Foundation models for linking brain functional representations and clinical outcomes across local intervention studies targeting cognitive aging: Validity and generalizability

Xinliang Zhou,^ Meishan Ai,^ Ehsan Adeli, Yu Zhang, F. Vankee-Lin,* Yang Merik Liu*

Abstract

The direct cognitive benefits of interventions targeting cognitive aging highly vary across individuals due to individual heterogeneity in aging-related comorbidities. This variation necessitates a better characterization of brain representations reflecting how older individuals uniquely benefit from interventions, thereby revealing the true effect of these interventions. Resting-state functional MRI (rsfMRI) has potential to uncover meaningful brain representation; however, traditional methods analyzing summary features derived from rsfMRI data are limited in capturing both within-individual variations across timepoints and between-individual difference, particularly in small and heterogeneous local intervention studies. Recent rsfMRI foundation models pretrained on large-scale observational cohorts, such ask UK Biobank, offer a potential alternative, yet their validity and generalizability in local intervention studies remain unclear. In this study, we systematically evaluated existing rsfMRI foundation models for classifying intervention responses to changes of episodic memory using data from two independent randomized controlled trials targeting cognitive aging among older adults with mild cognitive impairment. RsfMRI foundation models outperformed conventional machine learning and deep learning approaches, while performance differed across foundation models with different pretraining objectives. Clinically informed fine-tuning using an external Alzheimer's disease cohort further improved performance and demonstrated robustness to confounders (e.g., head motion during MRI data acquisition, data collection site, intervention arm). Interpretability analyses confirmed that brain representations learned by the model reflect the brain networks supporting cognitive aging or episodic memory. Together, these findings provide an empirical evaluation of current rsfMRI foundation models and support nuanced interpretation of heterogeneous intervention outcomes by linking individual differences in longitudinal brain dynamics to cognitive change.

1. Introduction

As the global population ages, developing and testing scalable and effective interventions aimed at managing and preventing cognitive decline with minimal side effects, such as cognitive training and aerobic exercise, have become a focus in clinical research[1,2]. However, the direct cognitive benefits of these interventions highly vary across individuals due to individual heterogeneity in aging-related comorbidities[3–5]. This variation necessitates a better characterization of brain representations reflecting how older individuals uniquely benefit from interventions (a.k.a., neuroplastic potential), thereby revealing the true effect of these interventions.

Resting-state functional magnetic resonance imaging (rsfMRI) provides an objective approach to examine individuals' neuroplastic potential[6,7]. However, older adults, especially those at risk for dementia, often display wider variation within individuals across different timepoints as well as between individuals in BOLD signals compared to younger, healthier adults. Factors contributing to this variance include head motion during data acquisition and intrinsic influences from biological aging[8–10]. Substantial progress has been made in applying statistical and machine learning (ML) approaches to rsfMRI, including network-based analyses, multivariate pattern analysis, and clinical phenotype modeling, partially accounting for inter-individual heterogeneity[11,12]. However, most statistical methods and ML techniques primarily depend on pre-defined or aggregated features[13] such as functional connectivity (FC)[14], which may overlook hidden, granular neural representations. This issue is amplified in local intervention studies, where subtle treatment related effects may be embedded in detailed, time-specific BOLD signals but are partially collapsed when summarized into static FC features. Conventional deep learning (DL) approaches usually require large datasets that are often unavailable in real-world clinical settings where local intervention studies take place[15]; limited data, on the other hand, may lead to oversimplification of temporal dependencies of overfitting to noisy BOLD signal when applying DL approaches[16].

Emerging foundation models (FMs) offer an alternative to address these challenges. FMs are large-scale, pre-trained models that serve as a base for a wide range of downstream tasks[17–19]. Typically trained on extensive datasets, they learn broad and generalizable representations that can be fine-tuned or adapted for specific applications. In the context of rsfMRI, FMs can be utilized to understand individual variations in response to interventions, creating the possibility of extracting rich brain representations directly from time-series BOLD signals, rather than relying only on summary patterns like static FC. Their massive pre-training provides a powerful prior, allowing them to learn robust representations of temporal dynamics without overfitting, while their architecture (often Transformer-based[20]) is inherently designed to capture the complex, long-range dependencies within a time-series that are lost when signals are aggregated into static FC patterns. By leveraging pre-training on thousands of hours of rsfMRI data from large, cross-sectional cohorts, rsfMRI FMs are expected to learn rich, generalizable neural representations that capture fine-grained spatial and temporal patterns of brain activity[21,22]. However, a systematic evaluation of the validity and generalizability for adapting rsfMRI FMs to local datasets, especially those local intervention studies targeting aging populations, remains a highly underexplored area.

In this paper, we addressed three questions regarding the use of rsfMRI FM in local intervention studies targeting cognitive aging: (Q1) Can rsfMRI FMs' embeddings improve the downstream models to learn from small and heterogeneous local intervention cohorts with limited labels? (Q2) Can FMs' pretrained spatiotemporal representations capture subtle between baseline (T1) and post-intervention (T2)

differences in BOLD signals despite most existing rsfMRI FMs being developed and validated on cross sectional tasks? (Q3) How well do rsfMRI FMs transfer to evaluate cognitive aging interventions under domain shift, where participant characteristics, MRI scanning protocols, and outcome definitions differ substantially from the pretraining data? To answer these questions, we evaluated two representative rsfMRI FMs: BrainLM[21], trained to model whole-brain activity as an inference sequence similar to a language model, and BrainJEPA[22], designed to learn shared embeddings from neighboring brain states. Both rsfMRI FMs were originally pre-trained on large-scale datasets, such as UK Biobank[23] and the Human Connectome Project (HCP)[24], which predominantly consist of younger, neurologically healthy adults. We further fine-tuned them using Alzheimer's Disease Neuroimaging Initiative (ADNI)[25] dataset to ensure the brain representations reflect the range of variability covering both normal and abnormal cognitive aging. We then tested the models on two local intervention studies targeting cognitive aging: ACT[26] study (NCT03313895), a 3-site randomized control trial (RCT) with four arms over 6-month exercise and cognitive training combined, alone, or active control, and CogTE[27] study (NCT02559063), a single-site RCT with two arms over 6-week cognitive training or active control. Both studies collected data on episodic memory (EM), the most vulnerable domain in both normal and abnormal cognitive aging, and rsfMRI at baseline and post intervention, targeting older adults with mild cognitive impairment (MCI). Across analyses, change in EM (ΔEM) was calculated as the difference between post-intervention and baseline, with responders defined as sustained/improved post-intervention from baseline (ΔEM>=0) vs. non-responders as declined cases (ΔEM<0). Upon addressing these questions, we also explored the interpretability of the best-performing rsfMRI FM. The overview of analytical flow is presented as **Figure 1**.

2. Result

2.1 Classification of change of episodic memory after intervention: Benchmark comparisons

To address Q1 of whether rsfMRI FMs can outperform conventional methods, a comprehensive benchmark was conducted. Evaluation was performed under stratified cross-validation, balancing outcome labels and key study factors (intervention arm and site when applicable). The performance of two FMs, BrainLM and BrainJEPA, was systematically compared with three widely used data-driven approaches: Support Vector Machine (SVM)[28], a ML method using statistical learning theory; Multilayer Perceptron (MLP)[29], a DL method capturing nonlinear relationships; and Graph Convolutional Network (GCN)[30], a DL method explicitly incorporating brain topology. These models were chosen to represent the spectrum of existing strategies commonly applied to classify outcomes, i.e., responders vs. nonresponders derived from ΔEM in our case. This evaluation was performed on ACT and CogTE respectively; sample characteristics and protocol details are presented in **Supplementary Materials Table S1**. All models were trained using rsfMRI collected at T1 and T2. For non-FM methods (SVM, MLP, and GCN), FC features were derived separately at each timepoint and concatenated to encode longitudinal information. This design follows standard practice in conventional rsfMRI modeling[31,32], where FC provides a stable, low-dimensional representation suitable for vector- or graph-based classifiers that do not operate directly on raw time-series data. The rsfMRI FMs used regional-averaged BOLD time-series as input, which were temporally segmented into fixed-length patches and embedded as tokens. See details in **Method**.

Figure 2a and **Supplementary Materials Table S2** summarize the model performance across benchmark comparisons. BrainLM outperformed other models, particularly on the CogTE dataset (*ACC: 80.3% ± 1%, F1: 88.2% ± 1%*). The performance advantage of BrainLM was also demonstrated on the multi-site ACT dataset (*ACC 65.2% ± 5%, F1 69.1% ± 5%*) compared to other models. In contrast, BrainJEPA did not achieve good performances across studies (*CogTE: ACC 60.4% ± 11%, F1 64.1% ± 7%; ACT: ACC 50.8% ± 13%, F1 53.4% ± 8%*). The performance difference between the two rsfMRI FMs may be explained by the difference of their representation learning manners: BrainJEPA relies on a built-in cross-view inference module that enforces consistency between neighboring spatiotemporal brain states during pre-training, which potentially overlooked the variability seen in older adults' data; whereas, BrainLM's autoregressive sequence modeling objective accommodate heterogeneous temporal patterns, making it more robust to modeling the neural dynamics among older adults. Overall, our findings support the promise of selected rsfMRI FMs for decoding subtle neural signals in local intervention studies, while emphasizing that careful adaptation and validation are needed before deployment.

2.2 Domain adaptive fine-tuning on ADNI dataset

To further examine whether the performance of rsfMRI FMs could be improved through domain adaptation (Q2), we next evaluated the impact of a clinically informed fine-tuning on an external aging-related dataset. Specifically, the best-performing BrainLM was fine-tuned on ADNI dataset using a supervised classification task distinguishing group with Alzheimer's disease (AD) from healthy control (HC), named as BrainLM-ADNI. The design was motivated by the assumption that intervention-related EM improvement reflects, at least in part, a reversal or modulation of neural patterns of EM difference between AD and HC, making the HC-to-AD contrast a biologically meaningful axis for adaptation. Moreover, the differences between AD and HC are more neurobiologically distinct and stable than those between preclinical AD, such as MCI, and HC, offering a stronger supervisory signal for domain adaptation.

The resulting BrainLM-ADNI model was then evaluated on the ACT and CogTE datasets using the same downstream inference protocol as in **Section 2.1**. As shown in **Figure 2b**, BrainLM-ADNI demonstrated a substantial performance boost, particularly on the more challenging ACT dataset (*CogTE: ACC 81.6% ± 5% [↑1.3%], F1 88.8% ± 3% [↑0.6%]); ACT: ACC 72.6% ± 13% [↑7.4%], F1 83.1% ± 10% [↑14%]*), which suggests that the knowledge about aging-related brain patterns learned from ADNI was successfully transferred, equipping the model with a more refined understanding of the neural basis and its response to intervention, supporting our hypothesis. The consistent improvement across both datasets provides evidence that a clinically relevant, supervised fine-tuning step may still be critical for existing rsfMRI FMs when applying to a specialized clinical study.

To further characterize how the ADNI-adapted BrainLM leverages longitudinal information, we additionally evaluated model performance under different combinations of timepoints. Specifically, we compared the default "*T1+T2*" setting with a "*T1-only*" setting, in which downstream classification was based solely on baseline rsfMRI embeddings, and a "*T2–T1*" setting, in which residuals between post-intervention and baseline rsfMRI embeddings were used as input. As shown in **Figure 2b**, across two datasets, the default "*T1+T2*" configuration achieved the strongest and most consistent performance (*CogTE: ACC 81.6% ± 5%, F1 88.8% ± 3%, ACT: ACC 72.6% ± 13%, F1 83.0% ± 10%;*), indicating that jointly modeling baseline and post-intervention brain states provides complementary information

beyond either baseline (*CogTE: ACC 68.3% ± 6%, F1 76.5% ± 5%, ACT: ACC 62.3% ± 11%, F1 70.4% ± 9%,*) or isolated change signals (*CogTE: ACC 72.0% ± 6%, F1 79.3% ± 5%, ACT: ACC 64.2% ± 5%,, F1 72.2% ± 9%,*) alone.

2.3 Robustness to neuroimaging confounders

Q3 examined the robustness of the adapted rsfMRI FM to common neuroimaging confounders, specifically inter-site variability and participant head motion. Using the multi-site ACT dataset, we evaluated the performance of BrainLM-ADNI under three preprocessing conditions: (1) original data, (2) data after site harmonization, and (3) data after both site harmonization and regression of head motion parameters. As shown in **Figure 2c**, explicit correction for these confounders resulted in only modest differences in classification performance (*original: ACC 72.6% ± 13%, F1 83.0% ± 10%; site-harmonized: ACC 73.2% ± 7%, F1 81.0% ± 8%; site + head-motion corrected: ACC 73.8% ± 8%, F1 82.7% ± 6%*). Overall, these results indicate that BrainLM-ADNI maintains stable performance across varying levels of confound control, suggesting robustness to site-related heterogeneity and motion artifacts.

Additionally, we examined whether classification performance differed across intervention arms (see **Figure 2d**) in both studies. We performed Chi-square test on intervention group assignment and classification performance (i.e., whether being correctly classified or not) for each participant. The classification performance displayed no significant difference across intervention arms for neither study (*CogTE:* $\chi^2=0.01$, $p=0.92$; *ACT:* $\chi^2=5.44$, $p=0.14$). These results suggest that BrainLM-ADNI maintains comparable classification performance across intervention arms, further supporting its robustness.

2.4 Biological interpretability of rsfMRI FM embeddings

Representational Similarity Analysis (RSA) was used here to assess whether inter-individual variability in learned model representations aligns with variability in region-wise BOLD activity[33]. For each timepoint (i.e., T1 and T2) and local intervention studies (i.e., ACT and CogTE), a similarity matrix was constructed from both the BrainLM-ADNI embedding space by computing pairwise inter-individual correlations, and separately for each brain region. For each brain region, the upper-triangular elements of the embedding and regional BOLD activity similarity matrix were vectorized and compared using Spearman rank correlation, yielding a measure of correspondence between representational geometry in the learned embedding space and regional BOLD activity across participants. Regions showing relatively higher correspondence (FDR-corrected $p<0.05$) were identified for each timepoint within each local intervention study. Brain regions exhibiting consistent correspondence across timepoints with studies, as well as across studies at each timepoint, are displayed in **Figure 2e**. At T1, regions showing high representational similarity were predominantly located in the visual cortex and superior parietal gyrus across both studies, corresponding to core components of the visual and dorsal attention networks, whose functions are closely related to episodic memory encoding and are commonly affected in early stages of cognitive decline. At T2, the pattern of high-similarity regions shifted toward the superior parietal gyrus, precuneus, and anterior cingulate cortex, which are key nodes of the dorsal attention and default mode networks. Importantly, overlapping regions exhibiting consistently high similarity across both ACT and CogTE studies were observed at each timepoint, indicating that the model captured reproducible neural patterns related to EM despite differences in intervention duration, modality, and study design. These findings are consistent with prior literature linking EM or cognitive aging to the default-mode and dorsal

attention networks[34,35]. Collectively, these findings support the interpretability of BrainLM-ADNI embeddings and indicate that the model leverages biologically meaningful neural characteristics when classifying intervention responses, providing a biologically grounded basis for relating individual brain changes to observed variability in cognitive intervention outcomes.

3. Discussion

3.1 Summary of key findings across research questions

Across two independent cognitive aging intervention datasets, rsfMRI FMs, particularly BrainLM, consistently outperformed well-established ML and DL approaches, including SVM, MLP, and GCN, despite the small sample sizes and high between- and within-individual heterogeneity inherent to intervention data. This performance advantage suggests that large-scale pretraining enables FMs to extract distributed and subtle spatiotemporal neural signals that are difficult to capture using FC or models trained from scratch. The performance advantage of BrainLM generalized across two intervention datasets with distinct durations, modalities, and study designs, indicating that pretrained rsfMRI representations capture neural features relevant to intervention responsiveness rather than task-specific artifacts. However, not all FMs generalized equally well: BrainJEPA underperformed relative to BrainLM, highlighting that architectural objectives and inductive biases critically shape downstream transferability. Importantly, supervised fine-tuning on the ADNI dataset further improved BrainLM's performance, particularly on the more challenging multi-site ACT dataset, demonstrating that clinically informed domain adaptation remains essential for bridging the gap between large-scale pretraining data on younger, healthier cross-sectional observational cohorts and small-scale intervention samples of older adults with cognitive impairment. Performance of the adapted BrainLM-ADNI remained stable under site harmonization and head motion regression and did not differ significantly across intervention arms in either dataset. This robustness suggests that large-scale pretraining on heterogeneous datasets may provide rsfMRI FMs with an inherent tolerance to common acquisition-related confounders, reducing (but not eliminating) the dependence on aggressive preprocessing strategies traditionally required in small-scale neuroimaging studies.

3.2 Implications of rsfMRI foundation models in local intervention studies

Our findings suggest that the advantage of rsfMRI FMs in local intervention studies arises from their ability to learn distributed, high-dimensional representations of whole-brain dynamics that go beyond predefined connectivity features. By leveraging large-scale pretraining, models such as BrainLM appear capable of encoding subtle spatiotemporal patterns that are difficult to recover when training models from scratch on small, heterogeneous intervention datasets. This mechanism is particularly relevant for aging populations, where intervention effects are often modest, temporally diffuse, and embedded within substantial inter-individual variability. From a clinical research perspective, this capacity enables brain-based characterization of individual differences in intervention response beyond group-average effects, a long-standing challenge in cognitive aging interventions. Importantly, our results also indicate that large-scale pretraining alone is insufficient for direct zero-shot transfer to intervention settings. Existing rsfMRI FMs are predominantly trained on cross-sectional data (e.g., UK Biobank) to characterize brain states within a single timepoint, without explicit supervision on longitudinal change. The substantial performance gains achieved through ADNI-based fine-tuning suggest that clinically informed domain adaptation acts as an essential intermediate scaffolding, aligning general neural representations learned

from healthy populations with aging- and pathology-related axes of variation. This alignment enables the model to bridge differences in acquisition protocols, population characteristics, and task objectives that would otherwise limit generalization. Finally, these insights should be interpreted in the context of reproducibility constraints. In this study, both BrainLM and BrainJEPA were evaluated using publicly available checkpoints (13-million and 22-million weights respectively) rather than the maximal-scale pretraining configurations described in their original reports. As a result, the lack of access to maximal-scale pretraining datasets and full model weights complicates replication and makes cross-study comparisons sensitive to factors beyond architectural design alone. Consequently, empirical evaluations using publicly available checkpoints, as in the present study, provide a realistic assessment of current deployable performance rather than theoretical upper bounds.

3.3 From findings to practice: recommendations and emerging trends

Building on the insights discussed above, we summarize a set of operational considerations for deploying rsfMRI FMs in local intervention studies. These considerations are directly relevant to clinical researchers designing and interpreting small-scale aging interventions, where limited sample sizes and heterogeneous responses often constrain drawing brain representation conclusions from neuroimaging measures. **Table 1** synthesizes key clinical constraints, corresponding technical solutions, and associated engineering effort, providing a concise implementation guide for translating FMs' advances into local intervention studies. Noticeably, recent work in rsfMRI FMs further indicates a diverse architectural and input choices. Beyond sequence-based models operating directly on voxel- or parcel-wise BOLD time series, emerging rsfMRI FMs increasingly explore graph-based formulations that explicitly encode brain network topology[36]. These models also vary in their choice of input representation, ranging from raw BOLD signals to static or dynamic FC matrices, each imposing distinct inductive biases on how neural interactions are modeled. BOLD-based models emphasize fine-grained temporal dynamics and non-stationary patterns, whereas FC-based or graph-based approaches prioritize relational structure and network-level organization. This heterogeneity reflects an active area of exploration, suggesting that no single representation is universally optimal and that model suitability may depend on the target population, task, and temporal scale of interest. In parallel, broader trends in FM research point toward the increasing importance of general-purpose and multimodal models (e.g., GPT[37], LLaVA[38], Qwen[39]) capable of integrating information across domains[40]. For brain aging and local intervention studies, such models offer a complementary pathway rather than a replacement for rsfMRI FMs. While rsfMRI FMs provide biologically grounded encodings of neural dynamics, general-purpose FMs can contribute higher-level semantic structure by aligning neuroimaging representations with behavioral, clinical, and linguistic information[17,41]. This cross-modal alignment has the potential to enhance interpretability, improve generalization across intervention contexts, and partially offset the data scarcity that currently limits purely neuroimaging-based models[42].

Of note, we used the present study as a showcase for demonstrating how to appropriately use emerging FM methods to dentify under-discovered, robust brain representations reflecting cognitive changes after interventions among aging populations. However, rsfMRI FM embeddings could also be used for other clinical scenarios, such as trial enrichment and stratified randomization by identifying participants most likely to show meaningful cognitive improvement, thereby increasing statistical power and reducing required sample sizes in small, heterogeneous cohorts. The same representations may also serve as a robust monitoring signal to track within person change across time, enabling earlier readouts of whether a

participant is on a favorable trajectory before long term outcomes are available. Finally, while we did not test between-group treatment effects in this work, our framework provides a foundation for future studies of treatment effect heterogeneity, where FM based representations can be evaluated as moderators to determine which individuals benefit more from specific interventions.

4. Methods

4.1 Overview of the analytical framework

To systematically evaluate the feasibility of rsfMRI FMs for classifying intervention responses in brain aging local intervention studies, we designed a multi-component framework (**Figure 1**). For each participant, rsfMRI data (both T1 and T2) were processed in a dual-stream architecture with shared model weights to extract latent embeddings for each timepoint. These embeddings were subsequently concatenated as input features to explicitly encode longitudinal neural change and classify intervention responses. To ensure a fair and controlled evaluation, all downstream analyses were conducted within a stratified five-fold cross-validation framework that balanced outcome labels, intervention arms, and study site when it is applicable. This design enabled direct comparison between FMs and conventional machine learning and deep learning approaches under identical data splits and evaluation protocols. Model performance was primarily assessed using Accuracy (ACC) and F1-score, reflecting overall correctness and the balance between precision and recall, respectively. To provide a complementary assessment of discriminative ability and class-specific performance, Area Under the ROC Curve (AUC), Sensitivity, and Specificity were also reported (see results in **Supplementary Materials Table S2**). Performance metrics were averaged across cross-validation folds, and the same evaluation protocol was applied consistently to all ML, DL, and rsfMRI FM approaches.

4.2 Study cohorts and preprocessing

Three independent datasets were included in this work for model deployment and evaluation. The ADNI dataset was used exclusively for domain-adaptive fine-tuning of the FM. The ADNI dataset comprises older adults spanning the cognitive spectrum from cognitively normal, MCI, to AD. In this study, ADNI participants diagnosed with AD and cognitively healthy controls were selected to construct a supervised fine-tuning task that captures aging- and pathology-related brain patterns. No ADNI data were included in downstream model evaluation. The ACT study is a multi-site randomized controlled trial involving older adults at risk for AD, with participants assigned to aerobic exercise, cognitive training, combined intervention, or active control arms over a six-month period. Resting-state fMRI and episodic memory assessments were collected at T1 and T2. The CogTE study is a single-site randomized controlled trial with a six-week cognitive training intervention and an active control condition, following a similar pre-post assessment design. Both ACT and CogTE datasets target older adults with elevated risk for cognitive decline and serve as independent cohorts for evaluating model generalizability across intervention paradigms and study designs (see **Figure 3a**). Both datasets exhibited substantial inter-individual variability in EM change, with samples of responders and non-responders in each intervention arm (see **Supplementary Materials Table S1**).

MRI acquisition details for ADNI, ACT, and CogTE have been reported previously and are summarized in **Supplementary Materials Table S1**. All imaging data were preprocessed using the standardized *fMRIprep* pipeline[43] (version 24.1.0), which is based on *Nipype* v1.8.6. For each participant, T1-weighted

anatomical images underwent skull stripping, tissue segmentation, surface reconstruction, and spatial normalization to Montreal Neurological Institute (MNI) space. RsfMRI images were corrected for head motion and slice timing, co-registered to the anatomical image, and normalized to MNI space. Susceptibility distortion correction was applied for the ACT dataset using estimated field maps from two echo-planar imaging references with *TOPUP*[44] that were available only for this study. Following *fMRIPrep*, additional denoising was performed using *Nilearn* (version 0.11.0). This included regression of twelve head motion parameters, five principal components derived from white matter and cerebrospinal fluid signals using the *aCompCor* approach, and temporal band-pass filtering (0.01–0.1 Hz). Finally, spatial smoothing was applied using a 6-mm full-width-at-half-maximum Gaussian kernel. Identical preprocessing procedures were applied across all datasets to minimize confounding differences arising from data handling and to ensure consistency in downstream model evaluation.

4.3 Resting-state fMRI foundation model deployment

BrainLM is designed to model whole-brain resting-state activity as an inference sequence, analogous to autoregressive language modeling. During pretraining, BrainLM learns to predict future spatiotemporal patterns of brain activity conditioned on past observations, enabling it to capture long-range temporal dependencies and distributed neural dynamics. BrainLM operates directly on preprocessed rsfMRI time-series data, where voxel- or parcel-wise BOLD signals are temporally segmented into fixed-length patches of 20 datapoints and embedded as input tokens to the transformer encoder. Although three pretrained variants (13-/111-/650-million weights) were described in the original paper, only the 13-million weight model is publicly released and was therefore used in this study.

BrainJEPA follows a joint embedding inference architecture. Its primary objective is to learn shared latent representations by enforcing consistency between embeddings derived from neighboring spatiotemporal views of the brain. BrainJEPA is trained to align representations of partially observed brain states, encouraging invariance across local temporal or spatial perturbations. As a result, it emphasizes representation consistency and local coherence rather than explicit temporal forecasting. BrainJEPA takes preprocessed rsfMRI time-series as input, formatting the signals into sequences of 160 time points and segmenting them into non-overlapping patches of size 16. Although three pretrained variants (22-/86-/307-million weights) were described in the original paper, only the 22-million weight model is publicly released and was therefore used in this study.

To deploy these rsfMRI FMs to local intervention studies without retraining the backbone, a linear probing strategy was employed. For each participant, rsfMRI data from T1 and T2 were independently passed through the frozen FM using shared weights. From each session, the final hidden representation corresponding to the Classification (CLS) token was extracted as a fixed-dimensional embedding. These embeddings summarize the participant's brain state at each timepoint in the learned representation space. The T1 and T2 embeddings were then concatenated to form a longitudinal feature vector explicitly encoding pre-post neural change. This fused representation was used as input to a lightweight, task-specific classifier trained to ΔEM. During linear probing, all FM weights remained frozen, and only the weights of the downstream classifier were optimized. This setup isolates the representational capacity of the pretrained rsfMRI FMs and allows a controlled comparison of their transferability to small-sample, longitudinal intervention datasets (**Figure 3b**). Detailed training configurations are described in **Supplementary Materials**.

4.4 Model fine-tuning

To adapt the rsfMRI FM to the domain of cognitive aging, supervised fine-tuning was performed on the ADNI dataset. This fine-tuning step was applied exclusively to BrainLM, based on its superior transfer performance in zero-shot and linear probing settings. Fine-tuning was formulated as a binary classification task distinguishing participants diagnosed with AD vs. HC. This task was selected to introduce aging- and pathology-relevant supervisory signals into the pretrained representation space, without altering the downstream intervention datasets. Only ADNI data were used during fine-tuning, and no ADNI subjects were included in any downstream evaluation. To reduce overfitting risk and computational cost, lightweight Low-Rank Adaptation (LoRA) modules were inserted into the query, key, value, and output projection layers of the self-attention blocks in the BrainLM transformer. The majority of pretrained FM weights remained frozen, while the final transformer block was unfrozen to allow limited adaptation of higher-level representations. It enables the model to incorporate domain-specific information while preserving the general spatiotemporal representations learned during large-scale pretraining. During fine-tuning, the downstream classifier consisted of a two-layer MLP operating on the CLS token embedding produced by the BrainLM. Model optimization was conducted using the AdamW optimizer with differential learning rates applied to the adapted rsfMRI FM weights and the downstream classifier. A cosine learning rate scheduler with warmup was used to stabilize training. Early stopping was applied based on validation loss to prevent overfitting, and the model checkpoint with the lowest validation loss was selected. After fine-tuning, the adapted BrainLM model (denoted as BrainLM-ADNI) was used as a frozen feature extractor in all downstream classifications. No further weight updates were performed during linear probing on the ACT and CogTE datasets (see **Figure 3c**).

4.5 Compared ML and DL methods

Support Vector Machine (SVM) served as a strong traditional machine learning method. For each participant, FC matrices were computed separately for T1 and T2 rsfMRI scans. The upper triangular elements of each FC matrix were flattened into feature vectors and concatenated across timepoints to explicitly encode longitudinal change. The resulting feature vector was used as input to a linear SVM classifier (see **Figure 3d**). A linear kernel was chosen to reduce overfitting risk under limited sample sizes. The regularization parameter C was optimized via grid search within the training folds of the cross-validation procedure.

Multilayer Perceptron (MLP) represents a vector-based deep learning method capable of modeling nonlinear relationships. The input features were same as those used for the SVM, consisting of concatenated upper-triangular FC vectors from T1 and T2. The MLP architecture comprised three fully connected layers with ReLU activation functions and dropout for regularization (see **Figure 3d**). Hyperparameters, including hidden layer dimensions, dropout rates, and learning rate, were tuned within the training folds to mitigate overfitting.

Graph Convolutional Network (GCN) was implemented to explicitly incorporate brain network topology and model longitudinal graph-structured data. Brain regions of interest (ROIs) were treated as nodes in a graph, with edges defined by a shared adjacency matrix constructed from a control-group functional connectivity template. For each participant, node features were derived from FC matrices at T1 and T2, resulting in two graph inputs representing baseline and post-intervention brain states. A dual-stream GCN architecture with shared weights was used to process the T1 and T2 graphs in parallel. Each

stream consisted of stacked graph convolutional layers followed by graph-level pooling to produce timepoint-specific embeddings. These embeddings were then fused by concatenation along with their difference to explicitly encode longitudinal change. The fused representation was passed to a fully connected output layer for classification (see **Figure 3d**). Model parameters were optimized using the Adam optimizer, and training was conducted within the same stratified cross-validation framework as all other models.

4.6 Post-hoc analyses

ROI-based Representational Similarity Analysis (RSA) was performed to quantify the correspondence between the high-dimensional representations learned by the model and the actual neural activation patterns of specific brain regions. For the model representational space, a 128-dimensional latent vector was extracted for each participant from the penultimate layer of the downstream classifier. This vector represents the fused longitudinal embedding combining T1 and T2 information used for intervention response classification. For the biological representational space, rsfMRI data were parcellated using the A424 brain atlas. For each ROI, the mean BOLD signal across time was computed separately for T1 and T2 scans and used as a summary physiological feature.

Representational Dissimilarity Matrices (RDMs) were constructed to characterize the inter-individual variability. For the model embedding $E \in \mathbb{R}^{N\times 128}$ (where N is the number of subjects), the model RDM (D_{model}) was computed using correlation distance: $D_{ij} = 1 - \text{corr}(v_i, v_j)$, where v_i and v_j are the feature vectors for subject i and j. For each ROI k, we constructed a biological RDM $\left(D_{ROI_k}\right)$ based on the scalar mean BOLD activation across subjects. Finally, to quantify the alignment between the model's learned representations and the biological features of specific brain areas, we calculated the Spearman rank correlation between the upper triangular elements of D_{model} and D_{ROI_k}. This process yielded a correlation score (ρ) for each ROI at both T1 and T2, indicating how well the inter-subject variability in that specific brain region explains the variance in the model's embedding space.

4.7 Data, Materials, and Software Availability

Data from the ADNI are publicly available and can be accessed through the ADNI data-sharing platform following approval of a data use application. Data from the ACT and CogTE intervention studies contain information that could compromise participant privacy and are therefore not publicly available; de-identified data may be made available from the corresponding authors upon reasonable request and subject to institutional approval. All imaging data were processed using publicly available software, including *fMRIPrep* and *Nilearn*. Publicly released checkpoints of BrainLM and BrainJEPA were used in this study. Code supporting the analyses and results reported here will be made publicly available upon publication.

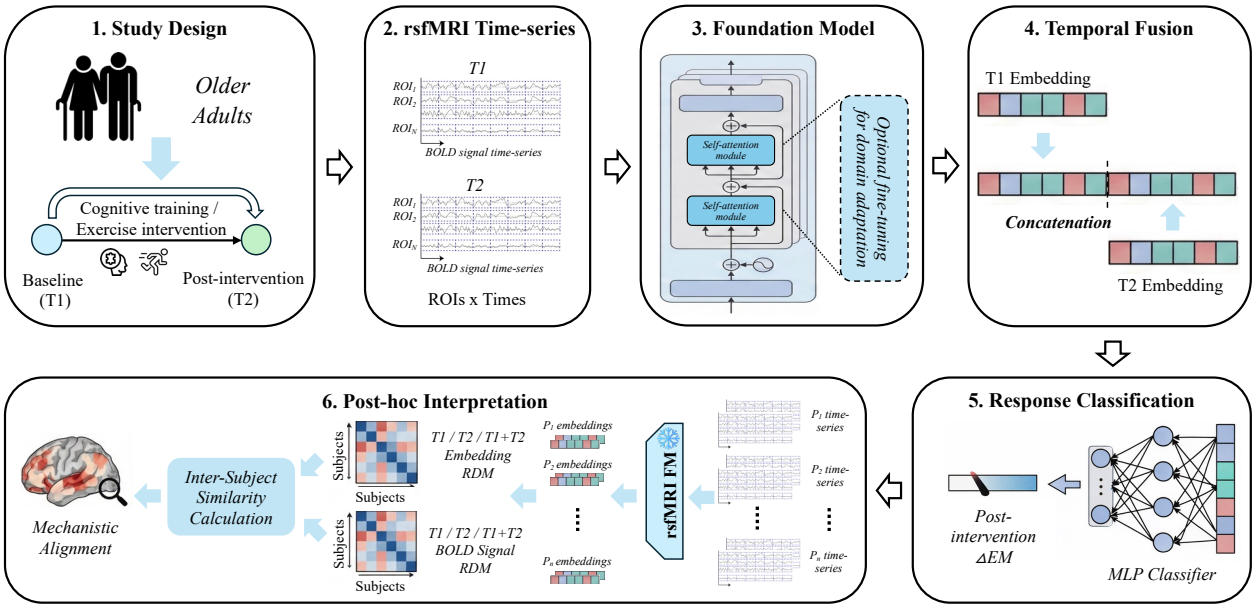

Figure 1. Schematic overview of the analytical pipeline. rsfMRI data collected at baseline (T1) and post-intervention (T2) were independently processed through a dual-stream architecture with shared weights. Pretrained FMs (BrainLM or BrainJEPA) operate directly on regional-averaged BOLD time-series, which are temporally segmented into patches to extract latent whole-brain representations at T1 and T2 respectively. The resulting T1 and T2 embeddings were concatenated to explicitly encode longitudinal neural change and used as input to a lightweight downstream classifier trained to infer ΔEM. Following model training and evaluation within a stratified cross-validation framework, post-hoc interpretability analysis was performed to quantify the correspondence between learned model embeddings and region-wise BOLD signal. **NOTE:** rsfMRI, resting-state functional magnetic resonance imaging; FM, foundation model; BOLD, blood-oxygen-level–dependent; T1/T2, baseline/post-intervention timepoints; MLP, multilayer perceptron; ΔEM, change of epsiodic memory; P, participant; RDM, representational dissimilarity matrix.

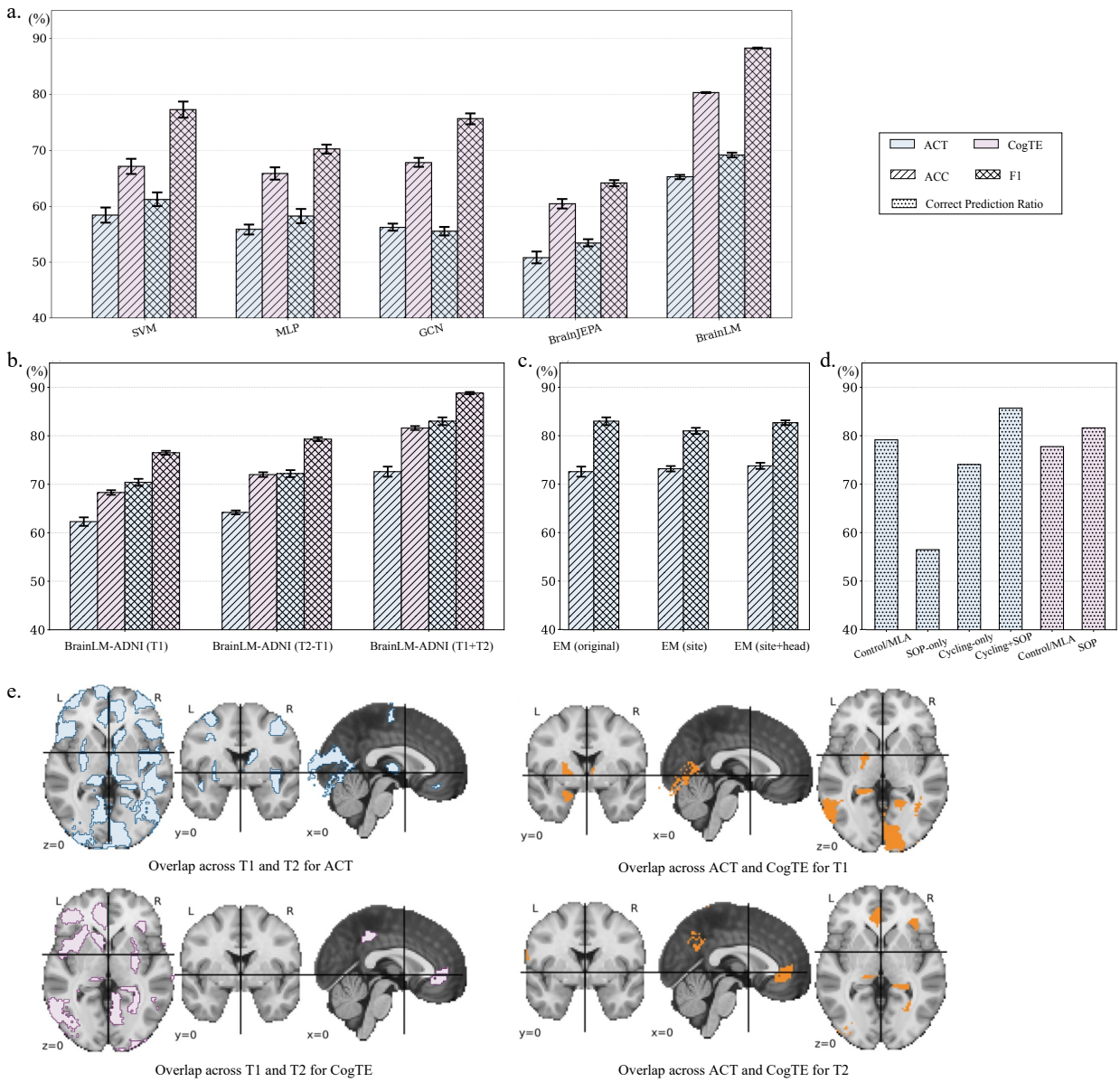

Figure 2. Evaluation of rsfMRI FMs for classifying intervention-related episodic memory response. **(a)** Benchmark comparison of classification performance between rsfMRI FMs (BrainLM, BrainJEPA) and conventional ML and DL approaches (SVM, MLP, GCN) across two independent intervention cohorts (CogTE and ACT). Models are trained to classify participants as episodic memory responders vs. non-responders, based on pre-post changes in EM, using baseline (T1) and post-intervention (T2) rsfMRI data. ACC and F1-score are reported as mean ± standard deviation across stratified cross-validation folds. **(b)** Performance of the BrainLM-ADNI under different timepoint combinations. Classification results for *T1 + T2*, *T1-only*, and *T2 − T1* input settings are shown for CogTE and ACT datasets (mean ± standard deviation across cross-validation folds). **(c)** Robustness of the adapted FM to common neuroimaging confounders in the multi-site ACT dataset. Classification performance of BrainLM-ADNI is shown for original data, data after site harmonization, and data after both site harmonization and head-motion regression. **(d)** Classification performance of BrainLM-ADNI stratified by intervention arm in both CogTE and ACT studies. Chi-square tests indicate no significant association between intervention

assignment and classification correctness. **(e)** Post-hoc interpretability analysis using RSA. Brain regions exhibiting significant correspondence (FDR-corrected $p < 0.05$) between BrainLM-ADNI embeddings and region-wise mean BOLD activity are visualized separately for T1 and T2 in both CogTE and ACT cohorts. **NOTE:** rsfMRI, resting-state functional magnetic resonance imaging; FM, foundation model; ML, machine learning; DL, deep learning; SVM, support vector machine; MLP, multilayer perceptron; GCN, graph convolutional network; ACT, Aerobic exercise and Cognitive Training study; CogTE, Cognitive Training and Exercise study; ADNI, Alzheimer's Disease Neuroimaging Initiative; ACC, accuracy; RSA, representational similarity analysis; FDR, false discovery rate; BOLD, blood-oxygen-level–dependent; EM, episodic memory; T1/T2, baseline/post-intervention timepoints.

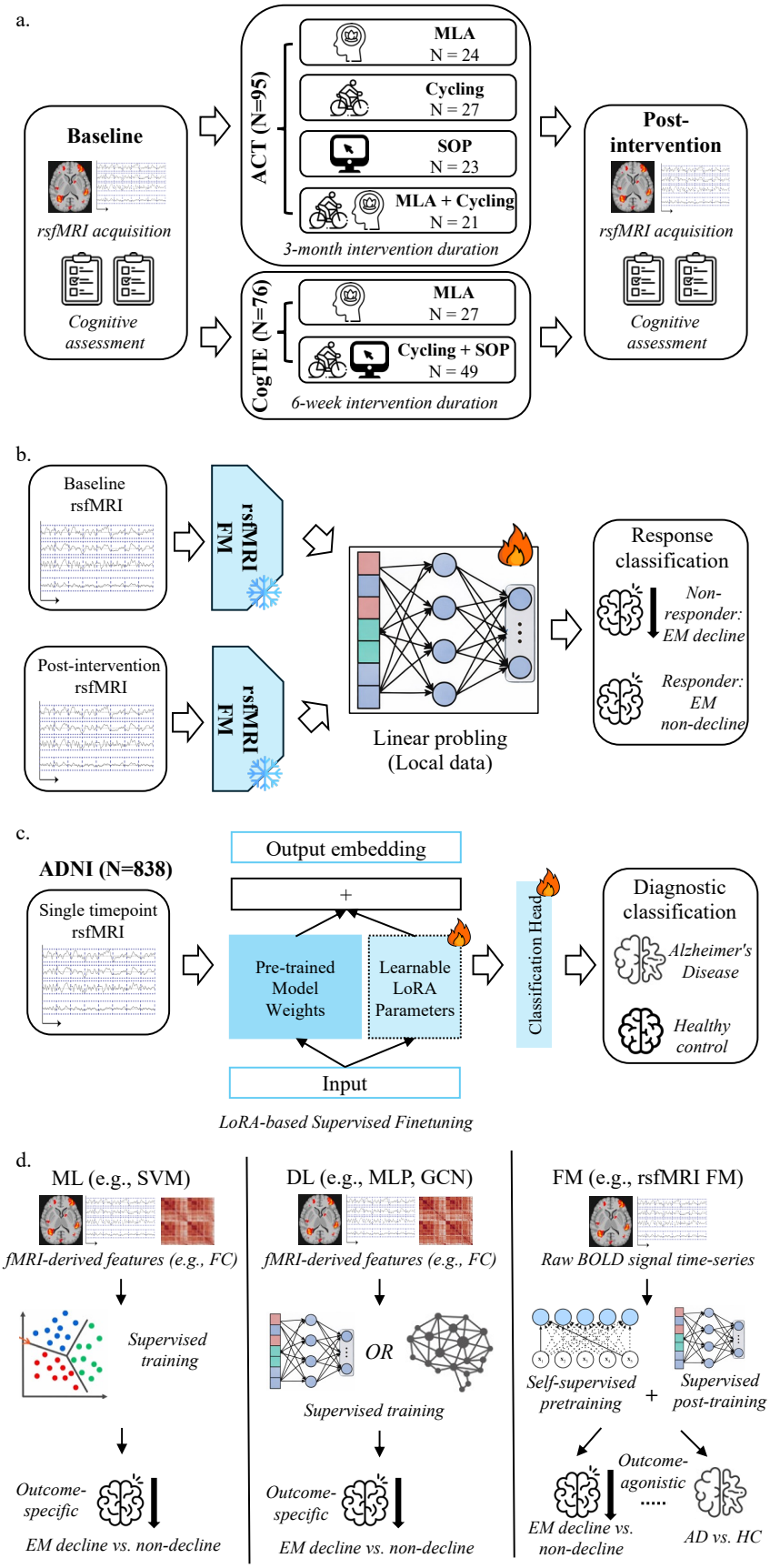
a.
Baseline
rsfMRI acquisition
Cognitive assessment
ACT (N=95)
MLA
N = 24
Cycling
N = 27
SOP
N = 23
MLA + Cycling
N = 21
3-month intervention duration
CogTE (N=76)
MLA
N = 27
Cycling + SOP
N = 49
6-week intervention duration
Post-intervention
rsfMRI acquisition
Cognitive assessment
b.
Baseline rsfMRI
Post-intervention rsfMRI
rsfMRI FM
rsfMRI FM
Linear probing
(Local data)
Response classification
Non-responder: EM decline
Responder: EM non-decline
c.
ADNI (N=838)
Single timepoint rsfMRI
Output embedding
+
Pre-trained Model Weights
Learnable LoRA Parameters
Input
Classification Head
Diagnostic classification
Alzheimer's Disease
Healthy control
LoRA-based Supervised Finetuning
d.
ML (e.g., SVM)
fMRI-derived features (e.g., FC)
Supervised training
Outcome-specific
EM decline vs. non-decline
DL (e.g., MLP, GCN)
fMRI-derived features (e.g., FC)
OR
Supervised training
Outcome-specific
EM decline vs. non-decline
FM (e.g., rsfMRI FM)
Raw BOLD signal time-series
Self-supervised pretraining
+
Supervised post-training
Outcome-agnostic
.....
EM decline vs. non-decline
AD vs. HC

Figure 3. Study datasets and implementation of rsfMRI FMs and comparison methods. **(a)** The ACT and CogTE studies serve as independent intervention cohorts for downstream evaluation and assessment of model generalizability across intervention paradigms and study designs. **(b)** Domain-adaptive finetuning of BrainLM using the ADNI dataset via LoRA. Supervised finetuning is performed to incorporate aging- and pathology-related brain patterns, yielding the adapted BrainLM-ADNI model that is subsequently used as a frozen feature extractor in all downstream intervention analyses. **(c)** Linear probing strategy for deploying pretrained rsfMRI FMs in longitudinal intervention studies. The sfMRI data collected at baseline (T1) and post-intervention (T2) are processed independently through a frozen FM with shared weights, and CLS-token embeddings from each timepoint are concatenated to explicitly encode longitudinal neural change for downstream inference of EM outcomes. **(d)** Architectures of conventional comparison models. FC features derived from rsfMRI data at T1 and T2 are used to train an SVM, an MLP, and a dual-stream GCN that explicitly models longitudinal brain network changes. **NOTE:** rsfMRI, resting-state functional magnetic resonance imaging; FM, foundation model; ACT, Aerobic exercise and Cognitive Training study; MLA, mental leisure activity; CogTE, Cognitive Training and Exercise study; ADNI, Alzheimer's Disease Neuroimaging Initiative; EM, episodic memory; FC, functional connectivity; T1/T2, baseline/post-intervention timepoints; SVM, support vector machine; MLP, multilayer perceptron; GCN, graph convolutional network; LoRA, low-rank adaptation; AD, Alzheimer's Disease; HC, healthy control.

Tabel 1 Implementation Guide for rsfMRI FMs in Intervention Studies

Dimension	Clinical Consideration	Technical Solution	Engineering Effort
Data Requirements	Small-sample, longitudinal intervention cohorts limit direct model training.	Reuse pretrained rsfMRI FMs; avoid end-to-end training from scratch.	Low-Medium
Domain Shift	Mismatch between large-scale pretraining cohorts and aging intervention populations.	Apply clinically informed domain-adaptive fine-tuning using aging- or pathology-relevant datasets (e.g., ADNI).	High
Temporal Structure	Intervention effects manifest as within-subject neural change rather than static states.	Explicit temporal modeling of baseline and post-intervention representations.	Medium
Population Heterogeneity	Increased variability and dynamics in cognitively impaired older adults.	Prefer sequence-based or flexible temporal modeling objectives over strict consistency constraints.	Medium
Generalizability	Heterogeneity across sites, scanners, and intervention modalities.	Use stratified validation and cross-study evaluation to assess robustness and transferability.	Medium
Interpretability	Latent FM representations are high-dimensional and difficult to interpret clinically.	Perform post-hoc analyses linking embeddings to neurobiological signals.	Medium–High
Reproducibility	Limited access to FM pretraining data and weights.	Emphasize transparent adaptation protocols and report deployable performance using public checkpoints.	Medium

References

1. Krellman JW, Mercuri G. Cognitive Interventions for Neurodegenerative Disease. *Curr Neurol Neurosci Rep*. 2023;23(9):461-468. doi:10.1007/s11910-023-01283-1

2. Xiang C, Zhang Y. Comparison of Cognitive Intervention Strategies for Individuals With Alzheimer's Disease: A Systematic Review and Network Meta-analysis. *Neuropsychol Rev*. 2024;34(2):402-416. doi:10.1007/s11065-023-09584-5

3. Chan ATC, Ip RTF, Tran JYS, Chan JYC, Tsoi KKF. Computerized cognitive training for memory functions in mild cognitive impairment or dementia: a systematic review and meta-analysis. *Npj Digit Med*. 2024;7(1):1. doi:10.1038/s41746-023-00987-5

4. Ciria LF, Román-Caballero R, Vadillo MA, et al. An umbrella review of randomized control trials on the effects of physical exercise on cognition. *Nat Hum Behav*. 2023;7(6):928-941. doi:10.1038/s41562-023-01554-4

5. Traut HJ, Guild RM, Munakata Y. Why Does Cognitive Training Yield Inconsistent Benefits? A Meta-Analysis of Individual Differences in Baseline Cognitive Abilities and Training Outcomes. *Front Psychol*. 2021;12. doi:10.3389/fpsyg.2021.662139

6. Stern Y. A framework for concepts of reserve and resilience in aging. *Alzheimers Dement*. 2024;20(S7):e087228. doi:10.1002/alz.087228

7. Belleville S, Clément F, Mellah S, Gilbert B, Fontaine F, Gauthier S. Training-related brain plasticity in subjects at risk of developing Alzheimer's disease. *Brain*. 2011;134(6):1623-1634. doi:10.1093/brain/awr037

8. Haller S, Monsch AU, Richiardi J, Barkhof F, Kressig RW, Radue EW. Head Motion Parameters in fMRI Differ Between Patients with Mild Cognitive Impairment and Alzheimer Disease Versus Elderly Control Subjects. *Brain Topogr*. 2014;27(6):801-807. doi:10.1007/s10548-014-0358-6

9. Tsvetanov KA, Henson RNA, Jones PS, et al. The effects of age on resting-state BOLD signal variability is explained by cardiovascular and cerebrovascular factors. *Psychophysiology*. 2021;58(7):e13714. doi:10.1111/psyp.13714

10. Geerligs L, Tsvetanov KA, Cam-CAN, Henson RN. Challenges in measuring individual differences in functional connectivity using fMRI: The case of healthy aging. *Hum Brain Mapp*. 2017;38(8):4125-4156. doi:10.1002/hbm.23653

11. Dennis EL, Thompson PM. Functional Brain Connectivity Using fMRI in Aging and Alzheimer's Disease. *Neuropsychol Rev*. 2014;24(1):49-62. doi:10.1007/s11065-014-9249-6

12. Nielsen AN, Barch DM, Petersen SE, Schlaggar BL, Greene DJ. Machine Learning With Neuroimaging: Evaluating Its Applications in Psychiatry. *Biol Psychiatry Cogn Neurosci Neuroimaging*. 2020;5(8):791-798. doi:10.1016/j.bpsc.2019.11.007

13. Ghahramani Z. Probabilistic machine learning and artificial intelligence. *Nature*. 2015;521(7553):452-459. doi:10.1038/nature14541

14. Bastos AM, Schoffelen JM. A tutorial review of functional connectivity analysis methods and their interpretational pitfalls. *Front Syst Neurosci*. 2016;9:175.

15. Warren SL, Moustafa AA. Functional magnetic resonance imaging, deep learning, and Alzheimer's disease: A systematic review. *J Neuroimaging*. 2023;33(1):5-18. doi:10.1111/jon.13063

16. Lecoq J, Oliver M, Siegle JH, Orlova N, Ledochowitsch P, Koch C. Removing independent noise in systems neuroscience data using DeepInterpolation. *Nat Methods*. 2021;18(11):1401-1408. doi:10.1038/s41592-021-01285-2

17. Zhang K, Zhou R, Adhikarla E, et al. A generalist vision–language foundation model for diverse biomedical tasks. *Nat Med*. 2024;30(11):3129-3141. doi:10.1038/s41591-024-03185-2

18. A comprehensive survey on pretrained foundation models: a history from BERT to ChatGPT | International Journal of Machine Learning and Cybernetics. Accessed January 30, 2025. https://link.springer.com/article/10.1007/s13042-024-02443-6

19. Moor M, Banerjee O, Abad ZSH, et al. Foundation models for generalist medical artificial intelligence. *Nature*. 2023;616(7956):259-265. doi:10.1038/s41586-023-05881-4

20. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *Adv Neural Inf Process Syst*. 2017;30.

21. Caro JO, de Oliveira Fonseca AH, Rizvi SA, et al. BrainLM: A foundation model for brain activity recordings.

22. Dong Z, Li R, Wu Y, et al. Brain-JEPA: Brain Dynamics Foundation Model with Gradient Positioning and Spatiotemporal Masking. *arXiv*. Preprint posted online September 28, 2024. doi:10.48550/arXiv.2409.19407

23. Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*. 2018;562(7726):203-209.

24. Van Essen DC, Ugurbil K, Auerbach E, et al. The Human Connectome Project: a data acquisition perspective. *Neuroimage*. 2012;62(4):2222-2231.

25. Jack Jr CR, Bernstein MA, Fox NC, et al. The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *J Magn Reson Imaging Off J Int Soc Magn Reson Med*. 2008;27(4):685-691.

26. Yu F, Lin FV, Salisbury DL, et al. Efficacy and mechanisms of combined aerobic exercise and cognitive training in mild cognitive impairment: study protocol of the ACT trial. *Trials*. 2018;19(1):700. doi:10.1186/s13063-018-3054-0

27. Lin FV, Tao Y, Chen Q, et al. Processing speed and attention training modifies autonomic flexibility: A mechanistic intervention study. *NeuroImage*. 2020;213:116730. doi:10.1016/j.neuroimage.2020.116730

28. Hearst MA, Dumais ST, Osuna E, Platt J, Scholkopf B. Support vector machines. *IEEE Intell Syst Their Appl*. 1998;13(4):18-28.

29. Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol Rev*. 1958;65(6):386.

30. Kipf T. Semi-supervised classification with graph convolutional networks. *ArXiv Prepr ArXiv160902907*. Published online 2016.

31. Filippi M, Basaia S, Sarasso E, et al. Longitudinal brain connectivity changes and clinical evolution in Parkinson's disease. *Mol Psychiatry*. 2021;26(9):5429-5440. doi:10.1038/s41380-020-0770-0

32. Sun L, Zhao T, Liang X, et al. Human lifespan changes in the brain's functional connectome. *Nat Neurosci*. 2025;28(4):891-901. doi:10.1038/s41593-025-01907-4

33. Vliet M van, Appelhoff S, Shimizu T, et al. MNE-RSA: Representational Similarity Analysis on EEG, MEG and fMRI data. *J Open Source Softw*. 2025;10(116):9148. doi:10.21105/joss.09148

34. Grady C, Sarraf S, Saverino C, Campbell K. Age differences in the functional interactions among the default, frontoparietal control, and dorsal attention networks. *Neurobiol Aging*. 2016;41:159-172. doi:10.1016/j.neurobiolaging.2016.02.020

35. Mevel K, Chételat G, Eustache F, Desgranges B. The Default Mode Network in Healthy Aging and Alzheimer′s Disease. *Int J Alzheimer's Dis*. 2011;2011(1):535816. doi:10.4061/2011/535816

36. Wei X, Zhao K, Jiao Y, He L, Zhang Y. A Brain Graph Foundation Model: Pre-Training and Prompt-Tuning for Any Atlas and Disorder. *ArXiv Prepr ArXiv250602044*. Published online 2025.

37. Achiam J, Adler S, Agarwal S, et al. Gpt-4 technical report. *ArXiv Prepr ArXiv230308774*. Published online 2023.

38. Liu H, Li C, Wu Q, Lee YJ. Visual instruction tuning. *Adv Neural Inf Process Syst*. 2023;36:34892-34916.

39. Bai J, Bai S, Chu Y, et al. Qwen technical report. *ArXiv Prepr ArXiv230916609*. Published online 2023.

40. Mitchell M. How do we know how smart AI systems are? *Science*. 2023;381(6654):eadj5957. doi:10.1126/science.adj5957

41. Wang C, Jiang Y, Peng Z, et al. Towards a general-purpose foundation model for fMRI analysis. *arXiv*. Preprint posted online June 11, 2025. doi:10.48550/arXiv.2506.11167

42. Qiu Q, Li S, Zhang J, et al. Lab-in-the-loop machine learning for brain-targeting delivery system design. *Cell Biomater*. Published online 2025.

43. Esteban O, Markiewicz CJ, Blair RW, et al. fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nat Methods*. 2019;16(1):111-116. doi:10.1038/s41592-018-0235-4

44. Andersson JLR, Skare S, Ashburner J. How to correct susceptibility distortions in spin-echo echo-planar images: application to diffusion tensor imaging. *NeuroImage*. 2003;20(2):870-888. doi:10.1016/S1053-8119(03)00336-7