Prior Text-Informed Gate Attention Framework for Multimodal Psychiatric Disorder Diagnosis

Wensheng Zhai[1†] **Kaizhong Zheng**[1†] **Difei Mei**[2] **Liangjun Chen**[1*] **Badong Chen**[1]

[1]State Key Laboratory of Human-Machine Hybrid Augmented Intelligence,
Institute of Artificial Intelligence and Robotics,
Xi'an Jiaotong University

[2]China Mobile Xiong'an Information and Communication Technology Co., Ltd.

zhaiws@stu.xjtu.edu.cn, kzzheng@xjtu.edu.cn, fanyihong@cmict.chinamobile.com,
chenbd@mail.xjtu.edu.cn, liangjunchen@xjtu.edu.cn

Abstract

Diagnosis of psychiatric disorders from functional magnetic resonance imaging (fMRI) is hindered by the scarcity of clinically meaningful text paired with individual scans, high inter-subject variability, and limited interpretability of multimodal fusion. We propose a prior text-informed gated attention framework (PTGA) that converts region-of-interest (ROI) level neuroimaging measures into statistically grounded textual priors via percentile-based soft binning. PTGA combines hemodynamic response function (HRF) aware spatiotemporal encoding with static graph topology modeling, and unifies variable-length fMRI time series through masked global temporal aggregation. It further uses text semantics as an inductive bias for prior-guided gated fusion, improving robustness under cross-modal noise. We compare PTGA with eight strong graph and multimodal baselines under a unified protocol. PTGA achieves up to 10.46% and 8.01% absolute improvements in autism spectrum disorder (ASD) and major depressive disorder (MDD) diagnosis, respectively, outperforming SOTA baselines. The learned gating also provides biologically plausible evidence: MDD related signals are concentrated in sensorimotor/visual/default-mode systems, while ASD related evidence is centered on the cerebellar network with joint contributions from default-mode, limbic, and dorsal-attention networks.

1 Introduction

Diagnosing psychiatric disorders such as major depressive disorder (MDD) and autism spectrum disorder (ASD) remains constrained by the subjectivity of behavioral assessments and the lack of reliable physiological biomarkers [1, 2]. Although functional magnetic resonance imaging (fMRI) offers valuable insights into brain functional organization [3], learning robust neural signatures from fMRI is challenging due to low signal-to-noise ratios and pronounced inter-subject variability, which together hinder the extraction of discriminative and generalizable representations [4]. These limitations highlight the need for more robust computational frameworks to enable objective and automated psychiatric diagnosis.

From a methodological perspective, current fMRI classification research is primarily categorized into three paradigms: (i) ***Image-Only***, focusing on learning discriminative features directly from voxel-level or region-of-interest (ROI) time series [5]; (ii) ***Image-Anatomy***, which integrates structural

[†]These authors contributed equally to this work and share first authorship.
[*]Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

priors (such as atlas-based partitioning) to ensure biological plausibility [6]; and (iii) ***Image-Text***, which aligns neuroimaging with clinical semantics to balance diagnostic accuracy and interpretability [7]. Among these, Image-Text paradigm demonstrates immense theoretical potential, yet multimodal fusion methodologies remain in their infancy. Limited availability of paired fMRI–text data and pervasive semantic noise have restricted prior multimodal methods to early proof-of-concept studies. As a result, there remains a substantial gap in principled and scalable frameworks that can deliver both reliable diagnostic gains and clinically meaningful interpretability [8, 9].

Recently, fMRI text modeling has increasingly focused on prompt learning and alignment based foundation models. Despite promising progress, three limitations remain [10, 11]. ***First***, current text generation is often weakly grounded in subject-level imaging markers, such as regional homogeneity, amplitude of low frequency fluctuation, and functional connectivity, which produces clinically inconsistent descriptions [12, 13]. ***Second***, coarse fusion strategies, including global alignment and direct concatenation, neglect ROI-level heterogeneity across spatial and semantic dimensions, limiting effective noise suppression [14, 15]. ***Third***, representation alignment alone often provides insufficient task relevant evidence for robust diagnosis in complex psychiatric disorders [16, 17].

We present PTGA, a multimodal framework for psychiatric diagnosis that integrates statistical priors into fMRI text fusion, and evaluate it on four multi-site cohorts with 7,536 subjects. Our main contributions are:

- **Statistically Constrained Clinical Prior Construction:** We construct subject-specific textual priors by mapping ROI-level regional homogeneity (ReHo), amplitude of low-frequency fluctuation (ALFF), and functional connectivity (FC) measures to percentile-bin labels and exact percentile positions relative to healthy controls in the training split. We use these labels and percentile values to constrain a large language model (LLM) during text generation to reduce semantic drift and improve consistency between generated descriptions and subject-specific imaging evidence.
- **Prior Informed Multiplicative Gating:** We introduce a gated attention mechanism that uses textual semantics as an inductive bias to modulate spatiotemporal and semantic feature interactions at ROI resolution, enabling robust cross-modal denoising under high heterogeneity.
- **Clinically Aligned Interpretability:** We provide a three level interpretability pipeline, including blinded expert evaluation of generated priors, cross LLM consistency assessment, and ROI to network gate attribution, and we observe disease specific patterns consistent with established clinical neuroanatomy for MDD and ASD.

2 Related Work

2.1 Brain Disease Diagnostic Model

Recent progress in brain disease diagnosis primarily leverages Graph Neural Networks (GNN) and Transformers to decode complex connectivity patterns from fMRI [18]. Representative GNN-based frameworks such as HGNN [19], IBrainGNN [20], and multicenter graph convolutional network (GCN) [21] focus on integrating heterogeneous modalities and phenotypic priors, while MAF-GNN [22] introduces multi-atlas fusion via cross-attention to mitigate information loss inherent in single template analysis. To capture the brain's hierarchical organization, BrainHGT [23] and CAGT [24] employ community-aware mechanisms to balance local interactions with long-range dependencies, while Fuse-Former [25] focuses on multi-scale temporal blood-oxygenation-level-dependent (BOLD) fluctuations. Furthermore, ContrastPool [26] and HoloGraph [27] enhance clinical interpretability and structural robustness through contrastive learning and bio-inspired oscillatory synchronization mechanisms. Despite these advancements, existing models struggle to explicitly ground neuroimaging features in clinical statistical facts and lack fine-grained, text-guided mechanisms for cross-modal feature filtering, often resulting in semantic misalignment and limited diagnostic reliability.

2.2 fMRI-Text Collaborative Model

In recent years, the integration of fMRI data with textual information has emerged as a promising frontier for advancing the diagnosis of brain disorders [28]. Current research primarily follows three pathways: prompt-driven enhancement, multimodal semantic alignment, and automated text generation. Specifically, BrainPrompt [29] and fTSPL [30] leverage LLM to design multi level

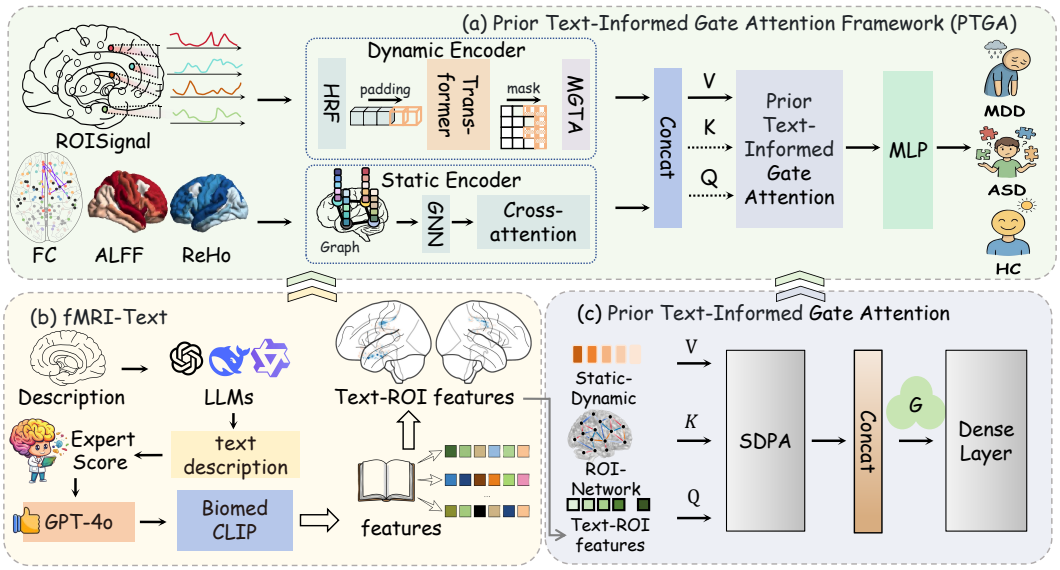


Figure 1: Overview of the PTGA Framework. (a) The framework of PTGA. MGTA represents the masked global temporal aggregation, and G represents prior text-informed gate attention. (b)Text generation and encoding module. (c) Gate attention guided by text priors. Two neuroimaging experts independently evaluated the generated texts.

or collaborative prompts, integrating external medical knowledge to guide brain network analysis. On the alignment front, fMRI-LM [14] and BrainTGA [31] attempt to bridge the modal gap and enhance cross-task transferability by projecting fMRI-derived features into a shared semantic space via foundational models or wavelet image transformers. To address the inherent scarcity of fMRI-text paired data, RTGMFF [32] introduces an ROI-based text generation mechanism to provide descriptive labels for imaging features. Despite these advances, several critical limitations remain: (i) existing text generation mechanisms often rely on hard threshold matching (e.g., binary activation states), overlooking continuous pathological changes within individuals and potentially leading to factual medical hallucinations. (ii) multimodal fusion in frameworks like PMIL [33] remains relatively static, lacking fine-grained ROI-level adaptive gate mechanisms for cross-modal noise filtering. (iii) many models sacrifice temporal integrity by pruning time points or employing sliding windows, rather than capturing full-range dynamic variations through dedicated convolutional layers.

3 Proposed Method

PTGA integrates medical-statistical text priors with spatiotemporal fMRI features via a biologically inspired multiplicative gate mechanism (Figure 1). The framework comprises three core components: (a) ***Spatiotemporal Encoding***, which models BOLD dynamics using hemodynamic response function (HRF)-shaped temporal filtering followed by Transformer encoding and masked global temporal aggregation (MGTA), while capturing static topology via cyclic cross-attention graph encoding; (b) ***Fact-Grounded Text Priors***, which maps ReHo/ALFF/FC to clinical semantics via a percentile-based soft-interval (P_{bins}) strategy; (c) ***Prior-Guided Gate***, where a multiplicative gate $\mathbf{g}(\cdot)$ adaptively fuses modalities by aligning clinical semantics with functional network priors.

3.1 Problem Definition

Given a cohort $\mathcal{S}$, each subject $i \in \mathcal{S}$ is characterized by a dynamic multivariate time series $\mathbf{X}_i^{\text{dyn}} \in \mathbb{R}^{R \times T_i}$ and a set of static brain features $\mathcal{F}_i = \{\mathbf{x}_i^{\text{ReHo}}, \mathbf{x}_i^{\text{ALFF}}, \mathbf{A}_i^{\text{FC}}\}$, where R denotes the number of regions of interest (ROIs), T_i is the subject-specific temporal length, and $\mathbf{A}_i^{\text{FC}} \in \mathbb{R}^{R \times R}$ is the functional connectivity adjacency matrix. Let $\mathcal{D}_i = \{d_{i,1}, \ldots, d_{i,R}\}$ denote the ROI-wise clinical textual priors and $\mathbf{P}$ define a subject-invariant anatomical prior matrix. The task is formulated as a

binary classification problem to learn a parameterized mapping $f_\theta : (\mathbf{X}_i^{\text{dyn}}, \mathcal{F}_i, \mathcal{D}_i; \mathbf{P}) \to \hat{y}_i \in \{0, 1\}$, where $\hat{y}_i$ is the predicted diagnostic label.

3.2 Text Generation and Encoding

To mitigate hallucination and factual drift in deep learning-based medical text generation, we propose a fact-locking mechanism anchored in medical statistics. We constrain the LLM with a structured instruction template that explicitly injects metric-grounded statistical evidence.

Statistical Binning and Semantic Mapping. For an ROI-level static metric $u \in \{\text{ReHo, ALFF, FC}\}$, we define a set of percentile thresholds $\mathcal{Q} = \{q_5, q_{25}, q_{50}, q_{75}, q_{95}\}$ calculated from the healthy control distribution Table 1. In practice, the percentile thresholds $\mathcal{Q}$ are estimated only from healthy controls in the training split. The resulting $\mathcal{Q}$ is then kept fixed when generating texts for validation/test subjects, preventing any information leakage. We define a mapping function $\mathcal{M}(u_{i,r})$ that discretizes continuous neural activity into symbolic medical facts.

Instruction Tuning with Percentile Distribution. Beyond discrete labels, we compute the exact percentile $p_{i,r} = F(u_{i,r})$ for each ROI, where $F(\cdot)$ is the cumulative distribution function (Figure 2). This p-distribution is injected into the LLM via a curated prompt template $\mathcal{I}$. The model is instruction-tuned to synthesize these anchors into a clinical description $d_{i,r} = \text{LLM}(\mathcal{M}(u_{i,r}), p_{i,r}, \text{ROI}_r)$. Finally, the generated descriptions are encoded into a unified latent space using a pre-trained BiomedCLIP model, denoted as $\mathbf{h}_{i,r}^{\text{text}} = \text{BiomedCLIP}(d_{i,r})$ and $\mathbf{H}_i^{\text{text}} = [\mathbf{h}_{i,1}^{\text{text}}, \ldots, \mathbf{h}_{i,R}^{\text{text}}]^\top$.

Table 1: Percentile-based soft-bin mapping for $\mathcal{M}(u_{i,r})$.

Condition on $u_{i,r}$	**Mapped label $\mathcal{M}(u_{i,r})$**
$u_{i,r} \le q_5$	Extremely Low
$q_5 < u_{i,r} \le q_{25}$	Relatively Low
$q_{25} < u_{i,r} \le q_{50}$	Slightly Low
$q_{50} < u_{i,r} \le q_{75}$	Slightly High
$q_{75} < u_{i,r} \le q_{95}$	Relatively High
$u_{i,r} > q_{95}$	Extremely High

3.3 Graph Encoding

Dynamic Encoder. We encode the full-length ROI BOLD sequence using fixed HRF-shaped temporal filtering, Transformer sequence modeling, and masked global temporal aggregation (MGTA). The canonical double-gamma HRF is sampled according to the repetition time (TR) and used as a fixed temporal kernel. Forward convolution of observed BOLD provides a physiologically motivated filtering/smoothing inductive bias; it is not deconvolution and does not recover latent neuronal activity. Same-length filtering and masking retain the number of time points remaining after preprocessing while accommodating variable sequence lengths:

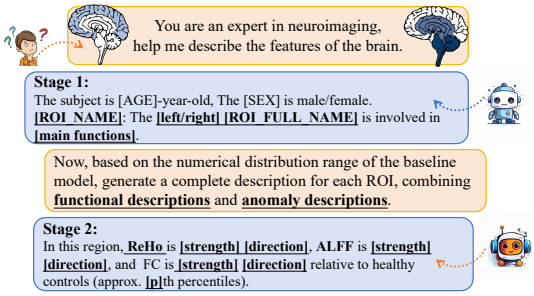


Figure 2: Example of text generation based on medical statistics.

$$\mathbf{h}_{i,r}^{\text{dyn}} = \text{MGTA}(\text{Transformer}(\mathbf{x}_{i,r} * h(\tau))), \tag{1}$$

where $\mathbf{h}_{i,r}^{\text{dyn}} \in \mathbb{R}^d$. The vector $\mathbf{x}_{i,r} \in \mathbb{R}^{T_i}$ denotes the ROI-specific time series of region r for subject i, extracted from $\mathbf{X}_i^{\text{dyn}}$. Detailed mathematical derivations and batch modeling steps are provided in Appendix B.

Static Encoder. To capture multi-scale neurobiological signatures, we construct independent graph views for ReHo, ALFF, and FC and encode them with modality-specific GCNs. We then implement a cyclic cross-query fusion scheme where Query (Q), Key (K), and Value (V) roles are rotated among modality embeddings to achieve mutual refinement and high-order coupling, yielding three interaction outputs $\{\mathbf{Z}_i^{(m)}\}_{m=1}^3$. To preserve the complementary cross-modal correlation patterns captured by

different rotations, we adopt a concatenation-projection fusion as the final static representation:

$$\mathbf{H}_i^{\text{sta}} = \mathrm{LN}\Big(\Big[\mathbf{Z}_i^{(1)} \parallel \mathbf{Z}_i^{(2)} \parallel \mathbf{Z}_i^{(3)}\Big]\,\mathbf{W}_F + \mathbf{b}_F\Big), \tag{2}$$

where $\mathrm{LN}(\cdot)$ denotes Layer Normalization, and $\parallel$ denotes feature concatenation. This concatenation-projection fusion is used in all experiments. Appendix C details static graph network construction.

3.4 Text Guided Multiplicative Gate Mechanism

Inspired by the efficacy of gate mechanisms in modeling non-linear interactions and enhancing training stability, we propose a Text-Prior Guided Multiplicative Gate module. This module adaptively integrates high-level clinical semantics with anatomical priors to refine the fusion of brain functional features. We map heterogeneous modalities into a shared d-dimensional space to enable ROI-wise fine-grained interaction. Let $r \in \{1, 2, \ldots, R\}$ index ROIs. For subject i and ROI r, we denote the BiomedCLIP-encoded textual embedding as $\mathbf{h}_{i,r}^{\text{text}} \in \mathbb{R}^d$, the static brain feature embedding as $\mathbf{h}_{i,r}^{\text{sta}} \in \mathbb{R}^{d_s}$, and the dynamic brain feature embedding as $\mathbf{h}_{i,r}^{\text{dyn}} \in \mathbb{R}^{d_d}$.

We also assume an anatomical prior vector $\mathbf{p}_r \in [0,1]^9$, the ROI r and functional networks are encoded, including the seven canonical Yeo functional networks: Visual network (VIS), Somatomotor network (SMN), Dorsal attention network (DAN), Ventral attention network (VAN), Limbic network (LIM), Frontoparietal control network (FPN), Default mode network (DMN) [34]. In addition, we include the Salience network (SN) and the Cerebellar network (Cere) to form a total of nine functional networks. Dice similarity between these networks and each ROI is then calculated. We first construct a semantic query, an anatomical key, and a spatiotemporal value for each ROI, all projected to $\mathbb{R}^d$:

$$\mathbf{q}_{i,r} = \mathbf{W}_Q\,\mathbf{h}_{i,r}^{\text{text}} + \mathbf{b}_Q, \tag{3}$$

$$\mathbf{k}_r = \mathbf{W}_K\,\mathbf{p}_r + \mathbf{b}_K, \tag{4}$$

$$\mathbf{v}_{i,r} = \mathrm{LN}\Big(\mathbf{W}_V\,[\mathbf{h}_{i,r}^{\text{sta}} \oplus \mathbf{h}_{i,r}^{\text{dyn}}] + \mathbf{b}_V\Big), \tag{5}$$

where $\mathbf{W}_Q \in \mathbb{R}^{d\times d}$, $\mathbf{W}_K \in \mathbb{R}^{d\times 9}$, and $\mathbf{W}_V \in \mathbb{R}^{d\times(d_s+d_d)}$ are learnable matrices, $\mathbf{b}_Q, \mathbf{b}_K, \mathbf{b}_V \in \mathbb{R}^d$ are biases, $\oplus$ denotes concatenation. $\mathbf{k}_r$ is subject-invariant, serving as an anatomical constraint derived from atlas-to-network overlap, while $\mathbf{q}_{i,r}$ is subject-specific, summarizing fact-anchored clinical semantics from text.

We perform scaled dot-product attention (SDPA) by anchoring the semantic query to the anatomical key across all ROIs, thereby capturing whole-brain dependencies constrained by anatomical priors. Denote $\mathbf{q}_{i,r}$ as the query for ROI r in subject i and $\mathbf{k}_j$ as the key for ROI j. The attention weights are:

$$\alpha_{i,r,j} = \frac{\exp\Big(\frac{\mathbf{q}_{i,r}^\top \mathbf{k}_j}{\sqrt{d}}\Big)}{\sum_{\ell=1}^{R}\exp\Big(\frac{\mathbf{q}_{i,r}^\top \mathbf{k}_\ell}{\sqrt{d}}\Big)}, \tag{6}$$

the SDPA output for ROI r is then the weighted sum of values:

$$\mathbf{y}_{i,r} = \sum_{j=1}^{R} \alpha_{i,r,j}\,\mathbf{v}_{i,j}, \tag{7}$$

to implement SDPA output gate (i.e., $\mathbf{Y}' = \mathbf{Y} \odot \sigma(\cdot)$), we compute a multiplicative gate from the element-wise alignment between semantic intent and anatomical constraint at the same ROI index r:

$$\mathbf{m}_{i,r} = \mathbf{q}_{i,r} \odot \mathbf{k}_r, \tag{8}$$

$$\mathbf{g}_{i,r} = \sigma(\mathbf{W}_g\,\mathbf{m}_{i,r} + \mathbf{b}_g), \tag{9}$$

where $\odot$ is the Hadamard product, $\sigma(\cdot)$ is the sigmoid, $\mathbf{g}_{i,r} \in (0,1)^d$, $\mathbf{W}_g \in \mathbb{R}^{d\times d}$, $\mathbf{b}_g \in \mathbb{R}^d$ are learnable parameters. Finally, the gate SDPA output is:

$$\mathbf{y}'_{i,r} = \mathbf{y}_{i,r} \odot \mathbf{g}_{i,r}, \tag{10}$$

this design ensures that prior-driven global interaction is injected only when clinical semantics and anatomical roles are mutually consistent: if $\mathbf{q}_{i,r}$ contradicts $\mathbf{k}_r$, the interaction $\mathbf{m}_{i,r}$ yields a small

gate, suppressing the SDPA output for that ROI and preventing unreliable prior injection. To preserve discriminative local spatiotemporal signatures and improve optimization stability, we add a residual pathway from the raw neural value and form the final ROI representation as:

$$\mathbf{o}_{i,r} = \mathrm{LN}\big(\mathbf{v}_{i,r} + \mathbf{y}'_{i,r}\big)\,, \tag{11}$$

stacking $\{\mathbf{o}_{i,r}\}_{r=1}^{R}$ yields $\mathbf{O}_i \in \mathbb{R}^{R\times d}$. We then obtain a subject-level diagnostic embedding via an attention pooling operator:

$$\boldsymbol{\beta}_i = \mathrm{softmax}(\mathbf{O}_i\,\mathbf{w}_p) \in \mathbb{R}^R, \tag{12}$$

$$\mathbf{Y}_i = \sum_{r=1}^{R} \boldsymbol{\beta}_i[r]\,\mathbf{o}_{i,r} \in \mathbb{R}^d, \tag{13}$$

where $\mathbf{w}_p \in \mathbb{R}^d$ is learnable and $\mathrm{softmax}(\cdot)$ is applied over the R ROI. In this way, the model adaptively emphasizes ROI whose textual evidence is consistent with their anatomical priors, while still retaining robust local signals through the residual connection. When atypical pathological patterns cause semantic-anatomical mismatch, the SDPA output gate naturally suppresses prior context for the affected ROIs, thereby mitigating prior-induced hallucinations and improving robustness.

3.5 Loss Function

The PTGA model is optimized using a classification loss $\mathcal{L}_{\mathrm{cls}}$, an ROI-level alignment loss $\mathcal{L}_{\mathrm{align}}$, and a prior-consistency loss $\mathcal{L}_{\mathrm{prior}}$. Let B denote the batch size, $R = 116$ the number of ROIs, and d the common latent dimension. The classification term is the batch-mean binary cross-entropy between the classification logits and diagnostic labels. The text, static, and dynamic ROI representations share the same d-dimensional space. The alignment loss encourages agreement between the textual representation and the mean of the static and dynamic imaging representations:

$$\mathcal{L}_{\mathrm{align}} = \frac{1}{BR}\sum_{i=1}^{B}\sum_{r=1}^{R}\left[1 - \cos\left(\mathbf{h}_{i,r}^{\mathrm{text}}, \frac{\mathbf{h}_{i,r}^{\mathrm{sta}} + \mathbf{h}_{i,r}^{\mathrm{dyn}}}{2}\right)\right], \tag{14}$$

where $\cos(\cdot,\cdot)$ denotes cosine similarity. Let $g_{i,r,c}$ denote the activation of gate channel c, $\mathbf{q}_{i,r}$ the semantic query, and $\mathbf{k}_r$ the projected anatomical key. The prior-consistency loss regularizes the channel-mean gate activation toward their sigmoid-transformed cosine compatibility:

$$\mathcal{L}_{\mathrm{prior}} = \frac{1}{BR}\sum_{i=1}^{B}\sum_{r=1}^{R}\left[\frac{1}{d}\sum_{c=1}^{d} g_{i,r,c} - \sigma(\cos(\mathbf{q}_{i,r}, \mathbf{k}_r))\right]^2, \tag{15}$$

where $\sigma(\cdot)$ denotes the sigmoid function. The total training objective is:

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{cls}} + \lambda_1\mathcal{L}_{\mathrm{align}} + \lambda_2\mathcal{L}_{\mathrm{prior}}, \tag{16}$$

where λ_1 and λ_2 control the contributions of ROI-level alignment and prior-consistency regularization, respectively.

4 Experiments

4.1 Experimental Setup

Datasets and Preprocessing. We evaluated our framework on four large-scale, multi-site fMRI datasets: ABIDE-I [35], ABIDE-II, RESTMDD-I [36], and RESTMDD-II [37]. The integrated cohort comprises a total of 7,536 participants, with detailed demographic statistics summarized in Table 2. The preprocessing process mainly includes discarding the first 10 time points, time layer correction, head movement correction, interference covariate regression, and spatial normalization. Detailed technical parameters can be

Table 2: Dataset statistics

Dataset	Gender (M/F)	Age (mean±std)	Total	Case/Healthy
ABIDE-I	918/157	16.85±7.97	1075	517/558
ABIDE-II	847/243	14.45±9.01	1090	518/572
RESTMDD-I	925/1455	36.20±15.11	2380	1276/1104
RESTMDD-II	1169/1822	32.67±13.17	2991	1650/1341

Note: M and F denote Male and Female.

Table 3: Classification performance results. The best results are in **bold**, and the second best are underlined. * indicates statistical significance after false discovery rate (FDR)-corrected paired t-tests (adjusted p<0.05).

Model	Text-Prompt	ABIDE-I				ABIDE-II			
		ACC(%)	SEN(%)	SPE(%)	AUC(%)	ACC(%)	SEN(%)	SPE(%)	AUC(%)
HGNN (2025)		60.42±2.91	57.10±2.95	63.80±2.99	63.22±2.88	61.08±2.88	58.40±2.75	64.04±2.81	64.33±2.90
MAF-GNN (2025)		59.44±3.12	58.21±3.44	60.69±2.15	62.11±2.88	58.62±4.89	55.98±4.03	61.54±6.68	58.76±3.93
IBrainGNN (2025)		62.12±0.12	68.50±0.19	56.00±0.26	62.25±0.16	65.91±0.18	63.02±0.25	69.08±0.19	66.05±0.16
BrainHGT (2026)		62.79±2.80	62.66±5.80	62.93±5.21	62.80±3.90	64.81±2.36	67.53±4.60	61.81±3.56	64.67±2.92
fTSPL (2024)	✓	62.33±2.81	58.34±2.95	66.34±2.99	65.34±2.88	67.57±0.15	70.06±0.18	64.05±0.35	67.89±0.22
RTGMFF (2025)	✓	66.29±0.18	61.00±0.24	72.00±0.21	66.50±0.16	67.82±2.32	73.05±2.98	63.46±2.33	68.26±1.89
PMIL (2025)	✓	61.11±2.88	57.12±2.99	65.12±2.95	64.12±2.91	65.64±2.44	61.88±2.77	69.88±2.62	65.88±1.91
BrainPrompt (2025)	✓	66.57±1.35	71.79±1.88	61.55±2.71	66.67±1.65	69.39±1.14	71.50±2.56	67.66±1.32	74.20±1.78
Ours	✓	**74.52±2.08***	**74.15±1.74***	**75.78±2.54***	**76.27±1.50***	**79.85±1.35***	**80.11±2.82***	**79.73±2.06***	**82.32±2.46***
		RESTMDD-I				RESTMDD-II			
HGNN (2025)		63.52±2.84	60.12±2.67	67.44±2.88	65.31±2.41	62.40±2.76	59.20±2.88	66.34±2.91	64.15±2.33
MAF-GNN (2025)		58.48±3.86	56.76±5.42	60.48±5.56	60.17±3.12	60.17±1.38	60.12±5.35	58.02±7.50	60.44±3.87
IBrainGNN (2025)		64.05±2.62	60.34±2.81	68.34±2.91	65.89±2.18	68.15±2.22	64.56±2.45	72.56±2.33	69.87±1.95
BrainHGT (2026)		66.82±1.82	70.89±3.64	60.36±3.84	66.42±2.17	65.89±1.13	66.72±6.85	63.10±8.27	65.83±4.62
fTSPL (2024)	✓	61.27±2.88	57.44±2.99	65.68±2.95	62.89±1.87	64.93±2.42	61.40±2.55	69.28±2.62	66.72±1.64
RTGMFF (2025)	✓	66.08±2.30	64.61±2.23	67.61±2.15	67.43±1.92	66.76±2.45	63.45±2.77	70.79±2.62	68.53±2.01
PMIL (2025)	✓	63.91±2.62	61.22±2.81	67.02±2.77	65.47±1.63	64.71±2.55	61.12±2.88	69.12±2.81	66.37±1.78
BrainPrompt (2025)	✓	69.97±2.09	**74.69±2.37***	64.13±2.45	70.28±1.54	71.02±1.91	68.34±2.11	74.34±1.99	72.89±1.64
Ours	✓	**75.18±2.16***	72.75±2.07	**76.09±1.67***	**77.34±1.82***	**79.03±1.84***	**76.70±1.25***	**79.93±1.37***	**80.82±2.10***

found in Appendix A. For each dataset, subjects were partitioned into training, validation, and testing subsets following $7:1:2$. To ensure that the clinical distribution was preserved across all subsets, a stratified sampling strategy was employed, maintaining a balanced ratio of patients to healthy controls in each split.

Implementation Details. Our PTGA framework and all baseline models were implemented in PyTorch and executed on an NVIDIA RTX PRO 6000. Model optimization was conducted using the Adam optimizer with a learning rate of 1×10^{-3}, a batch size of 64, and a weight decay of 1×10^{-4} to prevent overfitting. We trained the models for a maximum of 200 epochs, employing an early stopping mechanism with a patience of 20 epochs based on validation accuracy. The model checkpoint achieving the peak validation performance within each run was selected for final assessment on the held-out test set.

Metrics. Model performance was evaluated using Accuracy (ACC), Sensitivity (SEN), Specificity (SPE), and area under the curve (AUC). To minimize partitioning bias, we report the mean and standard deviation over 10 independent runs with stratified shuffling.

Baseline Models. We compare the proposed PTGA against two categories of state-of-the-art baselines: (i) Graph-based Models, which focus on brain network topology and functional connectivity, including HGNN [19], MAF-GNN [22], IBrainGNN [20], and BrainHGT [23]. (ii) Multimodal Models, which integrate fMRI data with textual information, including fTSPL [30], RTGMFF [32], PMIL [33], and BrainPrompt [29]. For fair comparison, all methods are evaluated using the same datasets, preprocessing pipeline.

4.2 Experimental Results

Classification Results. Table 3 summarizes the classification results across four datasets, where PTGA achieves the best performance on most metrics. As shown, compared to the second-best baseline models, our model achieves ACC improvements of 7.95%, 10.46%, 5.21%, and 8.01% on the ABIDE-I, ABIDE-II, RESTMDD-I, and RESTMDD-II datasets, respectively. These improvements are statistically significant (adjusted p-value <0.05) according to paired t-tests with FDR correction. This strongly indicates that integrating a statistically constrained text prior with ROI-level prior-guided gating enables more effective filtering of cross-modal noise and enhances discriminative representation learning. The aforementioned significant and consistent performance gains further demonstrate that the proposed framework exhibits superior robustness and generalization capabilities in multi-site heterogeneous scenarios.

Hyperparameter Results. We perform one-at-a-time sensitivity analysis for λ_1 (semantic alignment) and λ_2 (prior consistency), each swept over $\{0.01, 0.1, 0.5, 1.0\}$ while fixing the other parameter. As shown in Figure 3, PTGA is stable across four datasets and consistently performs best at the balanced setting $\lambda_1 = \lambda_2 = 0.5$. Moving either weight to extreme values degrades performance, indicating

that diagnosis quality depends on balanced semantic supervision and anatomical prior regularization rather than either term alone.

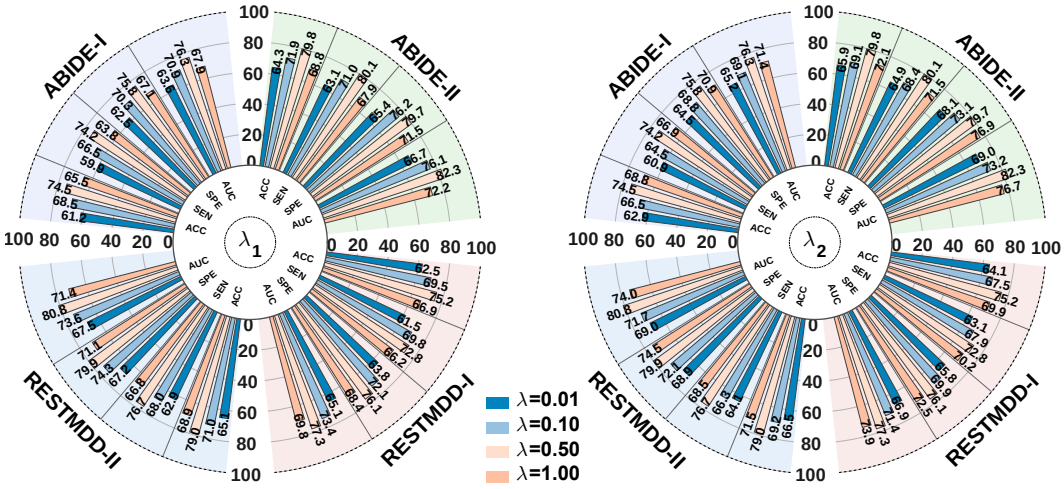


Figure 3: Hyperparameter analysis of PTGA performance relative to the semantic alignment weight λ_1 and prior-consistency weight λ_2.

4.3 Ablation Study

Ablation Analysis. To evaluate the contribution of each core component in the PTGA framework, we conduct a series of ablation experiments by comparing the full model with five derived variants: **(i) w/o Temporal**: We replace the HRF-based temporal processing with standard timepoint truncation and linear projection to evaluate the overall contribution of the temporal processing design. **(ii) w/o Text**: We remove all textual modalities and the text-guided branch, reducing PTGA to a pure Image-Anatomy framework, to quantify the diagnostic contribution of clinical semantic priors. **(iii) w/o P_{bins}**: We replace the proposed percentile-based soft-bin grounding with unconstrained LLM-generated descriptions to evaluate the contribution of statistics-grounded prompting to classification performance. **(iv) w/o Gate**: We remove the multiplicative gate and use standard SDPA to evaluate the contribution of multiplicative gating to cross-modal fusion and classification performance. **(v) w/o HRF**: We replace only the fixed HRF operator with the identity, keeping sequence length, projection, Transformer, masking, MGTA, all downstream modules, data splits, and optimization unchanged to isolate the contribution of HRF-shaped temporal filtering.

Ablation Results. PTGA achieves the highest mean accuracy across all four datasets (Table 4). The compound temporal ablation (w/o Temporal) reduces ACC by 10.98–13.71 percentage points, while removing text guidance (w/o Text) decreases ACC by 9.47–12.43 percentage points. Replacing percentile-bin grounding (w/o P_{bins}) reduces ACC by 7.68–11.47 percentage points, and removing multiplicative gating (w/o Gate) decreases ACC by 4.64–10.70 percentage points. Removing only the HRF filter (w/o HRF) yields a smaller ACC reduction of 1.92–2.87 percentage points, supporting a modest benefit from HRF-shaped temporal filtering. Together, these results support the contributions of temporal processing, imaging-derived text guidance, statistical grounding, and gated fusion to classification performance.

4.4 Interpretability Study

Interpretable Analysis. We examine PTGA along three complementary dimensions: (i) the clinical plausibility and inter-rater reliability of generated textual priors; (ii) their semantic consistency across LLM generators; and (iii) gate-based attribution of textual and imaging information across ROIs and functional networks.

- **Expert rating of clinical prior quality.** For each of the 116 ROIs, we generate one description using each of GPT-4o, DeepSeek-V3.1, and Qwen3 under the same standardized prompt. Two neuroimaging experts, blinded to generator identity, independently rate each text for medical

Table 4: Ablation study on ABIDE and RESTMDD datasets. The best results are in **bold**, and the second best are underlined.

Setting	ABIDE-I				ABIDE-II			
	ACC(%)	SEN(%)	SPE(%)	AUC(%)	ACC(%)	SEN(%)	SPE(%)	AUC(%)
w/o Temporal	62.55±3.10	61.25±3.28	63.85±2.54	65.02±2.30	66.14±3.41	65.80±3.22	66.48±2.95	68.61±2.42
w/o Text	65.05±2.99	64.20±2.88	65.90±3.11	67.52±2.35	67.42±2.11	66.12±3.51	68.72±2.22	69.89±2.30
w/o Gate	69.88±3.17	71.86±3.48	67.91±2.76	72.35±2.47	69.15±3.15	68.20±3.15	70.10±2.51	71.62±2.24
w/o P_{bins}	65.40±2.23	63.85±2.40	66.96±3.47	67.88±2.34	70.33±2.67	69.45±2.41	71.21±3.32	72.80±2.28
w/o HRF	72.60±2.55	71.80±2.81	73.34±2.61	74.18±2.11	76.98±1.91	76.40±2.52	77.51±2.21	79.54±2.74
PTGA (Full)	**74.52±2.08**	**74.15±1.74**	**75.78±2.54**	**76.27±1.50**	**79.85±1.35**	**80.11±2.82**	**79.73±2.06**	**82.32±2.46**
	RESTMDD-I				RESTMDD-II			
w/o Temporal	63.10±2.44	62.40±2.71	63.80±3.42	65.57±2.42	68.05±2.84	67.45±2.66	68.65±3.61	70.52±2.49
w/o Text	64.28±3.21	63.55±3.01	65.01±2.88	66.75±2.31	67.34±2.33	63.34±2.66	71.34±2.55	69.81±2.05
w/o Gate	69.86±3.59	70.79±3.40	68.93±2.64	72.33±2.39	71.42±3.44	70.12±3.11	72.72±2.11	73.89±2.09
w/o P_{bins}	67.50±2.20	70.43±3.41	64.58±2.54	69.97±2.36	67.56±2.34	64.55±2.66	70.57±2.50	70.03±2.45
w/o HRF	73.10±2.48	71.00±2.86	75.53±2.52	75.50±2.17	76.72±2.36	74.60±2.34	79.33±2.81	78.10±2.53
PTGA (Full)	**75.18±2.16**	**72.75±2.07**	**76.09±1.67**	**77.34±1.82**	**79.03±1.84**	**76.70±1.25**	**79.93±1.37**	**80.82±2.10**

Table 5: Expert evaluation results of three LLMs.

Model	Medical Accuracy	Semantic Coherence	Clinical Relevance	Overall Performance	Weighted Score
GPT-4o	4.41	4.32	4.33	4.35	**4.36**
DeepSeek-V3.1	4.11	4.03	4.01	4.05	4.04
Qwen3	3.80	3.74	3.72	3.75	3.74

accuracy, semantic coherence, and clinical relevance on a five-point scale. Inter-rater agreement is assessed using weighted Cohen's κ (Appendix D). As shown in Table 5, GPT-4o receives the highest mean ratings across the three criteria, supporting its selection as the default prior generator. This analysis assesses expert-rated clinical plausibility and the reliability of the ratings.

- **Cross-LLM semantic consistency.** We compare generated descriptions with reference texts using BERTScore-F1, Sentence-BERT (SBERT) cosine similarity, and Consensus-based Image Description Evaluation (CIDEr). Scores are computed for matched subject–ROI samples and aggregated by generator. Pairwise differences are assessed using two-sided paired t-tests with FDR correction. As shown in Figure 4, GPT-4o achieves higher BERTScore-F1 and SBERT cosine similarity than DeepSeek-V3.1 and Qwen3, while maintaining competitive CIDEr scores. These metrics quantify agreement with the reference descriptions across generators.
- **Gate-weight interpretability of multimodal fusion.** Based on prior clinical neuroimaging studies, we adopt a predefined functional-network template for ROI-level interpretation. We analyze TextPriorWeight (pre-gate text prior), GateTextWeight (post-gate text contribution), and GateImgWeight (post-gate imaging contribution), and map ROI-level weights to template-defined networks, including VIS, SMN, VAN, DAN, FPN, DMN, LIM, SN, and Cere systems.

Interpretable Results. Across the three interpretability analyses, we obtain consistent evidence for PTGA. ***First***, expert evaluation (Table 5) ranks GPT-4o highest in overall clinical quality, supporting its use as the default prior generator. ***Second***, in cross-LLM hallucination assessment (Fig. 4), GPT-4o achieves the strongest semantic fidelity, with higher BERTScore-F1 and SBERT cosine than DeepSeek-V3.1 and Qwen3 while maintaining competitive CIDEr; these gains remain significant after FDR correction. ***Third***, gate-weight analysis (Fig. 5) shows disease-specific ROI attribution patterns: for MDD, top GateImgWeight ROIs are mainly distributed in SMN/VIS/DMN with coordinated LIM/DAN

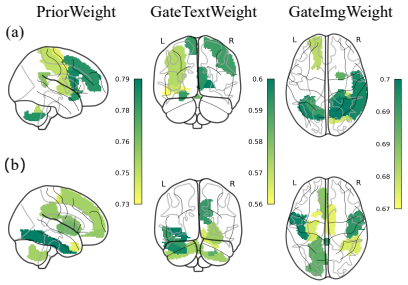


Figure 5: ROI-level modality attribution before and after gating. (a) illustrates the MDD group, and (b) illustrates the ASD group. TextPriorWeight denotes pre-gate text-prior contribution, GateTextWeight denotes post-gate text contribution, and GateImgWeight denotes post-gate imaging contribution.

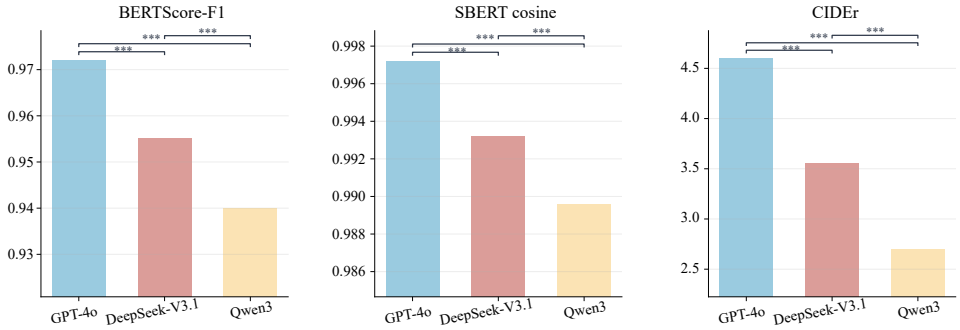


Figure 4: Cross-LLM semantic consistency on ROI-level generated priors, evaluated by BERTScore-F1, SBERT cosine, and CIDEr. Significance markers denote pairwise comparisons after FDR correction: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

involvement, whereas for ASD, high-weight ROIs are concentrated in Cere with coordinated DMN/LIM/DAN contributions. These patterns provide biologically plausible network-level evidence for PTGA.

5 Limitations

Limitations include the dependence on atlas-based ROI partitioning and histogram-based voxel distribution modeling, as well as sensitivity to site-specific preprocessing and LLM prior quality. In addition, the current evaluation scope is limited to four datasets, binary ASD/MDD tasks, and 10 shuffled runs, without leave-site-out or longitudinal validation. Future work will explore voxel- or surface-level parcellations, more expressive distribution estimators, and broader validation across additional disorders and longitudinal settings.

6 Conclusion

We presented PTGA, a multimodal framework that integrates statistically grounded textual priors with spatiotemporal fMRI representations for psychiatric disorder classification. PTGA constructs ROI-level descriptions using percentile-based statistical anchors and combines text semantics with anatomical priors to guide multiplicative gated fusion. Evaluation on four multi-site datasets demonstrates consistent accuracy gains over eight competitive baselines for ASD and MDD classification. Ablation studies support the contributions of temporal modeling, percentile-grounded text construction, and multiplicative gating, with a smaller benefit from HRF-shaped temporal filtering. Expert evaluation and cross-LLM semantic comparisons assess the quality of the generated priors, while gate-based attribution identifies ROI and network patterns consistent with previous neuroimaging findings. Future work will examine generalization to unseen sites and longitudinal cohorts.

Acknowledgments and Disclosure of Funding

This work was supported by Noncommunicable Chronic Diseases-National Science and Technology Major Project (2024ZD0527800) and the National Natural Science Foundation of China (32541016, 62436005, U25A20540, 62473303).

Competing Interests. The authors declare no competing interests.

References

[1] Samira M Epp, Gabriel Castrillón, Beijia Yuan, Jessica Andrews-Hanna, Christine Preibisch, and Valentin Riedl. Bold signal changes can oppose oxygen metabolism across the human cortex. *Nature neuroscience*, pages 1–12, 2025.

[2] Benjamin Yost Hayden, Sarah R Heilbronner, and Seng Bum Michael Yoo. Rethinking the centrality of brain areas in understanding functional organization. *Nature Neuroscience*, pages 1–12, 2025.

[3] Andrea I Luppi, S Parker Singleton, Justine Y Hansen, Keith W Jamison, Danilo Bzdok, Amy Kuceyeski, Richard F Betzel, and Bratislav Misic. Contributions of network structure, chemoarchitecture and diagnostic categories to transitions between cognitive topographies. *Nature Biomedical Engineering*, 8(9):1142–1161, 2024.

[4] Yinuo Zhang, Demao Liu, Zhichao Liang, Jiani Cheng, Kexin Lou, Jinqiao Duan, Ting Gao, Bin Hu, and Quanying Liu. Artificial intelligence as a surrogate brain: Bridging neural dynamical models and data. *National Science Review*, page nwaf457, 2025.

[5] Weikang Qiu, Zheng Huang, Haoyu Hu, Aosong Feng, Yujun Yan, and Rex Ying. Mindllm: A subject-agnostic and versatile model for fmri-to-text decoding. *arXiv preprint arXiv:2502.15786*, 2025.

[6] Yanchao Tan, Hang Lv, Yunfei Zhan, Guofang Ma, Bo Xiong, and Carl Yang. Boxlm: Unifying structures and semantics of medical concepts for diagnosis prediction in healthcare. In *Forty-second International Conference on Machine Learning*.

[7] Abhijeet Parida, Zhifan Jiang, Roger J Packer, Robert A Avery, Syed M Anwar, and Marius G Linguraru. Quantitative metrics for benchmarking medical image harmonization. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2024.

[8] Tatiana Medvedeva, Irina Knyazeva, Ruslan Masharipov, Alexander Korotkov, Denis Cherednichenko, and Maxim Kireev. Benchmarking resting state fmri connectivity pipelines for classification: Robust accuracy despite processing variability in cross-site eye state prediction. *bioRxiv*, pages 2025–10, 2025.

[9] Yanting Yang and Xiaoxiao Li. Neurolip: Interpretable and fair cross-modal alignment of fmri and phenotypic text. *arXiv preprint arXiv:2503.21964*, 2025.

[10] Chengsheng Mao, Jie Xu, Luke Rasmussen, Yikuan Li, Prakash Adekkanattu, Jennifer Pacheco, Borna Bonakdarpour, Robert Vassar, Li Shen, Guoqian Jiang, et al. Ad-bert: Using pre-trained language model to predict the progression from mild cognitive impairment to alzheimer's disease. *Journal of Biomedical Informatics*, 144:104442, 2023.

[11] Jie Sun, Tangsheng Lu, Xuexiao Shao, Ying Han, Yu Xia, Yongbo Zheng, Yongxiang Wang, Xinmin Li, Arun Ravindran, Lizhou Fan, et al. Practical ai application in psychiatry: historical review and future directions. *Molecular Psychiatry*, pages 1–10, 2025.

[12] Runda Li. Integrative diagnosis of psychiatric conditions using chatgpt and fmri data. *BMC psychiatry*, 25(1):145, 2025.

[13] Liang Peng, Songyue Cai, Zongqian Wu, Huifang Shang, Xiaofeng Zhu, and Xiaoxiao Li. Mmgpl: Multimodal medical data analysis with graph prompt learning. *Medical Image Analysis*, 97:103225, 2024.

[14] Yuxiang Wei, Yanteng Zhang, Xi Xiao, Chengxuan Qian, Tianyang Wang, and Vince D Calhoun. fmri-lm: Towards a universal foundation model for language-aligned fmri understanding. *arXiv preprint arXiv:2511.21760*, 2025.

[15] Xiaoyu Chen, Changde Du, Che Liu, Yizhe Wang, and Huiguang He. Bp-gpt: Auditory neural decoding using fmri-prompted llm. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

[16] Weihao Xia, Raoul de Charette, Cengiz Oztireli, and Jing-Hao Xue. Umbrae: Unified multimodal brain decoding. In *European Conference on Computer Vision*, pages 242–259. Springer, 2024.

[17] Wanaiu Huang. Brainchat: Decoding semantic information from fmri using vision-language pretrained models. *arXiv preprint arXiv:2406.07584*, 2024.

[18] Qihe Huang, Lei Shen, Ruixin Zhang, Shouhong Ding, Binwu Wang, Zhengyang Zhou, and Yang Wang. Crossgnn: Confronting noisy multivariate time series via cross interaction refinement. *Advances in Neural Information Processing Systems*, 36:46885–46902, 2023.

[19] Yuheng Gu, Shoubo Peng, Yaqin Li, Linlin Gao, and Yihong Dong. Fc-hgnn: A heterogeneous graph neural network based on brain functional connectivity for mental disorder identification. *Information Fusion*, 113:102619, 2025.

[20] Gang Qu, Ziyu Zhou, Vince D Calhoun, Aiying Zhang, and Yu-Ping Wang. Integrated brain connectivity analysis with fmri, dti, and smri powered by interpretable graph neural networks. *Medical Image Analysis*, 103:103570, 2025.

[21] Xuegang Song, Feng Zhou, Alejandro F Frangi, Jiuwen Cao, Xiaohua Xiao, Yi Lei, Tianfu Wang, and Baiying Lei. Multicenter and multichannel pooling gcn for early ad diagnosis based on dual-modality fused brain network. *IEEE Transactions on Medical Imaging*, 42(2):354–367, 2022.

[22] Zhihong Chen, Jiang Wu, Li Pu, Shaoqing Li, Yuhao Zhang, Dezhong Yao, and Hongmei Yan. Maf-gnn: Graph neural network-based multi-atlas brain functional information fusion for major depressive disorder diagnosis with rs-fmri. *Information Processing & Management*, 63(3):104524, 2026.

[23] Jiajun Ma, Yongchao Zhang, Chao Zhang, Zhao Lv, and Shengbing Pei. Brainhgt: A hierarchical graph transformer for interpretable brain network analysis. *arXiv preprint arXiv:2511.17604*, 2025.

[24] 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*, pages 4191–4199, 2025.

[25] Jiayu Ye, Yanting Li, An Zeng, Dan Pan, Alzheimer's Disease Neuroimaging Initiative, et al. Fuseformer: An interpretability analysis model for rs-fmri based on multi-scale information fusion interaction. *Biomedical Signal Processing and Control*, 105:107471, 2025.

[26] Jiaxing Xu, Qingtian Bian, Xinhang Li, Aihu Zhang, Yiping Ke, Miao Qiao, Wei Zhang, Wei Khang Jeremy Sim, and Balázs Gulyás. Contrastive graph pooling for explainable classification of brain networks. *IEEE Transactions on Medical Imaging*, 43(9):3292–3305, 2024.

[27] Tingting Dan, Jiaqi Ding, and Guorong Wu. Explore brain-inspired machine intelligence for connecting dots on graphs through holographic blueprint of oscillatory synchronization. *Nature Communications*, 16(1):9425, 2025.

[28] Arunkumar Kannan, Martin A Lindquist, and Brian Caffo. Brainmt: A hybrid mamba-transformer architecture for modeling long-range dependencies in functional mri data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 150–160. Springer, 2025.

[29] Jiaxing Xu, Kai He, Yue Tang, Wei Li, Mengcheng Lan, Xia Dong, Yiping Ke, and Mengling Feng. Brainprompt: Multi-level brain prompt enhancement for neurological condition identification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 172–182. Springer, 2025.

[30] Pengyu Wang, Huaqi Zhang, Zhibin He, Zhihao Peng, and Yixuan Yuan. ftspl: Enhancing brain analysis with fmri-text synergistic prompt learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 564–574. Springer, 2024.

[31] Wenfeng Tan, Hong Yang, Shanshan Ye, Ji Zhang, Peng Zhang, and Zhi Qiao. Text-graph alignment based on llms instruction learning and wavelet graph transformers for brain disease analysis. *Neurocomputing*, page 131355, 2025.

[32] Junhao Jia, Yifei Sun, Yunyou Liu, Cheng Yang, Changmiao Wang, Feiwei Qin, Yong Peng, and Wenwen Min. Rtgmff: Enhanced fmri-based brain disorder diagnosis via roi-driven text generation and multimodal feature fusion. *arXiv preprint arXiv:2509.03214*, 2025.

[33] Hyoungshin Choi, Jonghun Kim, Jiwon Chung, Bo-yong Park, and Hyunjin Park. Pmil: Prompt enhanced multimodal integrative analysis of fmri combining functional connectivity and temporal latency. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 543–553. Springer, 2025.

[34] BT Thomas Yeo, Fenna M Krienen, Jorge Sepulcre, Mert R Sabuncu, Danial Lashkari, Marisa Hollinshead, Joshua L Roffman, Jordan W Smoller, Lilla Zöllei, Jonathan R Polimeni, et al. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of neurophysiology*, 2011.

[35] Cameron Craddock, Yassine Benhajali, Carlton Chu, Francois Chouinard, Alan Evans, András Jakab, Budhachandra Singh Khundrakpam, John David Lewis, Qingyang Li, Michael Milham, et al. The neuro bureau preprocessing initiative: open sharing of preprocessed neuroimaging data and derivatives. *Frontiers in Neuroinformatics*, 7(27):5, 2013.

[36] Xiao Chen, Bin Lu, Hui-Xian Li, Xue-Ying Li, Yu-Wei Wang, Francisco Xavier Castellanos, Li-Ping Cao, Ning-Xuan Chen, Wei Chen, Yu-Qi Cheng, et al. The direct consortium and the rest-meta-mdd project: towards neuroimaging biomarkers of major depressive disorder. *Psychoradiology*, 2(1):32–42, 2022.

[37] Xiao Chen, Bin Lu, Yu-Wei Wang, Xue-Ying Li, Zi-Han Wang, Hui-Xian Li, Yi-Fan Liao, Daniel M Blumberger, Francisco Xavier Castellanos, Eduardo A Garza-Villarreal, et al. Subgenual anterior cingulate cortex functional connectivity abnormalities in depression: insights from brain imaging big data and precision-guided personalized intervention via transcranial magnetic stimulation. *Science Bulletin*, 70(16):2676–2690, 2025.

A Datasets and preprocessing

We used four multi-site resting-state fMRI datasets, including ABIDE-I, ABIDE-II, RESTMDD-I, and RESTMDD-II. For transparency, we report both the original cohort scale and the final included sample size after our dataset-specific cleaning. ABIDE-I was originally aggregated from 17 international sites and contains 1075 subjects (517 ASD and 558 typical controls; age range 7-64 years, median 14.7 years across groups). After excluding subjects with failed spatial registration and evidently poor image quality (e.g., severe artifacts or incomplete coverage), we retained ABIDE-I with N=1075. ABIDE-II was contributed by 19 sites (including charter institutions and new members) with 1090 subjects in total (518 ASD and 572 controls; age range 5-64 years); applying the same QC criteria (registration failure and low-quality scans) yielded a final ABIDE-II cohort of N=1090.

For MDD, RESTMDD-I corresponds to DIRECT Phase-I (REST-meta-MDD), where 25 cohorts were shared from 17 participating hospitals in China, comprising R-fMRI data from 1276 MDD patients and 1104 healthy controls. During harmonization, we removed the entire Site 4 because its data overlapped with Site 14 (duplicated/overlapping samples), which could otherwise introduce data leakage; the final RESTMDD-I sample size was N=2380. RESTMDD-II corresponds to DIRECT Phase-II, which assembled a substantially expanded dataset including 23 cohorts with 1650 MDD patients and 1341 healthy controls. For RESTMDD-II, subjects without valid years of education were excluded to ensure completeness of key demographic covariates, resulting in a final RESTMDD-II cohort of N=2991.

To reduce variability introduced by heterogeneous preprocessing pipelines, we adopted a unified preprocessing workflow for all datasets using DPABI (`http://www.rfmri.org/dpabi`). Specifically, we discarded the first 10 volumes, performed slice-timing correction, and realigned each subject's fMRI time series using a 6-parameter rigid-body transformation. Each subject's T1-weighted image was registered to the mean functional image (6 degrees of freedom, without resampling) and segmented into gray matter, white matter, and cerebrospinal fluid. We then used DARTEL to estimate the deformation field from native space to MNI152. A linear trend regressor was included to compensate for BOLD signal drift. Finally, the brain was divided into 116 ROIs using the Automated Anatomical Labeling (AAL) atlas, and all voxels were time segmented. Pearson correlation was used to construct the FC matrix, and Fisher z-transform was used to normalize the obtained correlation matrix to an approximate Gaussian distribution, ensuring the statistical validity of subsequent machine learning applications.

B Dynamic Feature Encoding

Let $\mathbf{x}_{i,r} = [x_{i,r}(1), \ldots, x_{i,r}(T_i)]^\top$ denote the preprocessed BOLD time series of subject i in ROI r, where T_i is the subject-specific sequence length after preprocessing. We encode each variable-length sequence into a fixed-dimensional dynamic representation through HRF-shaped temporal filtering, Transformer sequence modeling, and masked global temporal aggregation (MGTA).

B.1 HRF-informed Temporal Filtering

We apply a fixed double-gamma HRF-shaped kernel to introduce a hemodynamically motivated temporal inductive bias into BOLD sequence modeling. The kernel is sampled according to the acquisition repetition time (TR) and normalized before filtering. Let $h_i[k]$ denote the resulting discrete kernel for subject i, with length L_i. The kernel is shared across ROIs within each subject and remains fixed during training. For each ROI, temporal filtering is defined as

$$\tilde{x}_{i,r}[n] = \sum_{k=0}^{L_i-1} h_i[k]\, x_{i,r}[n-k], \qquad n = 1, \ldots, T_i, \tag{17}$$

where $x_{i,r}[m] = 0$ for indices outside the observed sequence. We compute the full discrete convolution and retain its first T_i output samples, yielding $\tilde{\mathbf{x}}_{i,r} \in \mathbb{R}^{T_i}$. This produces a filtered sequence with the same length as the preprocessed input. Filtering is performed independently for each ROI before mini-batch padding, and the resulting sequence is passed to the Transformer encoder.

B.2 Transformer Sequence Modeling with Attention Masking

To enable efficient mini-batch training, we right-pad each filtered sequence $\tilde{\mathbf{x}}_{i,r}$ with zeros to the maximum sequence length within the batch, $T_{\max}$. Let $\mathbf{z}_{i,r} \in \mathbb{R}^{T_{\max}}$ denote the padded filtered sequence. We define a padding indicator mask as

$$m_i(t) = \begin{cases} 0, & 1 \le t \le T_i, \\ -\infty, & T_i < t \le T_{\max}. \end{cases} \tag{18}$$

We first project the scalar input to d dimensions:

$$\mathbf{e}_{i,r,t} = \mathbf{W}_{\text{in}}\, z_{i,r}(t) + \mathbf{b}_{\text{in}} + \mathbf{p}_t, \tag{19}$$

where $\mathbf{p}_t$ is a positional encoding. Stacking $\mathbf{e}_{i,r,t}$ yields $\mathbf{E}_{i,r} \in \mathbb{R}^{T_{\max} \times d}$.

We employ a Transformer encoder with an additive attention mask $\mathbf{M}_i \in \mathbb{R}^{T_{\max} \times T_{\max}}$, defined by $M_i(t,s) = m_i(s)$, where t indexes queries and s indexes keys. This prevents valid query positions from attending to padded positions. For a single self-attention layer:

$$\text{Attn}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{Softmax}\!\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}} + \mathbf{M}_i\right)\mathbf{V}, \tag{20}$$

where the softmax is applied over key positions and $\mathbf{M}_i$ assigns $-\infty$ to logits corresponding to padded keys. Outputs at padded query positions are excluded from the subsequent MGTA pooling. The Transformer outputs hidden states

$$\mathbf{H}^{\text{seq}}_{i,r} = \text{Transformer}(\mathbf{E}_{i,r}, \mathbf{M}_i), \tag{21}$$

with $\mathbf{H}^{\text{seq}}_{i,r} \in \mathbb{R}^{T_{\max} \times d}$, while only the first T_i positions correspond to valid observations.

B.3 Masked Global Temporal Aggregation (MGTA) for Dimensional Unification

Since T_i varies across subjects, we aggregate the sequence $\mathbf{H}^{\text{seq}}_{i,r}$ into a fixed-dimensional ROI-wise dynamic embedding via masked attention pooling. We compute a scalar importance score for each time step:

$$u_{i,r,t} = \mathbf{w}_a^\top \mathbf{h}^{\text{seq}}_{i,r,t}, \tag{22}$$

and convert scores to normalized weights using a masked softmax:

$$\alpha_{i,r,t} = \frac{\exp\big(u_{i,r,t} + m_i(t)\big)}{\sum_{j=1}^{T_{\max}} \exp\big(u_{i,r,j} + m_i(j)\big)}. \tag{23}$$

Finally, the unified dynamic feature is obtained by a weighted sum over time:

$$\mathbf{h}^{\text{dyn}}_{i,r} = \sum_{t=1}^{T_{\max}} \alpha_{i,r,t}\, \mathbf{h}^{\text{seq}}_{i,r,t}. \tag{24}$$

where i indexes subjects and r indexes ROIs. T_i is the original sequence length and $T_{\max}$ is the maximum padded length in a mini-batch. d is the Transformer hidden dimension. $\mathbf{h}^{\text{seq}}_{i,r,t} \in \mathbb{R}^d$ is the hidden state at time t. $\mathbf{w}_a \in \mathbb{R}^d$ is a learnable pooling vector. The mask $m_i(t)$ ensures that padded positions have zero contribution both in the softmax normalization and in the final aggregation, preventing padding-induced bias while preserving full-length temporal information.

C Static Encoder

C.1 Graph construction and encoding

For each subject i, we construct modality-specific graphs $\mathcal{G}_i^{(k)} = (\mathcal{V}, \mathcal{E}^{(k)})$ sharing the same node set $\mathcal{V} = \{v_1, \ldots, v_R\}$, where each node corresponds to one ROI.

For functional connectivity (FC), we compute the ROI-to-ROI Pearson correlation matrix from BOLD time series and use it as the weighted adjacency matrix. Since the physiological interpretation of

negative correlations is still controversial, we set all negative correlations to zero and retain only positive connections for subsequent GCN encoding. The node feature of each ROI is defined as its FC connectivity profile (i.e., the vector of its positive correlations to all other ROIs).

For ReHo and ALFF, we construct an ROI-to-ROI adjacency matrix using voxel-level information. Specifically, for each subject and modality $m \in \{\mathrm{ReHo}, \mathrm{ALFF}\}$, we collect voxel-wise values within each ROI r to form a voxel-level sample set $X_r^{(m)}$. We z-score voxel-wise values within each subject and modality before histogram estimation. For each ROI, we estimate an empirical probability distribution $P_r^{(m)}$ from its voxel samples using a histogram with 64 equal-width bins. To ensure numerical stability for Kullback–Leibler (KL) divergence computation, we add a small constant (e.g., ϵ) to avoid zero bins and renormalize the histogram, and discard non-finite divergence values if any. Using these voxel-level distributions, we compute a symmetric KL divergence between every ROI pair (j, k) and convert it into a bounded similarity to form the ReHo/ALFF adjacency matrix:

$$A_{jk}^{(m)} = \frac{1}{1 + D_{\mathrm{KL}}\left(P_j^{(m)} \,\|\, P_k^{(m)}\right) + D_{\mathrm{KL}}\left(P_k^{(m)} \,\|\, P_j^{(m)}\right)}, \tag{25}$$

where $m \in \{\mathrm{ReHo}, \mathrm{ALFF}\}$, larger $A_{jk}^{(m)}$ indicates higher similarity between the voxel-level distributions of ROIs j and k. We use the node strength (weighted degree) as the node feature:

$$c_r^{(m)} = \sum_{j=1}^{R} A_{rj}^{(m)}, \quad \mathbf{F}_i^{(m)} = [c_1^{(m)}, \ldots, c_R^{(m)}]^\top, \tag{26}$$

For each modality $k \in \{\mathrm{FC}, \mathrm{ReHo}, \mathrm{ALFF}\}$, we add self-loops and apply symmetric normalization:

$$\tilde{\mathbf{A}}_i^{(k)} = \mathbf{A}_i^{(k)} + \mathbf{I}, \qquad \tilde{\mathbf{D}}_{i,rr}^{(k)} = \sum_{j=1}^{R} \tilde{\mathbf{A}}_{i,rj}^{(k)}. \tag{27}$$

The GCN layer updates node representations as:

$$\mathbf{E}_i^{(k)} = \sigma\left((\tilde{\mathbf{D}}_i^{(k)})^{-\frac{1}{2}} \tilde{\mathbf{A}}_i^{(k)} (\tilde{\mathbf{D}}_i^{(k)})^{-\frac{1}{2}} \mathbf{F}_i^{(k)} \mathbf{W}_{\mathrm{gcn}}^{(k)}\right), \tag{28}$$

where $\sigma(\cdot)$ is an activation function and $\mathbf{W}_{\mathrm{gcn}}^{(k)}$ denotes the modality-specific trainable weights.

C.2 Cyclic cross-modal interaction

Given $\mathbf{E}_i^{(k)} \in \mathbb{R}^{R \times d}$, we couple the three biomarkers via a cyclic Query-Key-Value mechanism that rotates modality roles, ensuring that each view is contextualized by the other two:

$$\mathbf{Z}_i^{(1)} = \mathrm{Softmax}\left(\frac{\mathbf{E}_i^{\mathrm{ReHo}} (\mathbf{E}_i^{\mathrm{ALFF}})^\top}{\sqrt{d}}\right) \mathbf{E}_i^{\mathrm{FC}}, \tag{29}$$

$$\mathbf{Z}_i^{(2)} = \mathrm{Softmax}\left(\frac{\mathbf{E}_i^{\mathrm{ALFF}} (\mathbf{E}_i^{\mathrm{FC}})^\top}{\sqrt{d}}\right) \mathbf{E}_i^{\mathrm{ReHo}}, \tag{30}$$

$$\mathbf{Z}_i^{(3)} = \mathrm{Softmax}\left(\frac{\mathbf{E}_i^{\mathrm{FC}} (\mathbf{E}_i^{\mathrm{ReHo}})^\top}{\sqrt{d}}\right) \mathbf{E}_i^{\mathrm{ALFF}}. \tag{31}$$

Instead of simple averaging, which may lead to information washout, we employ a concatenation-based fusion strategy to preserve the unique cross-modal correlation patterns distilled in each cycle. Specifically, the intermediate representations are concatenated along the feature dimension and projected back to the original d-dimensional space:

$$\mathbf{Z}_i^{\mathrm{concat}} = \left[\mathbf{Z}_i^{(1)} \,\|\, \mathbf{Z}_i^{(2)} \,\|\, \mathbf{Z}_i^{(3)}\right], \tag{32}$$

$$\mathbf{H}_i^{\mathrm{fused}} = \mathrm{LayerNorm}\left(\mathbf{Z}_i^{\mathrm{concat}} \mathbf{W}_F + \mathbf{b}_F\right), \tag{33}$$

where $\|$ denotes the concatenation operation, and $\mathbf{W}_F \in \mathbb{R}^{3d \times d}$ and $\mathbf{b}_F \in \mathbb{R}^d$ are trainable parameters that learn the relative importance of different cross-modal views. The final static node embedding for each ROI r is defined as $\mathbf{h}_{i,r}^{\mathrm{sta}} = \left(\mathbf{H}_i^{\mathrm{fused}}\right)_r$.

C.3 Aggregation and normalization

The three interaction outputs $\{\mathbf{Z}_i^{(m)}\}_{m=1}^3$ provide complementary, cross-validated views of each ROI under different query–key–value rotations. To preserve the complementary correlation patterns captured by each rotation while keeping the final representation dimension fixed, we adopt a concatenation–projection fusion followed by Layer Normalization, consistent with the main text.

Specifically, we first concatenate the three interaction streams along the feature dimension:

$$\mathbf{Z}_i^{\text{concat}} = \left[\mathbf{Z}_i^{(1)} \parallel \mathbf{Z}_i^{(2)} \parallel \mathbf{Z}_i^{(3)}\right] \in \mathbb{R}^{R\times 3d}. \tag{34}$$

This step explicitly preserves modality-specific and rotation-specific evidence instead of forcing premature information mixing.

Since direct concatenation increases dimensionality and may introduce scale imbalance across interaction streams, we project the fused representation back to the shared latent space:

$$\mathbf{S}_i = \mathbf{Z}_i^{\text{concat}}\mathbf{W}_F + \mathbf{b}_F \in \mathbb{R}^{R\times d}, \tag{35}$$

where $\mathbf{W}_F \in \mathbb{R}^{3d\times d}$ and $\mathbf{b}_F \in \mathbb{R}^d$ are trainable parameters. The projection learns how to reweight and integrate the three cross-modal interaction views in a data-driven manner.

Because the magnitudes of different interaction components can still vary across modalities and subjects, we further apply Layer Normalization along the feature dimension for each ROI:

$$\mathbf{H}_i^{\text{sta}} = \text{LN}(\mathbf{S}_i) = \gamma \odot \frac{\mathbf{S}_i - \mu(\mathbf{S}_i)}{\sqrt{\sigma^2(\mathbf{S}_i) + \epsilon}} + \beta, \tag{36}$$

where $\mu(\mathbf{S}_i)$ and $\sigma^2(\mathbf{S}_i)$ denote the mean and variance computed over the d channels for each ROI independently, and $\gamma, \beta \in \mathbb{R}^d$ are learnable affine parameters shared across ROIs.

In this way, LN aligns the scale of different interaction terms, improves gradient conditioning, and makes the resulting static embedding less sensitive to subject-specific amplitude fluctuations. The final $\mathbf{H}_i^{\text{sta}} \in \mathbb{R}^{R\times d}$ thus serves as a unified yet biomarker-aware representation in which each ROI is supported by jointly-consistent evidence from regional intensity, local synchrony, and network-level topology. This definition is exactly consistent with Eq. (2) in the main text and is used in all experiments.

D Clinical Interpretability Verification

D.1 Standardized Generation of ROI-wise Medical Priors.

To validate the factual accuracy and clinical reliability of generated medical priors, we adopt a standardized generation pipeline. For $R = 116$ ROIs and their associated statistical indicators (e.g., P_{bins}), we prompt three LLM model(GPT-4o, DeepSeek-V3.1, and Qwen3) using an identical standardized prompt paradigm to produce ROI-specific descriptive texts. This yields R priors per model and $3R = 348$ texts in total, which are subsequently evaluated by experts and further analyzed for cross-model consistency. To ensure objective assessment and minimize bias, all generated clinical descriptions were independently scored by two board-certified neuroimaging specialists, each possessing eight years of clinical expertise. The evaluation followed a rigorous double-blind protocol, ensuring that the specialists remained masked to the specific models generating the content throughout the scoring process.

D.2 Double-blind Expert Rating with a 5-point Scale

To minimize subjective bias, two senior experts with neuroimaging background independently rate all 348 generated texts under a double-blind setting where the model identity is hidden. Each text is scored on a 5-point Likert scale (1=poor, 5=excellent) along three criteria: (1) Medical Accuracy, (2) Semantic Coherence, (3) Clinical Relevance. Let $s_{m,r,d}^{(k)}$ denote the score given by expert $k \in \{1, 2\}$ to the text generated by model $m \in \{\text{GPT-4o}, \text{DeepSeek}, \text{Qwen}\}$ for ROI r on criterion d, where $D = 3$. We define the overall model score as

$$\bar{s}_m = \frac{1}{R \cdot 2 \cdot D} \sum_{r=1}^{R} \sum_{k=1}^{2} \sum_{d=1}^{D} s_{m,r,d}^{(k)}. \tag{37}$$

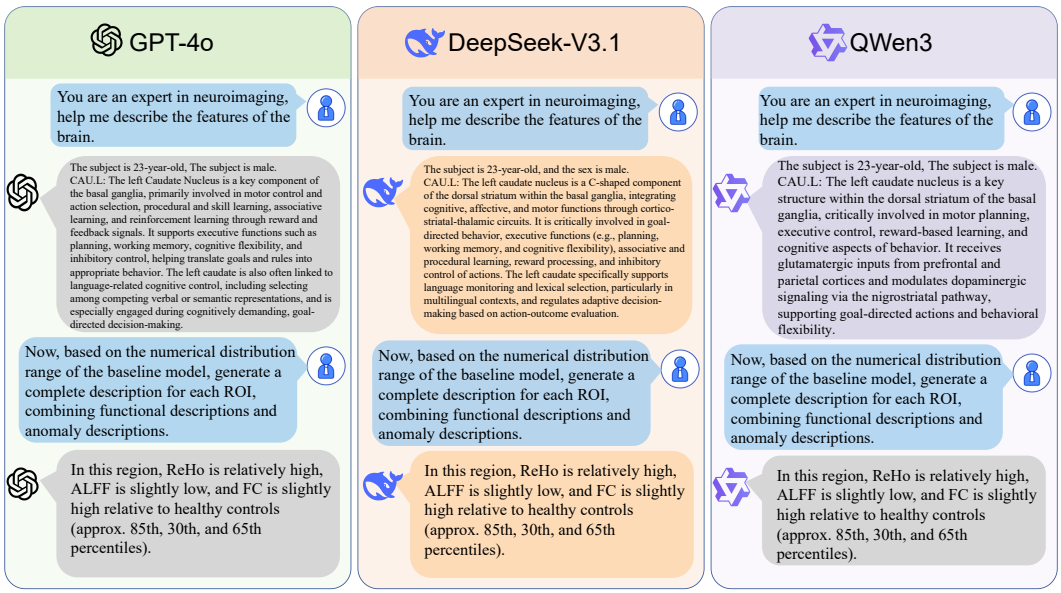


Figure 6: A case study of a brain region's answer.

We select the LLM with the highest $\bar{s}_m$ (and satisfactory agreement) as the default prior generator to be injected into the PTGA framework.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the core contributions (statistically grounded text priors, prior-guided gating, and multimodal fusion) and scope, and these claims are supported by the method and experiments (Sec. 1, Sec. 3, Sec. 4).

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses limitations including dependence on atlas-based ROI partitioning, distribution modeling choices, and sensitivity to preprocessing and LLM prior quality, and outlines future directions (Conclusion).

   Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper is primarily methodological and empirical; it does not present theorem–proof style theoretical results requiring formal assumptions/proofs.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide dataset descriptions, preprocessing pipeline, train/validation/test protocol, model architecture details, optimization settings, metrics, and repeated-run evaluation procedures (Sec. 4 and Appendix).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The study uses established datasets with cited sources, but full code/scripts are not publicly released in this submission version; release and detailed reproduction instructions will be provided in supplementary/final materials.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We report data split strategy, optimizer, learning rate, batch size, weight decay, epoch budget, early stopping, and hyperparameter sensitivity analysis needed to interpret and reproduce results (Sec. 4.1, Hyperparameter analysis).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report mean$\pm$std over 10 independent runs and perform paired statistical significance testing with FDR correction for key comparisons (Sec. 4, Table 3).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Hardware type is reported (NVIDIA RTX-Pro-6000), but full per-run wall-clock time and aggregate compute budget are not comprehensively listed and will be added in the revised version.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research follows NeurIPS Code of Ethics, including anonymized submission practices, transparent reporting, and responsible use of existing datasets/models.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses potential positive impacts (improved objective support for psychiatric assessment) and potential risks (bias, domain shift, and over-reliance in clinical settings), with caution on deployment scope (Conclusion/Discussion).

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: This work does not release high-risk generative models or scraped datasets that require dedicated controlled-access safeguards.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [No]

Justification: Original assets are cited, but an explicit per-asset license/terms table is not yet fully documented in the current manuscript and will be added in supplementary/final materials.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: The submission does not introduce a newly released dataset, model checkpoint, or software package as a new public asset at this stage.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The study does not involve crowdsourcing or new human-subject experiments; it analyzes previously collected de-identified datasets.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No new participant recruitment or intervention was conducted by the authors; analyses are based on existing datasets.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLMs are a core methodological component used to generate ROI-level clinical text priors from statistical anchors, which are then encoded and fused in PTGA (Sec. 3.2 and interpretability-related sections).

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.

- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.