BrainAnytime: Anatomy-Aware Cross-Modal Pretraining for Brain Image Analysis with Arbitrary Modality Availability

Guangqian Yang[1]*, Tong Ding[1]*, Wenlong Hou[1], Yue Xun[1], Ye Du[1], Qian Niu[2]✉, Shujun Wang[1,3]✉, and for the Alzheimer's Disease Neuroimaging Initiative

[1] Department of Biomedical Engineering, The Hong Kong Polytechnic University, Hong Kong SAR, China
[2] Department of Technology Management for Innovation, The University of Tokyo, Japan
[3] Department of Data Science and Artificial Intelligence, The Hong Kong Polytechnic University, Hong Kong SAR, China
qian.niu@weblab.t.u-tokyo.ac.jp, shu-jun.wang@polyu.edu.hk
✉ Corresponding authors.
*These authors contributed equally to this work.

Abstract. Clinical diagnostic workups typically follow a modality escalation pathway: after initial clinical evaluation, clinicians begin with routine structural imaging (*e.g.*, MRI), selectively add sequences such as FLAIR or T2 to refine the differential, and reserve molecular imaging (*e.g.*, amyloid-PET) for cases that remain uncertain after standard evaluation. Consequently, patients are observed with heterogeneous and often incomplete modality subsets. However, most current AI models assume fixed data modalities as the model inputs. In this paper, we present BrainAnytime, a unified pretraining framework pretrained on **34,899** 3D brain scans from five datasets that support brain image analysis under arbitrary modality availability spanning multi-sequence MRI and amyloid-PET. A single model accepts whatever imaging is available, from a lone T1 scan to a full multimodal workup. Pretraining learns structural-molecular correspondences between MRI and PET via cross-modal distillation (RCMD) and prioritizes disease-vulnerable anatomy via atlas-guided curriculum masking (PACM), all within a shared 3D masked autoencoder (Multi-MAE3D). Across **four** downstream tasks and **five** clinically motivated modality settings, BrainAnytime largely outperforms modality-specific models, missing-modality baselines, and large-scale brain MRI pretrained foundation models on most modality settings. Notably, it surpasses the strongest missing-modality baselines with relative improvements of **6.2%** and **7.0%** in average accuracy on CN vs. AD and CN vs. MCI classification, respectively. Code is available at `https://github.com/SDH-Lab/BrainAnytime`.

Keywords: Modality Escalation Pathway · Missing Modality · Alzheimer's Disease · Pretraining Framework · Anatomy-aware Learning

1 Introduction

When a patient presents with suspected cognitive decline, the diagnostic workup typically begins with clinical history review and neuropsychological screening [14]. Once these initial evaluations warrant neuroimaging, clinicians rarely order every modality upfront. Instead, imaging follows a staged escalation of evidence: T1-weighted MRI is acquired first, additional MRI sequences (*e.g.*, FLAIR and T2) are added as needed, and amyloid-PET, which measures β-amyloid plaque burden and serves as a core biomarker of AD pathology, is often reserved for cases where structural imaging alone is insufficient [4]. We refer to this sequential acquisition logic as the *clinical modality escalation pathway*. Consequently, modality availability is inherently incomplete and stage dependent: in major AD cohorts, only a small fraction of subjects have a complete set of all imaging modalities (Fig. 1 and Table 1), while most are observed through partial and heterogeneous modality subsets.

These practice patterns motivate a different pretraining paradigm. Rather than training separate models for fixed input configurations, we seek a unified pretraining model that supports brain image analysis under *arbitrary modality availability* spanning multi-sequence MRI and amyloid-PET. Such a framework should satisfy three requirements. **(i) Modality-flexible inference.** A single model should accept any subset of available modalities at test time without architectural switching or retraining. **(ii) Cross-modal structural-molecular correspondence.** Pretraining should encourage shared representations that capture correspondences between structural MRI and amyloid-PET, instead of learning each modality in isolation [5]. **(iii) Anatomy-aware learning.** Because AD exhibits selective regional vulnerability [13], pretraining should emphasize disease-salient anatomy rather than treating all locations uniformly.

However, existing brain imaging pretraining paradigms do not meet these requirements in the AD setting. Recent large-scale self-supervised models pretrained on brain neuroimaging show strong transfer across downstream tasks [19, 6, 17, 22, 7, 23, 9, 27]. Despite this progress, three limitations remain. First, most are **MRI-only** and therefore do not learn an explicit bridge to amyloid-PET, even though PET is a key source of pathological evidence along the escalation pathway. Second, pretraining and deployment commonly assume a **fixed modality** (or a fixed combination), mismatching real-world, stage-dependent missingness and hindering inference under arbitrary modality subsets. Third, standard MAE pretraining [12] relies on **spatially uniform masking**, providing no mechanism to prioritize anatomically vulnerable regions that are most informative for AD. These gaps motivate a cross-modal, modality-flexible, and anatomy-aware pretraining framework tailored to clinically realistic modality availability.

We present **BrainAnytime**, an anatomy-aware cross-modal pretraining framework for brain image analysis with *arbitrary modality availability*. BrainAnytime is pretrained on **34,899** 3D brain scans from **five** large-scale datasets spanning multi-sequence MRI and amyloid-PET, by randomly sampling modality subsets to mimic clinical missingness. BrainAnytime integrates three components:

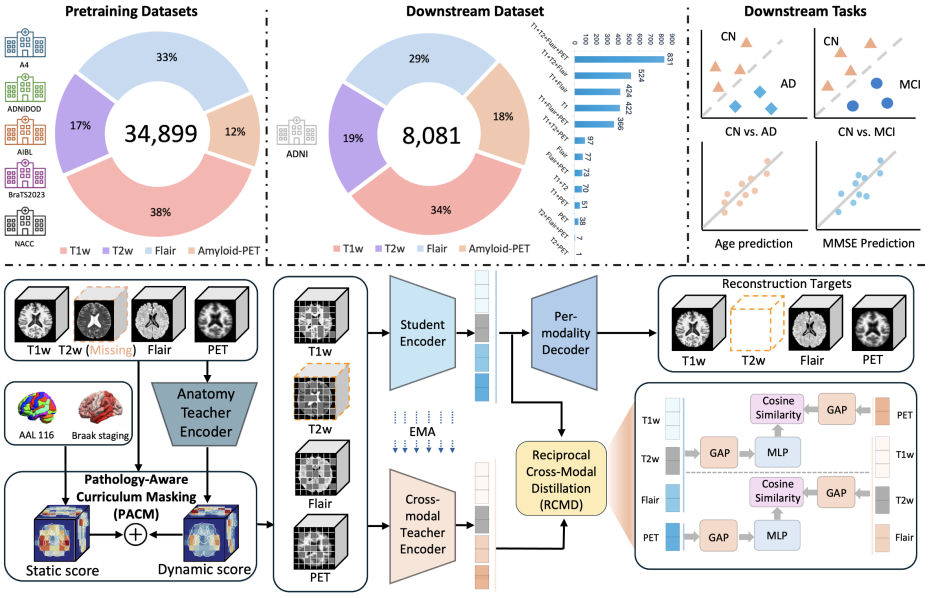


Fig. 1: Overall framework of BrainAnytime. EMA, GAP, and MLP denote exponential moving average, global average pooling, and the multilayer perceptron predictors, respectively.

Multi-MAE3D, a 3D multi-modal masked autoencoder with a shared Transformer encoder for any modality subset; **Reciprocal Cross-Modal Distillation (RCMD)**, an EMA-teacher distillation objective that aligns MRI and PET representations; and **Pathology-Aware Curriculum Masking (PACM)**, an atlas-guided curriculum masking strategy that emphasizes AD-relevant neuroanatomy during reconstruction. Together, these designs yield a unified pretrained model that is cross-modal, robust to missing modalities, and explicitly anatomy-aware.

2 Method

We propose *Anatomy-Aware Cross-Modal Pretraining*, a self-supervised framework for learning unified representations from heterogeneous 3D neuroimaging data with arbitrary missing modalities. As shown in Fig. 1, our framework consists of three components: (i) a 3D multi-modal masked autoencoder (**Multi-MAE3D**) that encodes any subset of modalities with a single Transformer; (ii) **Reciprocal Cross-Modal Distillation (RCMD)** that learns MRI–PET correspondences via an EMA teacher; and (iii) **Pathology-Aware Curriculum Masking (PACM)** that progressively focuses reconstruction on disease-relevant brain regions.

2.1 Multi-MAE3D: 3D Multi-Modal Masked Autoencoder

Given a set of 3D volumes $\{\mathbf{x}_m\}_{m\in\mathcal{M}}$ where $\mathcal{M} \subseteq \{\text{T1, T2, FLAIR, PET}\}$ denotes the available modalities, each volume $\mathbf{x}_m \in \mathbb{R}^{D\times H\times W}$ is partitioned into N non-overlapping patches of size p^3.

Per-modality tokenization. Each modality has a modality-specific patch embedding f_m with a 3D convolutional layer, that maps raw patches to d-dimensional tokens. Following PACM (Sec. 2.3), only a subset of visible tokens per modality is retained. Missing modalities are fully masked, and the visible-token budget is redistributed among observed modalities via $\text{Dir}(\alpha)$ [9] under a global mask ratio r, keeping the total token count constant regardless of modality availability.

Shared encoder. The visible tokens from all available modalities are concatenated with a learnable `[CLS]` token, augmented with 3D sinusoidal positional embeddings, and processed by a shared student encoder. An attention mask blocks tokens from missing modalities (set to $-\infty$), ensuring the encoder gracefully handles any modality subset at both training and inference time.

Per-modality decoders. Each modality has a lightweight decoder that takes the encoded visible tokens and learnable mask tokens as input. Cross-attention between mask tokens and encoder output is followed by self-attention blocks. The decoder reconstructs *all* N patches per modality; the training loss is computed only on masked patches using mean squared error with per-patch normalization [12]:

$$\mathcal{L}_{\text{MAE}} = \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}}\frac{1}{|\mathcal{S}_m|}\sum_{i\in\mathcal{S}_m}\|\hat{\mathbf{p}}_i^m - \bar{\mathbf{p}}_i^m\|_2^2\,, \tag{1}$$

where $\mathcal{S}_m$ is the set of masked patch indices for modality m, $\hat{\mathbf{p}}_i^m$ is the predicted patch, and $\bar{\mathbf{p}}_i^m$ is the per-patch normalized ground truth.

2.2 Reciprocal Cross-Modal Distillation (RCMD)

To explicitly learn MRI–PET correspondences, we introduce RCMD, a cross-modal prediction objective based on an exponential moving average (EMA) cross-modal teacher. The teacher shares the encoder architecture and is updated as $\theta_T \leftarrow \mu\,\theta_T + (1-\mu)\,\theta_S$, where the momentum μ follows a cosine schedule from μ_0 to 1.0. For paired samples containing both MRI and PET, we compute group-level representations by average-pooling the encoded tokens within two groups: $\mathbf{z}_{\text{MRI}}$ (T1+T2+FLAIR tokens) and $\mathbf{z}_{\text{PET}}$ (PET tokens), from both the student and teacher encoders. Two symmetric MLP predictors $g_{\text{M}\to\text{P}}$ and $g_{\text{P}\to\text{M}}$ are trained to predict cross-modal teacher features:

$$\mathcal{L}_{\text{RCMD}} = 1 - \frac{1}{2}\big[\cos\big(g_{\text{M}\to\text{P}}(\mathbf{z}^s_{\text{MRI}}),\,\mathbf{z}^t_{\text{PET}}\big) + \cos\big(g_{\text{P}\to\text{M}}(\mathbf{z}^s_{\text{PET}}),\,\mathbf{z}^t_{\text{MRI}}\big)\big]\,, \tag{2}$$

where superscripts s and t denote the student and teacher, respectively, and all representations are ℓ_2-normalized; $cos(\cdot)$ denotes cosine similarity.

2.3 Pathology-Aware Curriculum Masking (PACM)

Standard MAE treats all spatial locations equally during masking. In neuroimaging, however, certain regions carry greater diagnostic relevance (*e.g.*, hippocampus in AD). We introduce a PACM strategy that biases mask sampling toward clinically important regions through a temperature-scheduled curriculum.

Importance Scoring. We register the AAL116 atlas to the input space and compute a patch-region membership matrix $\mathbf{R} \in \mathbb{R}^{N \times K}$, where $R_{i,k}$ is the fraction of voxels in patch i belonging to region k. A *static* score is derived from neuropathological priors (Braak staging [3]): $s_i^{\mathrm{s}} = \sum_k R_{i,k}\, w_k$, where w_k reflects the pathological relevance of region k. A *dynamic* score s_i^{d} is obtained by periodically extracting the CLS-to-patch attention from an EMA anatomy teacher and aggregating it to region level via $\mathbf{R}$. The two are combined as $s_i = (1-\beta)\,\hat{s}_i^{\mathrm{s}} + \beta\,\hat{s}_i^{\mathrm{d}}$ after min-max normalization.

Curriculum Schedule. Importance scores are converted to mask probabilities via $p_i = \mathrm{softmax}(s_i/\tau)$. To avoid early training instability, we adopt a three-phase curriculum: (1) uniform random masking in the early stage; (2) τ anneals from τ_{start} to τ_{target} with cosine decay, progressively sharpening the distribution; (3) $\tau{=}\tau_{\mathrm{target}}$ for the remainder of training. Patches are then selected using the Gumbel-top-k trick [9]: $k_i = \log p_i + g_i,\ g_i{\sim}\mathrm{Gumbel}(0,1)$, where low-importance patches are kept visible and high-importance patches are masked. This progressively shifts the model's reconstructive effort toward pathologically critical structures.

2.4 Training Objective and Downstream Finetuning

The overall loss is $\mathcal{L} = \mathcal{L}_{\mathrm{MAE}} + \lambda \cdot \mathbb{1}[\text{paired}] \cdot \mathcal{L}_{\mathrm{RCMD}}$, where λ follows a linear warmup schedule and $\mathbb{1}[\text{paired}]$ activates the cross-modal term only when both MRI and PET are observed. For finetuning, we only use the student encoder with a lightweight task head (layer normalization + linear layer on the `[CLS]` token) and remove patch-level masking. Missing modalities remain zeroed and attention-blocked as in pretraining; modality dropout is applied during training for additional robustness.

3 Experimental Results

3.1 Dataset Description

We curate a multi-cohort neuroimaging collection from six publicly available datasets totaling 16,481 subjects (Tab. 1). Five datasets (A4 [18], DOD-ADNI [20], AIBL [8], BraTS23 [1], and NACC [2]; 13,500 subjects, 34,899 scans) spanning multi-sequence MRI and amyloid-PET are used for pretraining; an independent ADNI cohort [21] (2,981 subjects) is held out for evaluation. We designate five clinically common modality combinations (T1; T1+FLAIR; T1+FLAIR+PET; T1+T2+FLAIR; T1+T2+FLAIR+PET) for test reporting with a diagnosis-stratified train/val/test split (60/10/30); remaining low-frequency combinations

Table 1: Modality availability across pretraining datasets and downstream diagnostic classes, reported as count (percentage).

Split	Subjects	T1	T2	FLAIR	PET	Total
Pretraining Datasets						
A4	1,736	1,736 (100.0%)	1,719 (99.0%)	1,729 (99.6%)	1,526 (87.9%)	6,710
DOD-ADNI	543	499 (91.9%)	0 (0.0%)	0 (0.0%)	284 (52.3%)	783
AIBL	1,408	1,283 (91.1%)	801 (56.9%)	670 (47.6%)	407 (28.9%)	3,161
BraTS23	2,569	2,569 (100.0%)	2,569 (100.0%)	2,569 (100.0%)	0 (0.0%)	7,707
NACC	7,244	7,244 (100.0%)	773 (10.7%)	6,637 (91.6%)	1,884 (26.0%)	16,538
Subtotal	**13,500**	**13,331 (98.7%)**	**5,862 (43.4%)**	**11,605 (86.0%)**	**4,101 (30.4%)**	**34,899**
Downstream Dataset (ADNI)						
CN	1,235	1,143 (92.6%)	700 (56.7%)	1,050 (85.0%)	738 (59.8%)	3,631
MCI	1,133	1,069 (94.4%)	554 (48.9%)	869 (76.7%)	496 (43.8%)	2,988
AD	613	573 (93.5%)	276 (45.0%)	383 (62.5%)	230 (37.5%)	1,462
Subtotal	**2,981**	**2,785 (93.4%)**	**1,530 (51.3%)**	**2,302 (77.2%)**	**1,464 (49.1%)**	**8,081**

are used for train/val only (80/20). All volumes undergo skull stripping, affine registration to MNI152, min–max normalization, and resampling to $128 \times 128 \times 128$. We evaluate on four downstream tasks [7, 19]: CN vs. MCI, CN vs. AD, MMSE regression, and age prediction.

3.2 Implementation Details

The encoder is ViT-B (d=768, 12 heads, 12 layers, patch size p=16, N=512 patches); each decoder has 2 blocks with dimension 384. We pretrain with AdamW (lr 10^{-4}, weight decay 0.05) for 1000 epochs (40-epoch warmup, cosine decay) on 8×A100 GPUs, batch size 128. Mask ratio r=0.75, Dirichlet α=1.0; λ=0.1; μ_0=0.996. PACM uses importance weights $w_k \in \{3.0, 1.5, 0.3\}$ for AD-critical, other gray-matter, and non-brain regions (β=0.5), with curriculum phases at 20% and 70% of training (τ: 5.0→1.0). Data augmentation includes random flipping, affine transforms, and elastic deformation. For finetuning, we use AdamW (lr 10^{-5}, weight decay 0.05) for up to 100 epochs with early stopping (patience 15); the encoder is frozen for the first 5 epochs. Classification uses BCE; regression uses MSE.

3.3 Benchmarking with Clinically Driven Modality Escalation

To benchmark BrainAnytime under clinically realistic, stage-dependent modality availability, we compare it with four families of baselines: (i) single-modality models (3D ResNet [11], M3T [15]), (ii) dual-modality fusion methods (MCAD [26], MENet [16]), (iii) brain foundation models (BrainIAC [19], SAM-Brain3D [6]), and (iv) incomplete multimodal learning approaches (Flex-MoE [24], FuseMoE [10], MoE-retriever [25]). Following the escalation-motivated modality patterns in our cohort (Sec. 3.1), we report results on five mutually exclusive modality combinations. For classification, we use ACC/AUC/F1; for MMSE and age regression, we report MAE/RMSE/PCC [7]. All results are averaged over 3 seeds (mean(std)). As shown in Table 2, BrainAnytime achieves the best overall performance across four downstream tasks and five clinically motivated modality

Table 2: Results on ADNI. The best results are highlighted in pink, and the second-best results are underlined.

Modality	Method	CN vs. AD			CN vs. MCI			MMSE prediction			Age prediction (Years)		
		ACC↑	AUC↑	F1↑	ACC↑	AUC↑	F1↑	MAE ↓	RMSE ↓	PCC↑	MAE ↓	RMSE ↓	PCC↑
T1	3D Resnet50 [11]	84.2 (2.8)	93.1 (2.6)	87.2 (1.6)	62.6 (4.3)	67.5 (6.2)	72.7 (1.9)	2.102 (0.116)	2.730 (0.162)	60.0 (6.1)	3.759 (0.087)	4.851 (0.154)	78.5 (1.5)
	M3T [15]	70.3 (1.4)	81.2 (1.0)	75.9 (2.0)	63.4 (3.8)	54.7 (4.1)	75.5 (2.4)	2.370 (0.044)	2.964 (0.098)	43.5 (2.7)	5.230 (0.082)	6.512 (0.067)	57.9 (1.0)
	BrainIAC [19]	76.1 (3.4)	82.5 (1.5)	79.1 (2.9)	60.9 (2.6)	58.4 (0.5)	72.0 (1.6)	2.358 (0.084)	3.030 (0.160)	46.5 (0.9)	5.246 (0.067)	6.584 (0.082)	53.3 (3.4)
	SAM-Brain3D [6]	81.1 (3.6)	91.7 (1.9)	84.2 (2.2)	58.9 (1.9)	59.3 (2.4)	69.8 (2.6)	2.210 (0.106)	2.836 (0.144)	56.2 (2.6)	4.492 (0.154)	5.630 (0.092)	68.2 (1.5)
	Flex-MoE [24]	75.2 (3.9)	81.7 (7.4)	81.3 (2.3)	63.4 (2.9)	59.4 (3.3)	74.4 (4.6)	2.660 (0.100)	3.560 (0.200)	21.7 (1.7)	6.205 (0.667)	7.845 (0.923)	60.2 (5.6)
	FuseMoE [10]	74.8 (2.1)	80.2 (1.0)	80.4 (1.9)	60.9 (3.8)	57.8 (1.6)	73.8 (4.2)	2.560 (0.040)	3.260 (0.160)	27.8 (3.0)	4.871 (0.154)	6.256 (0.154)	58.9 (2.8)
	MoE-retriever [25]	69.4 (11.3)	72.0 (13.9)	79.9 (5.4)	65.8 (0.7)	60.9 (4.5)	79.0 (0.8)	2.363 (0.187)	3.007 (0.178)	29.2 (31.6)	5.076 (0.974)	6.410 (0.974)	62.1 (13.2)
	BrainAnytime	86.5 (2.7)	93.4 (2.4)	88.4 (2.9)	66.7 (2.5)	70.1 (1.5)	75.2 (3.4)	2.100 (0.040)	2.660 (0.080)	57.8 (0.4)	4.205 (0.051)	5.435 (0.103)	71.2 (0.6)
T1+Flair	MCAD [26]	87.8 (3.6)	93.9 (1.0)	67.9 (10.5)	63.3 (4.2)	65.6 (2.9)	54.4 (7.0)	1.910 (0.056)	2.814 (0.130)	52.6 (2.5)	4.205 (0.282)	5.461 (0.328)	75.3 (0.8)
	MENet [16]	89.6 (2.1)	95.4 (0.3)	73.5 (5.8)	62.1 (2.9)	64.5 (3.6)	52.6 (5.4)	1.924 (0.014)	2.784 (0.078)	53.1 (3.2)	4.102 (0.144)	5.317 (0.190)	75.8 (0.6)
	BrainIAC	80.2 (0.8)	87.8 (1.1)	58.9 (4.4)	62.1 (0.5)	67.6 (3.0)	59.0 (1.3)	1.964 (0.062)	2.958 (0.092)	43.9 (6.3)	5.087 (0.144)	6.440 (0.179)	63.4 (1.4)
	SAM-Brain3D	91.4 (3.4)	95.3 (1.6)	81.9 (6.6)	64.9 (1.4)	68.2 (2.2)	62.0 (2.6)	1.964 (0.038)	2.884 (0.052)	47.9 (3.1)	4.466 (0.303)	5.712 (0.472)	73.0 (2.8)
	Flex-MoE	88.3 (0.8)	95.5 (0.3)	75.9 (1.8)	58.5 (8.2)	62.7 (5.3)	50.7 (12.0)	2.320 (0.460)	3.140 (0.320)	41.7 (8.6)	4.974 (0.615)	6.307 (0.513)	68.1 (1.7)
	FuseMoE	90.1 (0.8)	94.3 (0.8)	78.3 (2.4)	65.8 (2.1)	69.9 (2.3)	61.5 (5.0)	1.980 (0.020)	3.000 (0.160)	43.0 (4.9)	4.769 (0.103)	6.205 (0.154)	64.8 (0.9)
	MoE-retriever	90.1 (2.1)	93.9 (2.6)	77.6 (4.2)	53.3 (3.7)	57.8 (3.7)	58.6 (6.8)	1.878 (0.024)	2.748 (0.165)	57.1 (2.4)	4.666 (0.308)	5.948 (0.308)	71.1 (4.8)
	BrainAnytime	91.9 (2.7)	97.0 (1.6)	81.3 (5.6)	73.3 (1.4)	78.8 (2.2)	70.0 (3.9)	1.920 (0.060)	2.700 (0.100)	55.9 (4.9)	4.256 (0.051)	5.538 (0.051)	74.2 (1.9)
T1+T2+ Flair	BrainIAC	76.6 (2.3)	82.1 (1.4)	62.8 (3.6)	54.5 (2.8)	56.2 (1.4)	54.0 (4.3)	2.578 (0.040)	3.666 (0.068)	35.4 (4.2)	4.953 (0.118)	6.246 (0.046)	67.8 (1.1)
	SAM-Brain3D	76.6 (5.6)	85.8 (4.6)	64.2 (7.5)	59.9 (1.2)	63.4 (1.8)	59.5 (3.3)	2.414 (0.158)	3.292 (0.274)	52.0 (7.6)	3.953 (0.174)	4.953 (0.205)	80.3 (2.0)
	Flex-MoE	77.7 (3.2)	88.8 (1.5)	61.8 (7.4)	47.7 (3.1)	61.6 (2.3)	12.7 (12.6)	2.500 (0.080)	3.400 (0.240)	47.7 (3.5)	4.666 (0.769)	5.743 (0.820)	77.5 (3.3)
	FuseMoE	76.9 (2.2)	83.0 (0.7)	60.5 (6.9)	55.3 (2.1)	58.6 (2.5)	44.5 (0.4)	2.960 (0.180)	4.120 (0.220)	38.4 (6.2)	4.564 (0.051)	5.846 (0.051)	73.1 (0.9)
	MoE-retriever	82.1 (5.1)	86.9 (6.2)	72.9 (7.9)	55.3 (2.2)	62.5 (4.5)	57.8 (12.2)	2.301 (0.062)	3.075 (0.209)	58.3 (3.1)	4.205 (0.308)	5.282 (0.308)	77.3 (4.2)
	BrainAnytime	84.2 (1.7)	90.6 (0.5)	77.0 (5.6)	59.6 (2.6)	65.4 (3.5)	54.0 (8.4)	2.260 (0.040)	3.140 (0.060)	63.3 (3.3)	3.795 (0.256)	4.871 (0.256)	82.6 (0.4)
T1+Flair+ PET	Flex-MoE	90.4 (1.8)	95.3 (0.3)	84.1 (2.7)	59.6 (5.6)	63.8 (11.9)	33.9 (20.5)	2.180 (0.180)	2.960 (0.080)	46.9 (1.0)	4.974 (0.205)	6.205 (0.308)	70.5 (2.2)
	FuseMoE	82.3 (2.3)	92.0 (1.9)	63.7 (6.4)	66.3 (1.7)	77.2 (1.0)	55.9 (3.9)	2.220 (0.220)	3.220 (0.260)	39.0 (6.5)	4.871 (0.051)	6.102 (0.103)	66.2 (1.2)
	MoE-retriever	84.8 (1.5)	92.5 (0.8)	73.1 (4.2)	65.9 (1.7)	72.3 (5.2)	58.1 (5.3)	1.956 (0.085)	2.802 (0.095)	54.2 (2.3)	4.410 (0.359)	5.640 (0.461)	71.3 (4.0)
	BrainAnytime	92.4 (1.5)	97.5 (0.8)	87.2 (2.7)	68.9 (2.9)	77.8 (1.4)	66.2 (1.6)	1.840 (0.060)	2.660 (0.100)	59.4 (2.3)	4.051 (0.154)	5.128 (0.205)	78.6 (1.8)
T1+T2+ Flair+PET	Flex-MoE	89.3 (1.8)	93.6 (0.2)	70.7 (3.3)	65.3 (2.7)	68.3 (1.0)	30.7 (12.5)	2.600 (0.500)	3.460 (0.640)	54.1 (3.1)	4.256 (0.103)	5.384 (0.205)	75.4 (0.6)
	FuseMoE	86.3 (0.6)	87.9 (0.4)	56.0 (1.6)	65.9 (1.7)	73.5 (1.0)	35.1 (3.2)	2.260 (0.340)	3.320 (0.320)	45.2 (7.3)	4.564 (0.051)	5.794 (0.103)	66.3 (1.0)
	MoE-retriever	89.1 (3.6)	90.6 (6.3)	69.9 (10.0)	62.8 (2.3)	64.7 (2.6)	38.9 (3.6)	1.828 (0.041)	2.710 (0.077)	62.9 (1.4)	4.153 (0.410)	5.128 (0.461)	75.8 (4.2)
	BrainAnytime	91.9 (0.9)	95.2 (0.6)	78.3 (1.6)	67.3 (2.6)	70.3 (0.8)	54.2 (5.6)	1.800 (0.080)	2.680 (0.120)	64.5 (2.1)	3.743 (0.154)	4.820 (0.154)	79.6 (0.6)
Average	Flex-MoE	84.2 (2.3)	91.0 (1.9)	74.8 (3.5)	58.9 (4.5)	63.2 (4.8)	40.5 (12.4)	2.452 (0.260)	3.304 (0.300)	42.4 (3.6)	5.015 (0.461)	6.297 (0.564)	70.3 (2.7)
	FuseMoE	82.1 (1.6)	87.5 (0.9)	67.8 (3.8)	62.8 (2.3)	67.4 (1.7)	54.2 (3.3)	2.400 (0.160)	3.384 (0.220)	38.7 (5.6)	4.728 (0.103)	6.041 (0.103)	65.9 (1.4)
	MoE-retriever	83.1 (4.7)	87.2 (6.0)	74.7 (6.3)	60.6 (2.1)	63.6 (4.1)	58.5 (5.5)	2.065 (0.053)	2.868 (0.093)	52.3 (6.8)	4.502 (0.461)	5.682 (0.513)	71.5 (6.1)
	BrainAnytime	89.4 (1.9)	94.7 (1.2)	82.4 (3.7)	67.2 (2.4)	72.5 (1.9)	63.9 (4.6)	1.984 (0.056)	2.768 (0.092)	60.2 (2.6)	4.010 (0.133)	5.158 (0.154)	77.2 (1.1)

settings, with the highest average ACC/AUC/F1 of 89.4/94.7/82.4 for AD diagnosis, outperforming incomplete multimodal learning baselines. Across settings, BrainAnytime ranks first or second on the majority of primary metrics for classification (AUC) and regression (MAE), demonstrating strong effectiveness under arbitrary modality availability. Because the five real-world modality groups in Table 2 are disjoint, cross-group differences may reflect cohort variation rather than modality effects. We thus conduct a within-subject robustness study on the fully observed (T1+T2+FLAIR+PET) test subset, where we synthetically drop modalities to recreate the same five combinations on identical subjects. As additional modalities are provided, performance improves from T1 to full input across all tasks (CN vs. AD F1: 65.0→80.4; CN vs. MCI ACC: 64.2→67.3; MMSE PCC: 61.6→64.5; Age PCC: 75.5→79.6; Fig. 2).

3.4 Ablation Study and Model Interpretability

Table 3 ablates each component. MultiMAE3D pretraining already yields a large gain over training from scratch; adding RCMD or PACM individually further improves results, while combining both (BrainAnytime) achieves the best ACC/F1 on CN vs. AD (89.4/82.7) and ACC/AUC on CN vs. MCI (67.2/72.5) averaged cross all five modality settings. Label efficiency analysis (Fig. 4) confirms reduced reliance on labeled data. We also visualize the PACM dynamic score evolution

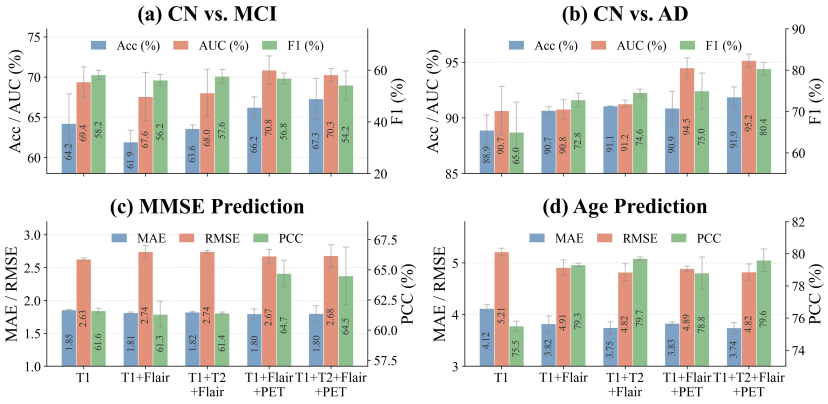
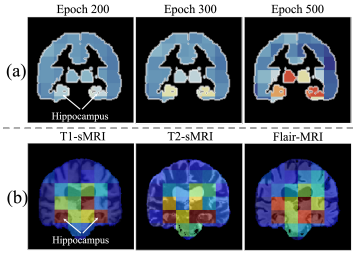


Fig. 2: Modality robustness analysis across four downstream tasks under simulated missing-modality conditions via artificial modality dropout.

Table 3: Ablation study

Task	Method	ACC	AUC	F1
CN vs. AD	ViT-B (scratch)	78.8 (2.2)	84.8 (1.1)	63.3 (5.7)
	Multi-MAE3D	86.9 (0.4)	92.8 (1.9)	75.6 (1.5)
	Multi-MAE3D+RCMD	88.8 (0.3)	94.1 (0.9)	81.3 (0.8)
	Multi-MAE3D+PACM	88.1 (1.6)	94.8 (0.5)	79.9 (3.7)
	BrainAnytime (Ours)	89.4 (0.5)	94.7 (0.5)	82.7 (1.5)
CN vs. MCI	ViT-B (scratch)	59.3 (2.1)	61.4 (1.4)	48.6 (9.2)
	Multi-MAE3D	64.4 (1.8)	68.1 (1.7)	61.2 (3.8)
	Multi-MAE3D+RCMD	65.7 (0.6)	71.3 (0.2)	64.2 (2.0)
	Multi-MAE3D+PACM	66.7 (1.1)	71.5 (1.1)	61.2 (2.4)
	BrainAnytime (Ours)	67.2 (1.4)	72.5 (0.9)	63.9 (3.6)

Fig. 3: (a) Dynamic score in PACM and (b) attention map.



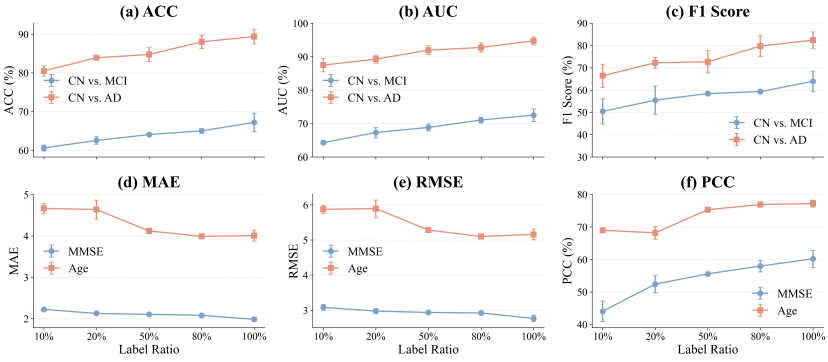


Fig. 4: Label efficiency analysis of BrainAnytime with 10%, 20%, 50%, 80%, and 100% of training data on downstream tasks.

(Fig. 3(a)) and the attention map of a predicted AD sample (Fig. 3(b)), showing that BrainAnytime attends to AD-relevant regions such as the hippocampus.

4 Conclusion

We presented BrainAnytime, a unified framework pretrained on 34,899 3D brain scans that supports brain image analysis under arbitrary modality availability. Through cross-modal distillation and atlas-guided curriculum masking, a single model flexibly accepts whatever imaging is available and consistently outperforms modality-specific, missing-modality, and foundation model baselines across four tasks and five modality settings.

Acknowledgments. This work was partially supported by RGC Collaborative Research Fund (No. C5055-24G), the Start-up Fund of The Hong Kong Polytechnic University (No. P0045999), the Seed Fund of the Research Institute for Smart Ageing (No. P0050946), and Tsinghua-PolyU Joint Research Initiative Fund (No. P0056509), and PolyU UGC funding (No. P0053716).

Disclosure of Interests. The authors have no competing interests.

References

1. Adewole, M., Rudie, J.D., Gbdamosi, A., Toyobo, O., Raymond, C., et al.: The brain tumor segmentation (brats) challenge 2023: Glioma segmentation in sub-saharan africa patient population (brats-africa). ArXiv pp. arXiv–2305 (2023)
2. Beekly, D.L., Ramos, E.M., Lee, W.W., Deitrich, W.D., Jacka, M.E., Wu, J., Hubbard, J.L., Koepsell, T.D., Morris, J.C., Kukull, W.A.: The national Alzheimer's coordinating center (NACC) database: The uniform data set. Alzheimer Disease & Associated Disorders **21**, 249–258 (2007)
3. Braak, H., Alafuzoff, I., Arzberger, T., Kretzschmar, H.A., Tredici, K.D.: Staging of Alzheimer disease-associated neurofibrillary pathology using paraffin sections and immunocytochemistry. Acta Neuropathologica **112**, 389 – 404 (2006)
4. Chen, K., Weng, Y., you Huang, Y., Zhang, Y., Dening, T., et al.: A multi-view learning approach with diffusion model to synthesize FDG PET from MRI T1WI for diagnosis of Alzheimer's disease. Alzheimer's & Dementia **21** (2024)
5. Chételat, G., Arbizu, J., Barthel, H., Garibotto, V., Law, I., Morbelli, S., Van De Giessen, E., Agosta, F., Barkhof, F., Brooks, D.J., et al.: Amyloid-PET and 18F-FDG-PET in the diagnostic investigation of Alzheimer's disease and other dementias. The Lancet Neurology **19**(11), 951–962 (2020)
6. Deng, Z., Wang, H., Huang, Z., Zhang, L., Aviles-Rivero, A.I., Liu, C., He, J., Kourtzi, Z., Schönlieb, C.B.: Brain foundation models with hypergraph dynamic adapter for brain disease analysis. Pattern Recognition p. 112595 (2025)
7. Ding, R., Lu, H., Liu, M.: DenseFormer-MoE: A dense transformer foundation model with mixture of experts for multi-task brain image analysis. IEEE Transactions on Medical Imaging **44**, 4037–4048 (2025)
8. Ellis, K.A., Bush, A.I., Darby, D., De Fazio, D., Foster, J., Hudson, P., Lautenschlager, N.T., Lenzo, N., Martins, R.N., Maruff, P., et al.: The Australian imaging, biomarkers and lifestyle (AIBL) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease. International psychogeriatrics **21**(4), 672–687 (2009)
9. Erdur, A.C., Beischl, C., Scholz, D., Pan, J., Wiestler, B., Rueckert, D., Peeken, J.C.: MultiMAE for brain MRIs: Robustness to missing inputs using multi-modal masked autoencoder. In: International Workshop on Machine Learning in Medical Imaging. pp. 572–582. Springer (2025)

10. Han, X., Nguyen, H., Harris, C., Ho, N., Saria, S.: FuseMoE: Mixture-of-experts transformers for fleximodal fusion. Advances in Neural Information Processing Systems **37**, 67850–67900 (2025)
11. Hara, K., Kataoka, H., Satoh, Y.: Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet? In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 6546–6555 (2018)
12. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15979–15988 (2022)
13. Hu, W., Guan, Z., Yang, P., Li, J., Liu, Y., Gan, S., Cai, T., Zhang, A., Zhang, T., Qu, J., et al.: Anatomy-guided multimodal graph networks for Alzheimer's disease: Integrative analysis of cross-modal brain connectivity signatures. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 66–75. Springer (2025)
14. Jack Jr, C.R., Andrews, J.S., Beach, T.G., Buracchio, T., Dunn, B., Graf, A., Hansson, O., Ho, C., Jagust, W., McDade, E., et al.: Revised criteria for diagnosis and staging of Alzheimer's disease: Alzheimer's association workgroup. Alzheimer's & Dementia **20**(8), 5143–5169 (2024)
15. Jang, J., Hwang, D.: M3T: three-dimensional medical image classifier using multi-plane and multi-slice transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20686–20697 (2022)
16. Leng, Y., Cui, W., Peng, Y., Yan, C., Cao, Y., Yan, Z., Chen, S., Jiang, X., Zheng, J., Initiative, A.D.N., et al.: Multimodal cross enhanced fusion network for diagnosis of Alzheimer's disease and subjective memory complaints. Computers in Biology and Medicine **157**, 106788 (2023)
17. Rui, S., Chen, L., Tang, Z., Wang, L., Liu, M., Zhang, S., Wang, X.: Multi-modal vision pre-training for medical image analysis. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5164–5174 (2025)
18. Sperling, R.A., Donohue, M.C., Raman, R., Sun, C.K., Yaari, R., Holdridge, K., Siemers, E., et al.: Association of factors with elevated amyloid burden in clinically normal older individuals. JAMA neurology **77**(6), 735–745 (2020)
19. Tak, D., Garomsa, B.A., Zapaishchykova, A., Chaunzwa, T.L., Climent Pardo, J.C., Ye, Z., Zielke, J., Ravipati, Y., Pai, S., et al.: A generalizable foundation model for analysis of human brain MRI. Nature Neuroscience pp. 1–12 (2026)
20. Weiner, M.W., Harvey, D., Hayes, J., Landau, S.M., Aisen, P.S., Petersen, R.C., Tosun, D., Veitch, D.P., Jack Jr, C.R., Decarli, C., et al.: Effects of traumatic brain injury and posttraumatic stress disorder on development of Alzheimer's disease in vietnam veterans using the Alzheimer's disease neuroimaging initiative: preliminary report. Alzheimer's & Dementia: Translational Research & Clinical Interventions **3**(2), 177–188 (2017)
21. Weiner, M.W., Veitch, D.P., Aisen, P.S., Beckett, L.A., Cairns, N.J., Green, R.C., Harvey, D., Jack, C.R., Jagust, W., Liu, E., et al.: The Alzheimer's disease neuroimaging initiative: a review of papers published since its inception. Alzheimer's & Dementia **9**(5), e111–e194 (2013)
22. Yang, G., Du, K., Yang, Z., Du, Y., Cheung, E.Y.W., Zheng, Y., Yang, M., Kourtzi, Z., Schönlieb, C.B., Wang, S., Initiative, A.D.N.: ADFound: A foundation model for diagnosis and prognosis of Alzheimer's disease. IEEE Journal of Biomedical and Health Informatics **29**(11), 8395–8408 (2025)
23. Yin, L., Ye, C., Liu, T., Wu, J., Yan, T.: UniCross: Balanced multimodal learning for Alzheimer's disease diagnosis by uni-modal separation and metadata-guided

cross-modal interaction. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 638–648. Springer (2025)
24. Yun, S., Choi, I., Peng, J., Wu, Y., Bao, J., Zhang, Q., Xin, J., et al.: Flex-MoE: Modeling arbitrary modality combination via the flexible mixture-of-experts. Advances in Neural Information Processing Systems **37**, 98782–98805 (2025)
25. Yun, S., Xin, J., Choi, I., Peng, J., Ding, Y., Long, Q., Chen, T.: Generate, then retrieve: Addressing missing modalities in multimodal learning via generative AI and MoE. In: Workshop on Large Language Models and Generative AI for Health at AAAI 2025 (2025)
26. Zhang, J., He, X., Liu, Y., Cai, Q., Chen, H., Qing, L.: Multi-modal cross-attention network for Alzheimer's disease diagnosis with multi-modality data. Computers in Biology and Medicine **162**, 107050 (2023)
27. Zhang, X., Ou, N., Basaran, B.D., Visentin, M., Qiao, M., Gu, R., Matthews, P.M., Liu, Y., Ye, C., Bai, W.: A foundation model for lesion segmentation on brain MRI with mixture of modality experts. IEEE Transactions on Medical Imaging **44**(6), 2594–2604 (2025)