit different regional patterns, while the corresponding functional connections further reveal distinct task-dependent network organization. These visualizations provide qualitative evidence that task-conditioned brain tokens capture task-relevant representations. We interpret the connection patterns within each dataset and do not treat edge appearances across ASD and ADHD panels as calibrated comparisons of connection importance.

5 Conclusion

We propose BrainTAL, a source-free target-adaptive framework that connects brain networks with LLMs through compact connectivity instructions and task-conditioned brain tokens. A lightweight calibration module adapts the brain–language interface using only unlabeled target data. Experiments on ABIDE and ADHD demonstrate improved cross-site performance across disease diagnosis, sex classification, and age prediction, while visualizations highlight task-relevant brain regions and connections, supporting the effectiveness and interpretability of BrainTAL.

AI USE STATEMENT

Generative AI tools were used to assist with language refinement, manuscript editing, and code checking. All AI-assisted content was reviewed and verified by the authors. All experimental results and scientific claims were independently validated by the authors, who take full responsibility for the final content of this work.

ETHICS STATEMENT

This study uses publicly available neuroimaging datasets, ABIDE and ADHD, and involves no new data collection. We use the datasets solely for research purposes and follow their corresponding data-use requirements.

REPRODUCIBILITY STATEMENT

We provide the model architecture and training objectives in Section 3, and report the datasets, evaluation protocol, and major implementation details in Section 4. All experiments follow the same leave-one-site-out protocol with repeated runs for reproducible evaluation.

REFERENCES

Adriana Di Martino, Chao-Gan Yan, Qingyang Li, Erin Denio, Francisco X Castellanos, Kaat Alaerts, Jeffrey S Anderson, Michal Assaf, Susan Y Bookheimer, Mirella Dapretto, et al. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular Psychiatry*, 19(6):659–667, 2014.

Nicola K Dinsdale, Mark Jenkinson, and Ana IL Namburete. Sfharmony: Source free domain adaptation for distributed neuroimaging analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11494–11505, 2023.

Rui Dong, Zitong Wang, Jiaxing Li, Weihuang Zheng, and Youyong Kong. Bleg: Llm functions as powerful fmri graph-enhancer for brain network analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5662–5672, 2026.

Yihang Duan, Shuo Huang, Li Zhang, and Meiling Wang. Iebgl: An interpretability-enhanced brain graph learning framework with llm-instructed topology and literature-augmented semantics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 35331–35340, 2026.

Yuqi Fang, Jinjian Wu, Qianqian Wang, Shijun Qiu, Andrea Bozoki, and Mingxia Liu. Source-free collaborative domain adaptation via multi-perspective feature enrichment for functional mri analysis. *Pattern Recognition*, 157:110912, 2025.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

Sam Gijsen, Marc-Andre Schulz, and Kerstin Ritter. Brain-semantoks: Learning semantic tokens of brain dynamics with a self-distilled foundation model. In *International Conference on Learning Representations*, volume 2026, pp. 13186–13208, 2026.

Wentao Gu, Yuquan Li, Xinyang Jiang, Zilong Wang, Dongsheng Li, Zehui Li, Zijian Dong, and Cai Zhao. Joint adaptation of uni-modal foundation models for multi-modal alzheimer's disease diagnosis. In *International Conference on Learning Representations*, volume 2026, pp. 14667–14693, 2026.

Hao Guan and Mingxia Liu. Domain adaptation for medical image analysis: a survey. *IEEE Transactions on Biomedical Engineering*, 69(3):1173–1185, 2021.

Hao Guan, Yunbi Liu, Erkun Yang, Pew-Thian Yap, Dinggang Shen, and Mingxia Liu. Multi-site mri harmonization via attention-guided deep domain adaptation for brain disorder identification. *Medical image analysis*, 71:102076, 2021.

Jinlong Hu, Yangmin Huang, Nan Wang, and Shoubin Dong. Brainnpt: Pre-training transformer networks for brain network classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 32:2727–2736, 2024.

Xingcan Hu, Wei Wang, and Li Xiao. Fcn-llm: Empower llm for brain functional connectivity network understanding via graph-level multi-task instruction tuning. *arXiv preprint arXiv:2603.01135*, 2026.

Yao Hu, Zhi-An Huang, Rui Liu, Xiaoming Xue, Xiaoyan Sun, Linqi Song, and Kay Chen Tan. Source free semi-supervised transfer learning for diagnosis of mental disorders on fmri scans. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):13778–13795, 2023.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, et al. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*, 2024.

Manoj Kumar Kanakasabapathy, Prudhvi Thirumalaraju, Hemanth Kandula, Fenil Doshi, Anjali Devi Sivakumar, Deeksha Kartik, Raghav Gupta, Rohan Pooniwala, John A Branda, Athe M Tsibris, et al. Adaptive adversarial neural networks for the analysis of lossy and domain-shifted datasets of medical images. *Nature Biomedical Engineering*, 5(6):571–585, 2021.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Jogendra Nath Kundu, Naveen Venkat, R Venkatesh Babu, et al. Universal source-free domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4544–4553, 2020.

Jaein Lee, Eunsong Kang, Da-Woon Heo, and Heung-Il Suk. Site-invariant meta-modulation learning for multisite autism spectrum disorders diagnosis. *IEEE Transactions on Neural Networks and Learning Systems*, 35(12):18062–18075, 2023.

Cheng-Yi Li, Kao-Jung Chang, Cheng-Fu Yang, Hsin-Yu Wu, Wenting Chen, Hritik Bansal, Ling Chen, Yi-Ping Yang, Yu-Chun Chen, Shih-Pin Chen, et al. Towards a holistic framework for multimodal llm in 3d brain ct radiology report generation. *Nature Communications*, 16(1):2258, 2025a.

Shanglin Li, Motoaki Kawanabe, and Reinmar Kobler. Spdim: Source-free unsupervised conditional and label shift adaptation in eeg. In *International Conference on Learning Representations*, volume 2025, pp. 101907–101931, 2025b.

Shengrong Li, Qi Zhu, Chunwei Tian, Li Zhang, Bo Shen, Chuhang Zheng, Daoqiang Zhang, and Wei Shao. Spatio-temporal evolutionary graph learning for brain network analysis using medical imaging. *IEEE Transactions on Image Processing*, 34(12):5860–5872, 2025c.

Shengrong Li, Qi Zhu, Chunwei Tian, Xinyang Zhang, Wei Shao, Jie Wen, and Daoqiang Zhang. Neuroh-tgl: Neuro-heterogeneity guided temporal graph learning strategy for brain disease diagnosis. *Advances in Neural Information Processing Systems*, 38:109349–109371, 2026.

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*, 74:102233, 2021.

Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International Conference on Machine Learning*, pp. 6028–6039. PMLR, 2020.

Yiheng Liu, Zhengliang Liu, Zihao Wu, Junhao Ning, Haiyang Sun, Sichen Xia, Yang Yang, Xiaohui Gao, Ning Qiang, Bao Ge, et al. Brain-inspired exploration of functional networks and key neurons in large language models. *arXiv preprint arXiv:2502.20408*, 2025. doi: https://doi.org/10.48550/arXiv.2502.20408.

Jinglei Lv, Binbin Lin, Qingyang Li, Wei Zhang, Yu Zhao, Xi Jiang, Lei Guo, Junwei Han, Xintao Hu, Christine Guo, et al. Task fmri data analysis based on supervised stochastic coordinate coding. *Medical Image Analysis*, 38:1–16, 2017.

Shengbing Pei, Jiajun Ma, Zhao Lv, Chao Zhang, and Jihong Guan. Community-aware graph transformer for brain disorder identification. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 4191–4199, 2025.

Matej Perovnik, Tomaž Rus, Katharina A Schindlbeck, and David Eidelberg. Functional brain networks in the evaluation of patients with neurodegenerative disorders. *Nature Reviews Neurology*, 19(2):73–90, 2023.

Liting Shi, Lei Shi, Zhijun Cui, Chengting Lin, Rui Zhang, Jiayi Zhang, Yechen Zhu, Wei Shi, Jianlin Wang, Yanlong Wang, et al. Transformer-based structural connectivity networks for adhd-related connectivity alterations. *Communications Medicine*, 5(1):296, 2025.

Ruoxian Song, Peng Cao, Guangqi Wen, Pengfei Zhao, Ziheng Huang, Xizhe Zhang, Jinzhu Yang, and Osmar R Zaiane. Braindas: Structure-aware domain adaptation network for multi-site brain network analysis. *Medical Image Analysis*, 96:103211, 2024.

Jianpo Su, Hui Shen, Limin Peng, and Dewen Hu. Few-shot domain-adaptive anomaly detection for cross-site brain images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(3): 1819–1835, 2021.

Nathalie Tzourio-Mazoyer, Brigitte Landeau, Dimitri Papathanassiou, Fabrice Crivello, Octave Etard, Nicolas Delcroix, Bernard Mazoyer, and Marc Joliot. Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain. *NeuroImage*, 15(1):273–289, 2002.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph Attention Networks. *International Conference on Learning Representations*, 2018.

Yan Wang, Jian Cheng, Yixin Chen, Shuai Shao, Lanyun Zhu, Zhenzhou Wu, Tao Liu, and Haogang Zhu. Fvp: Fourier visual prompting for source-free unsupervised domain adaptation of medical image segmentation. *IEEE Transactions on Medical Imaging*, 42(12):3738–3751, 2023.

Jianghao Wu, Guotai Wang, Ran Gu, Tao Lu, Yinan Chen, Wentao Zhu, Tom Vercauteren, Sébastien Ourselin, and Shaoting Zhang. Upl-sfda: Uncertainty-aware pseudo label guided source-free domain adaptation for medical image segmentation. *IEEE Transactions on Medical Imaging*, 42 (12):3932–3943, 2023.

Zhanghao Wu, Paras Jain, Matthew Wright, Azalia Mirhoseini, Joseph E Gonzalez, and Ion Stoica. Representing long-range context for graph neural networks with global attention. *Advances in Neural Information Processing Systems*, 34:13266–13279, 2021.

Wen Xiong, Junzhong Ji, and Jinduo Liu. Brainec-llm: Brain effective connectivity estimation by multiscale mixing llm. *Advances in Neural Information Processing Systems*, 38:123110–123143, 2026.

Yanwu Yang, Chenfei Ye, Xutao Guo, Tao Wu, Yang Xiang, and Ting Ma. Mapping multi-modal brain connectome for brain disorder diagnosis via cross-modal mutual learning. *IEEE Transactions on Medical Imaging*, 43(1):108–121, 2023.

Yanwu Yang, Chenfei Ye, Guinan Su, Ziyao Zhang, Zhikai Chang, Hairui Chen, Piu Chan, Yue Yu, and Ting Ma. Brainmass: Advancing brain network analysis for diagnosis with large-scale self-supervised learning. *IEEE Transactions on Medical Imaging*, 43(11):4004–4016, 2024.

Xu Yin, John Q Gan, and Haixian Wang. Cogformer: A unified multi-scale brain representation for visual decoding and reconstruction from fmri. *IEEE Transactions on Medical Imaging*, 45(6): 3021–3038, 2026.

Kunyu Zhang and Tianxiang Xu. Brainriem: Riemannian prototype learning for source-free cross-site brain network diagnosis. *arXiv preprint arXiv:2606.29200*, 2026.

Yingtao Zhang, Diego Cerretti, Jialin Zhao, Wenjing Wu, Ziheng Liao, Umberto Michieli, and Carlo Vittorio Cannistraci. Brain network science modelling of sparse neural networks enables transformers and llms to perform as fully connected. *Advances in Neural Information Processing Systems*, 38:48241–48277, 2026.

Zhen Zhang and Bingsheng He. Aggregate to adapt: Node-centric aggregation for multi-source-free graph domain adaptation. In *Proceedings of the ACM on Web Conference 2025*, pp. 4420–4431, 2025.

Hongyi Zhi, Tianyou Yu, Zhenghui Gu, Zhuobin Lin, Le Che, Yuanqing Li, and Zhuliang Yu. Supervised contrastive learning-based domain generalization network for cross-subject motor decoding. *IEEE Transactions on Biomedical Engineering*, 72(1):401–412, 2024.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4396–4415, 2022.

Qi Zhu, Ruting Xu, Ran Wang, Xijia Xu, Zhiqiang Zhang, and Daoqiang Zhang. Stacked topological preserving dynamic brain networks representation and classification. *IEEE Transactions on Medical Imaging*, 41(11):3473–3484, 2022.