Supplementary Materials

A. Detailed Explanations on fMRI Descriptors

This appendix summarizes the full set of imaging-based descriptors used in our analyses. For each metric, we provide both ROI-level and global-level definitions along with a brief description of the information captured by the feature. Most of the features are based on the Schaefer400 atlas with 7 functional networks: visual, somatomotor, dorsal attention, ventral attention, limbic, fronto-parietal control, and default-mode.

A.1. Functional Connectivity (FC) Descriptors

Functional connectivity captures the statistical synchronization between brain regions, reflecting how distributed neural populations co-fluctuate over time. It provides insight into large-scale communication patterns, network integration/segregation, and the overall organizational architecture of intrinsic brain activity.

ROI-Level

- **Network-Pair Connectivity**: Mean correlation strength between predefined network pairs (e.g., Default–Visual, Control–Limbic), z-scored relative to the cohort. *Captures directed pairwise interaction strength among large-scale systems.*

Global-Level

- **Top/Bottom Connectivity Patterns**: The three most elevated and three most reduced network-pair connections across the whole brain. *Summarizes extremes of functional integration and dissociation.*

Example:

The subject's functional connectivity profile reveals several notable patterns:- Among the most pronounced increases: SalVentAttn-SomMot coupling shows mild enhancement (z=+1.34), DorsAttn network exhibits heightened within-network connectivity (z=+1.38), and SalVentAttn-Vis interaction is moderately elevated (z=+1.18).- Regarding diminished connections: Cont-Limbic functional coupling falls within normative bounds (z=-0.85), DorsAttn-Limbic interaction remains in the typical range (z=-0.74), and Default-Limbic connectivity similarly shows no significant deviation (z=-0.64).

A.2. Functional Gradients Descriptors

Functional gradients describe the continuous, hierarchical organization of the cortex from low-level sensory processing to high-level transmodal cognition.

ROI-Level

- **Network Gradient Values**: Mean gradient values for each of the seven canonical networks (Control, Default Mode, Dorsal Attention, Ventral Attention, Limbic, Somatomotor, Visual). *Indexes each network's position along macroscale functional gradients.*

Global-Level

- **Principal Gradient Range**: Degree of separation between sensory and transmodal regions along the first gradient axis. *Reflects hierarchical organization of perception-to-association cortex.*
- **Second Gradient Range**: Extent of segregation among distinct sensory modalities. *Quantifies differentiation within sensory processing streams.*
- **Third Gradient Range**: Degree of distinction between control and association systems. *Captures higher-order specialization across cognitive control networks.*
- **Gradient Variance**: Overall dispersion/spread of functional gradients. *Represents global heterogeneity of cortical functional organization.*

Example:

The functional gradient profile for this subject reveals:- Principal gradient extent falls within typical limits (+0.3 SD)- Second gradient span remains within normative bounds (+0.3 SD)- Third gradient coverage aligns with typical variation (+0.2 SD)- All functional networks demonstrate principal gradient values within expected ranges.

A.3. Graph-Theoretic Descriptors

Graph metrics characterize the topological structure of the brain's functional network, quantifying how efficiently information flows and how modular or integrated the system is.

ROI-Level

- **Network Strength**: Weighted degree (sum of edge weights) for each of the seven networks. *Measures hub-like activity and overall connectedness of each network.*

Global-Level

- **Modularity**: Degree of segregation into distinct functional modules. *Measures community structure and functional specialization.*
- **Global Efficiency**: Capacity for parallel information transfer between distributed brain regions. *Indexes the overall integrative efficiency of the brain network.*
- **Average Clustering Coefficient**: Tendency for neighboring nodes to form tightly interconnected clusters. *Captures local segregation and community tightness.*

Example:

The brain's functional architecture exhibits the following graph theoretical characteristics:- Modularity: Network segregation patterns fall within the expected range relative to the reference population.- Global Efficiency: Information integration capacity across the entire brain appears typical.- Clustering Coefficient: Local connectivity clustering remains within normative bounds.- Regional Connectivity: At the nodal level, the Ventral Attention network demonstrates increased functional coupling strength.

A.4. ICA-Derived Descriptors

ICA features capture the temporal dynamics and intrinsic activity patterns of independent large-scale functional networks. The ICA is estimated with a pre-defined template, followed by cross-subject decomposition. We use another set of functional networks (different from the FC) to provide another aspect of functional information. These networks are: visual, default-mode, cognitive control, sensorimotor, subcortical, auditory, and cerebellum.

ROI-Level (Network-Level)

- **Network Temporal Amplitude**: Mean absolute activation magnitude within each ICA-defined network. *Represents the overall activity level of each network.*
- **Network Temporal Variability**: Standard deviation of temporal fluctuations in each network. *Captures dynamic stability vs. variability of network activity.*
- **Network Spectral Ratio**: Ratio of slow (0.01–0.1 Hz) to fast (0.1–0.25 Hz) oscillations. *Reflects dominant timescale of spontaneous network dynamics.*
- **Network Autocorrelation**: Lag-1 temporal coherence per network. *Measures persistence vs. rapid transitions in network activation.*
- **Network Transient Frequency**: Proportion of extreme activation events ($|z| > 3$). *Indexes occurrence of transient bursts or network "spikes".*
- **Network-Pair FNC**: Functional network connectivity among ICA-derived networks. *Represents coherence between independent components.*
- **Network fALFF**: Fractional amplitude of low-frequency fluctuations per network. *Measures relative contribution of spontaneous low-frequency oscillatory activity.*

Global-Level

- **Overall Network Engagement**: Mean absolute activation across all ICA networks. *Summarizes global activity level.*
- **Overall Temporal Dynamics**: Average variability of network activity. *Indexes brain-wide dynamic stability vs. fluctuation.*
- **Overall Spectral Balance**: Global ratio of slow to fast oscillatory power. *Reflects dominant timescale of whole-brain spontaneous activity.*

- **Overall Temporal Coherence**: Average lag-1 autocorrelation across networks. *Represents persistence of whole-brain states.*
- **Overall Transient Activity**: Frequency of intense activation bursts across the brain. *Measures prevalence of transient whole-brain activation events.*

Example:

The subject's independent component analysis reveals these functional network characteristics:- Temporal Dynamics: ICA timecourses demonstrate typical overall network engagement, standard temporal variability, prominent slow oscillation dominance, characteristic temporal coherence, and expected transient activity. At the network level, the Cerebellar system exhibits strong engagement in absolute mean amplitude and dynamic fluctuations in temporal variability. The Default Mode network displays slow oscillations in spectral ratio, while the Auditory network shows persistent activity in lag-1 autocorrelation. The Visual network demonstrates few transient events in outlier frequency.- Functional Coupling Patterns: Notable connectivity patterns include elevated coupling between Visual and Cognitive Control networks, within the Cerebellar network, and within the Default Mode network. Reduced connectivity is observed between Subcortical and Cognitive Control networks, and between Default Mode and Cerebellar networks.- Fractional ALFF: All networks show fractional amplitude of low-frequency fluctuations within typical ranges, suggesting balanced contributions across systems.

A.5. High-Level Semantic Descriptors

The high-level semantic descriptors cover the demographic, phenotypic, cognitive, and biomarker-related information of each subject. We listed the descriptors used in stage 3 in Sec. A.5 as well as what they measure and the available datasets.

Name	Meaning	Avail Datasets
Sex	Male/Female	All datasets
Age	age of subject in years	All datasets
BMI	Body Mass Index; an estimate of body fat based on weight relative to height.	UKB, HCP, HCP-A,
Blood Pressure	A measure of the force of blood against artery walls, given as diastolic values	UKB, HCP-A, ABCD
Cholesterol Level	Blood lipid profile indicating levels of total cholesterol; a marker of cardiovascular and metabolic health.	UKB, HCP-A
Fluid Intelligence Score	A measure of problem-solving ability, reasoning, and the capacity to think flexibly without relying on prior knowledge.	UKB
Fluid Composite Score	A combined measure of fluid cognitive abilities in the Human Connectome Project, capturing reasoning, abstraction, and novel problem solving.	HCP, HCP-A, ABCD
Flanker Score	An assessment of attention and inhibitory control; measures the ability to suppress distracting information.	HCP-A
APOE4 Status	Genetic marker indicating presence of the APOE $\epsilon 4$ allele, associated with increased Alzheimer's disease risk.	ADNI
AV45 (Florbetapir PET SUVR)	A PET imaging biomarker of β-amyloid deposition; higher values indicate greater amyloid burden.	ADNI
CDRSB (Clinical Dementia Rating – Sum of Boxes)	A clinician-rated measure of cognitive and functional impairment severity across multiple domains.	ADNI
MMSE (Mini-Mental State Examination)	A brief standardized test of global cognitive function, including memory, attention, and orientation.	ADNI
Verbal IQ	A measure of verbal reasoning, vocabulary knowledge, and language-based cognitive abilities.	ADHD200, ABIDE2
Performance IQ	A measure of nonverbal reasoning, visual–spatial processing, and perceptual problem-solving skills.	ADHD200, ABIDE2

B. Details on Datasets

The experimental targets for each dataset are detailed in Table 1, where "Cls" denotes a classification task and "Reg" denotes a regression task. In line with the methods described in the previous section, we discretized two continuous variables. The fluid intelligence status was created by binning the z-scores of fluid intelligence relative to the UKB dataset. Similarly, the fluid composite status was formed by discretizing the corresponding z-scores relative to the HCP and HCP-A datasets.

Table 1. Targets used in this study.

Dataset	Name	Task Type	Range/Possible Values
All Datasets	Sex	Cls	Male/Female
All Datasets	Age	Reg	Float, 10-100
All Datasets	Age Group	Cls	adolescent ($x < 18$), young adult ($18 < x < 30$), middle-aged adult ($30 < x < 60$), senior ($60 < x < 80$), elderly ($80 < x$)
UKB	Fluid Intelligence Score	Reg	Integer, 50-150
UKB	Fluid Intelligence Status	Cls	higher than usual ($z > 1.5$),average ($-1.5 < z < 1.5$), lower than usual ($z < -1.5$)
HCP,HCP-A	Fluid Composite Score	Reg	Integer, 50-150
HCP,HCP-A	Fluid Composite Status	Cls	higher than usual ($z > 1.5$),average ($-1.5 < z < 1.5$), lower than usual ($z < -1.5$)
ADNI	Alzheimer's Diagnosis	Cls	Cognitive Normal, Mild Cognition Impairment, Alzheimerś
ADNI	APOE4 status	Cls	APOE4 positive, APOE4 negative
ADHD200	ADHD Diagnosis	Cls	Control, ADHD
ABIDE2	Autism Spectrum Disorder Diagnosis	Cls	Control, Autism

C. Detailed Experimental Settings

Table 2. Hyperparameters of stage 1 training.

Hyperparameters		Values
fMRI Tokenizer (small/base/large)	Transformer encoder layers	12/12/24
	Transformer decoder layers	12/12/24
	Patch size	32/32/32
	Embedding dimension	384/768/1024
	Num heads	6/12/16
	Codebook size	8192
Batch size		8
Learning rate		1e-4
Minimal learning rate		1e-5
Learning rate scheduler		Cosine
Optimizer		AdamW
Total epochs		50

Table 3. Hyperparameters of stage 2 training.

Hyperparameters	Value
Batch size	16
Learning rate	6e-4
Minimal learning rate	6e-5
Learning rate scheduler	Cosine
Optimizer	AdamW
Total epochs	25
Gradient clip	1.0
Deepspeed stage	stage 2

Table 4. Hyperparameters of stage 3 training.

Hyperparameters	Value
Batch size	32, 128 (zero-shot)
Learning rate	5e-4
Minimal learning rate	5e-5
Learning rate scheduler	Cosine
Optimizer	AdamW
Total epochs	20
Gradient clip	1.0
Deepspeed stage	stage 2

D. Instruction Tuning Tasks and Paradigms

D.1. Single-question Single-answer Paradigm

> ### Question: {QUESTION}. ### Answer: {ANSWER}

Here, `QUESTION` depends on the dataset and target variable. For example, for UK Biobank (UKB) sex prediction:

> ### Question: What is the sex of this subject? ### Answer: Male

To enhance robustness and reduce overfitting to specific phrasings, 200 distinct rewrites of each `QUESTION` type are generated to increase linguistic diversity.

D.2. Multi-question Multi-answer Paradigm

For a single fMRI scan, multiple attributes may be queried simultaneously, each requiring an independent answer. The general template is:

> ### Question: {QUESTION1}, {QUESTION2}. ### Answer: {ANSWER1} {ANSWER2}

The number of questions and answers is flexible. For example, in UKB sex and fluid-intelligence prediction:

> ### Question: What is the sex of this subject? How about its fluid intelligence status? ### Answer: Male | Below average

Answers are separated using the "|" delimiter. Question rewrites also include fused forms in which multiple questions are asked within a single sentence:

> ### Question: Can you specify the subject's sex and level of fluid intelligence (lower, mean, higher) from the fMRI scan? ### Answer: Male — Below average

D.3. Open-Ended Question Paradigm

In the open-ended setting, a broad question is posed without restricting the answer to any predefined set of targets. An example is:

> ### Question: What specifics about the subject can you infer from the fMRI scan, such as demographics, cognitive abilities, or disease information? ### Answer: This is a senior male subject who is likely to have Alzheimer's disease, with a positive APOE4 biomarker.

To quantitatively assess correctness, a structured target schema is used, such as:

{Sex: Male; Age Group: Senior; AD Diagnosis: AD; APOE4: Positive}

This structured reference is provided to an LLM-based evaluator, which scores each field for correctness. An overall correctness score is additionally computed, defined as correct only if all fields match the ground truth.

E. Results on Discretized Targets

To better align with the language-modeling objective and enable more robust classification capabilities, several continuous or composite targets (e.g., age, fluid intelligence, and flanker performance) were discretized into categorical groups. Table 5 reports the performance of fMRI-LM compared with two strong baselines (SWiFT and Brain-JEPA) across these discretized tasks.

Table 5. Performance of fMRI-LM (base, GPT2) and baselined trained on discretized targets.

	UKB (age group)		UKB (fluid intel status)		HCP (fluid comp status)		HCP-A (age group)		HCP-A (fluid comp status)		HCP-A (flanker status)	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
SWiFT	80.26	63.25	81.82	30.24	77.95	37.24	55.28	39.05	96.21	88.29	83.22	61.05
Brain-JEPA	81.46	64.98	80.12	35.29	78.29	30.25	70.19	48.74	93.28	88.18	88.91	62.19
fMRI-LM	81.58	65.24	82.95	37.54	85.42	38.12	69.52	48.29	97.10	89.12	89.65	64.26

F. Results on Targets Independently Trained per Dataset

Across the datasets used in this study, several targets are shared or aligned across sites. For example, UKB, HCP, and HCP-A all include sex classification; the UKB fluid-intelligence status corresponds closely to the fluid-composite status in HCP and HCP-A; and the diagnostic labels for control subjects overlap across ADNI, ADHD200, and ABIDE-II. Such overlap introduces potential sources of interference when these targets are trained jointly in a multi-task setting—particularly when datasets differ in sample size, demographic composition, or preprocessing pipelines.

To isolate the effect of multi-task training from dataset-specific signal and to better understand how fMRI-LM performs on each target, we train separate models independently on each dataset–target pair and compare them with the jointly trained multi-target model. This experiment allows us to assess (i) whether joint training introduces positive transfer, negative interference, or neither, and (ii) how consistently fMRI-LM generalizes when trained on focused versus shared objectives.

Table 6 summarizes the performance differences between the jointly trained model and the independently trained models across all targets. Results are reported in terms of accuracy and AUC (when available). In the table, joint training generally maintains or slightly improves performance for UKB and HCP fluid-related tasks, suggesting positive transfer from shared cognitive phenotypes. However, for HCP-A, independent training yields higher accuracy in fluid-composite and flanker tasks, indicating dataset-specific specialization may be beneficial for older-adult cohorts. For clinical datasets, joint training provides modest improvements for ADNI and ABIDE-II, whereas ADHD200 shows minimal difference between settings. These results suggest that diagnostic targets benefit from shared representational features learned across diverse datasets, despite label heterogeneity.

Overall, the results demonstrate that multi-task joint training rarely harms performance, improves several clinical and cognitive targets, and preserves competitive performance on sex classification across datasets.

Table 6. Performance of fMRI-LM (base, GPT2) trained independently on each target and jointly on all targets.

	UKB (sex)		UKB (fluid intel status)		HCP (sex)		HCP (fluid comp status)		HCP-A (sex)	
	Acc	AUC	Acc	AUC	Acc	AUC	Acc	AUC	Acc	AUC
Joint	94.89	94.90	82.95	-	89.58	89.13	85.42	-	89.58	88.98
Independent	94.72	94.73	83.03	-	78.74	79.14	85.83	-	86.81	87.01
	HCP-A (fluid comp status)		HCP-A (flanker status)		ADNI (AD)		ADHD200 (ADHD)		ABIDE2 (ASD)	
	Acc	AUC	Acc	AUC	Acc	AUC	Acc	AUC	Acc	AUC
Joint	97.10	97.25	89.65	89.74	77.92	79.91	72.92	68.72	65.97	68.72
Independent	95.41	96.02	92.11	93.05	70.83	71.52	71.35	70.15	59.90	70.21

G. Further Results on Ablations

G.1. Ablation on weights of stage 1 and 2's loss terms

Here we present the performance of fMRI-LM on different loss weights, including λ **??** and α, β of **??**. The results are in Sec. G.1.

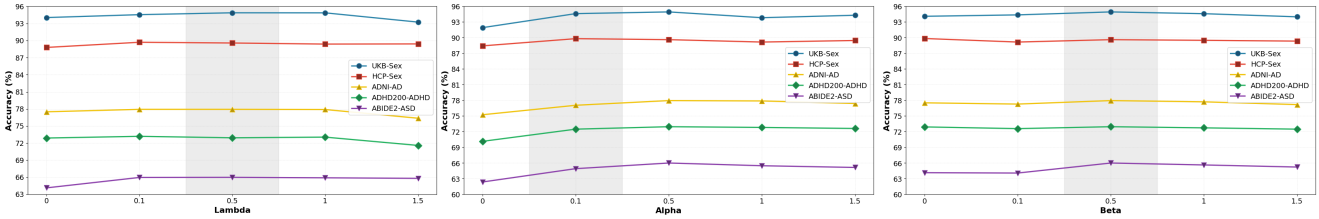

Figure 1. Ablations on the weights of loss terms in stage 1 and 2. The values in the main experiments are marked in gray. The Changes in values of λ, α, β have no notable effect in the final performance, while changing them to 0 can have a negative effect.

G.2. Ablation on different sizes of fMRI tokenizers and LLMs

We further show the performances using different sizes of fMRI tokenizers (small, base, large) and LLMs (GPT2, from 124M to 1.5B). The results are in Sec. G.2.

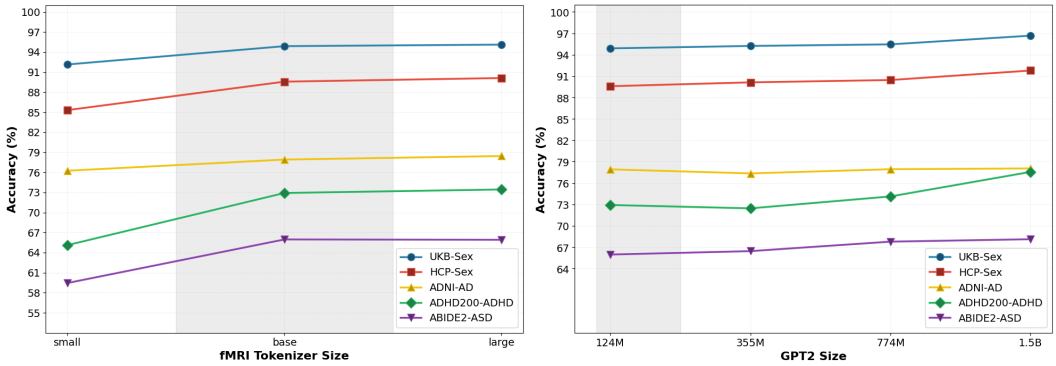

Figure 2. Compare performance with different sizes of fMRI tokenizers and GPT2 model.

H. Limitations and Future Work

In this work, we align a universal fMRI representation with a large language model (LLM). To compensate for the lack of natural fMRI–text pairs, we construct large-scale fMRI descriptors derived from imaging-based features and fine-tune the LLM to interpret these representations for diverse downstream applications. A key limitation is that we do not yet fully capture or evaluate the model's reasoning process over fMRI data. Ideally, the LLM would be able to explicitly identify the fMRI substrates (e.g., specific functional networks or regions) that support a given diagnosis or prediction, providing more transparent and mechanistic explanations. Achieving this will require more rigorous construction of instruction-tuning datasets, including carefully curated and validated annotations by domain experts, to minimize factual errors and support reliable, interpretable reasoning over brain signals. On the other hand, such imaging biomarkers are not fully revealed in the medical literature, so it will be difficult to examine the correctness of the reasoning.