MODELING LOCAL, GLOBAL, AND CROSS-MODAL CONTEXT IN MULTIMODAL 3D MRI

Anonymous authors
Paper under double-blind review

ABSTRACT

Brain MRIs pose a fundamental challenge for machine learning: models must learn from high-dimensional 3D data spanning multiple co-registered modalities, despite the limited sample sizes typical of neuroimaging studies relative to the diversity in anatomy, pathology, and acquisition conditions. While multimodal imaging provides complementary information critical for clinical interpretation, effectively integrating these signals remains difficult. We propose **M**ultimodal **I**ntra- and **C**ross-Context **Vi**sion **T**ransformer (MICViT), a 3D Vision Transformer (ViT) that explicitly models both modality-specific representations and cross-modal interactions across local and global contexts. Concretely, MICViT combines four attention mechanisms: modality-specific local and global attention for intra-modal feature learning, and cross-modal local and global attention to capture interactions between modalities. We evaluate MICViT on brain age prediction across four heterogeneous datasets (UK Biobank, SOOP, ADNI, Cam-CAN) using multiple MRI modalities (*e.g.,* T1, FLAIR, DWI, SWI). MICViT consistently achieves better results than established Convolutional Neural Network (CNN) and Transformer baselines in 3D settings. Notably, it benefits more strongly from multimodal inputs, yielding larger performance gains as additional modalities are incorporated. Beyond age prediction, we further evaluate the learned representations on disease-onset prediction in UK Biobank and Alzheimer's disease classification in ADNI, where MICViT shows competitive to strong performance across clinically relevant endpoints. These results suggest that explicitly modeling intra- and cross-modal interactions is a promising direction for representation learning in multimodal neuroimaging. Code is available at https://anonymous.4open.science/r/MICViT-C057.

1 INTRODUCTION

Multimodal neuroimaging provides radiologists with non-invasive, high-resolution insights into the human brain, facilitating the identification of pathologies and supporting precise clinical interventions. In radiological practice, image interpretation is inherently multi-scale and multimodal: localized findings within a specific brain region are not assessed in isolation but are continuously contextualized with the global brain structure within the same modality, as well as with corresponding regions across other MRI modalities (Gillies et al., 2016). In diagnostic tasks, a single MRI modality (e.g., T1-weighted MRI (T1)) can provide valuable information about anatomical structures. However, its diagnostic potential is substantially enhanced when combined with additional modalities. For instance, T2-weighted MRI (T2) is sensitive to increased water content, making it particularly effective for visualizing edema and inflammation, Fluid-Attenuated Inversion Recovery (FLAIR) improves visibility of white matter lesions, Diffusion-weighted imaging (DWI) characterizes white matter tracts, and Susceptibility Weighted Imaging (SWI) reveals microbleeds and iron deposits (Haacke et al., 1999). In clinical practice, radiologists routinely integrate such complementary information to improve diagnostic accuracy. Similarly, machine learning models that effectively combine multimodal brain MRI data can achieve more robust and accurate predictions (Schulz et al., 2024).

Recent advances in transformer-based architectures have demonstrated strong potential for modeling long-range dependencies in medical imaging (Chen et al., 2021). In particular, ViTs enable flexible

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

context aggregation beyond the locality constraints of CNNs (Takahashi et al., 2024). However, existing approaches face two key limitations when applied to multimodal 3D MRI data.

First, while prior work has explored cross-modal interactions, many architectures do not explicitly disentangle intra-modality and cross-modality processing within a unified and structured 3D MRI architecture. Early fusion combines modality-specific features before predictive modeling, whereas late fusion aggregates independently generated predictions without explicitly modeling feature-level interactions across modalities.(Huang et al., 2020). This is a particular limitation of late fusion in multimodal MRI, where spatially aligned modalities provide complementary tissue contrasts at corresponding anatomical locations that can be exploited through fine-grained cross-modal feature interactions. At the same time, radiologists interpret local findings in the context of global brain anatomy, an interplay that is not explicitly modeled in prior work. Second, while unimodal 2D architectures have demonstrated the importance of jointly modeling local and global context, such designs are rarely extended to multimodal 3D MRI, limiting the ability to capture fine-grained anatomical structures alongside global patterns across modalities (Liu et al., 2021; Hatamizadeh et al., 2023). This principle extends beyond healthcare; for example, in autonomous driving, the fusion of camera, LiDAR, and radar data enables more reliable perception under challenging environmental conditions (Chen et al., 2017).

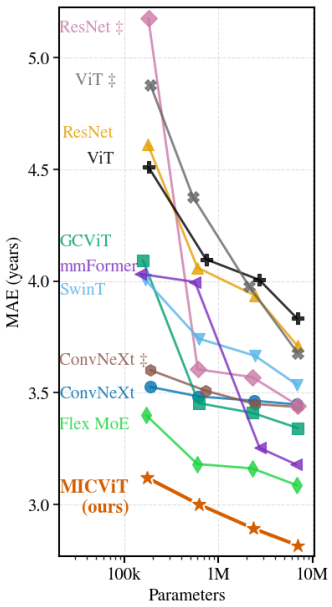

Figure 1: Scaling behavior on UKBB as a function of model size using MRI modalities (T1, FLAIR, SWI, and DWI). MICViT consistently achieves the lowest error across model sizes. Late-fusion variants are marked with ‡.

Motivated by clinical radiological reasoning, our approach explicitly models intra- and cross-modality interactions across both local and global contexts within a unified architecture. We introduce the Multimodal Intra- and Cross-Context Vision Transformer (MICViT), a novel multimodal ViT for 3D brain MRI that captures how radiologists integrate localized findings with global brain context across modalities. While inspired by the unimodal Global Context Vision Transformer (GC ViT) (Hatamizadeh et al., 2023) and its combination of 2D image-based local self-attention with global context aggregation, our model incorporates four dedicated attention mechanisms for multimodal 3D MRI data: Separated Local (SL) and Separated Global (SG) for modality-specific representation learning, as well as Cross Local (CL) and Cross Global (CG) for explicit cross-modal interaction, enabling joint modeling of local structure, global context, and cross-modal dependencies. Experiments across multiple neuroimaging datasets, including the large-scale UK Biobank (UKBB) (Sudlow et al., 2015), suggest that performance in multimodal 3D neuroimaging is not solely determined by model size or data scale, but critically depends on how spatial and cross-modal interactions are modeled.

The main contributions of this work are as follows:

- We introduce MICViT, a multimodal 3D ViT for brain MRI that explicitly separates modality-specific representation learning from cross-modal feature interaction.
- We propose four complementary attention mechanisms, SL, SG, CL, and CG, to jointly model local structure, global anatomical context, and spatially aligned cross-modal dependencies.
- We demonstrate across neuroimaging datasets, model scales, and modality configurations that structured intra- and cross-modal attention yields robust performance and stronger gains from additional MRI modalities than standard 3D baselines.

2 Related Work

Multimodal Learning in Brain MRI Multimodal learning can improve representation quality by combining complementary information across data sources (Ngiam et al., 2011; Baltrušaitis et al.,

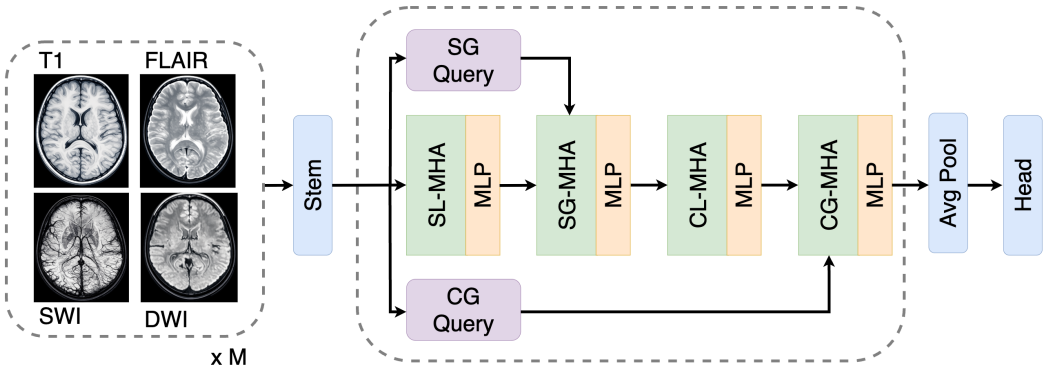

Figure 2: Architecture of the proposed MICViT. Multiple imaging modalities (*e.g.,* T1, FLAIR, SWI, DWI) are processed via a stem that generates patch embeddings. The model sequentially applies Separated Local (SL) and Separated Global (SG) attention for intra-modality representation learning, followed by Cross Local (CL) and Cross Global (CG) attention to enable explicit cross-modality fusion. Dedicated global query generators are used for SG and CG.

2017; Sun et al., 2024). This is particularly relevant in brain MRI, where different contrasts and imaging modalities capture distinct aspects of anatomy, tissue properties, and pathology (Huang et al., 2020; Luo et al., 2024; Li et al., 2024; Jirsaraie et al., 2023). Prior work has shown that multiple imaging modalities can substantially improve phenotype prediction, sometimes yielding gains comparable to increasing dataset size (Schulz et al., 2024). However, many multimodal neuroimaging models rely on input concatenation, shared encoders, or late fusion, which limits their ability to explicitly model spatially aligned intra- and cross-modality dependencies throughout the network.

Convolutional Neural Networks for Brain MRI Analysis CNNs remain strong baselines for MRI analysis due to their locality and translation-related inductive biases (Kayhan & Gemert, 2020; Mienye et al., 2025). Early 2D CNN approaches reduce computation by processing slices or patches, but only approximate volumetric anatomical context (Gupta et al., 2021; Bashyam et al., 2020; Jönemo et al., 2023). 3D CNNs better preserve spatial structure and have achieved strong performance in brain age prediction and related neuroimaging tasks (Peng et al., 2021; Dinsdale et al., 2021; Lee et al., 2022; Leonardsen et al., 2022). Yet many CNN-based methods remain unimodal or combine modalities through input stacking, multi-stream designs, or isolated fusion modules (He et al., 2021; Liu et al., 2022). As a result, cross-modal interaction is often not modeled explicitly and repeatedly across spatial scales.

Vision Transformers in Medical Imaging Transformers provide a natural mechanism for modeling long-range dependencies through self-attention (Vaswani et al., 2023; Dosovitskiy et al., 2021) and have been widely adopted in medical imaging (Chen et al., 2021; Li et al., 2022). Several MRI transformer models approximate volumetric structure using 2D slices or multi-view representations, which limits explicit 3D context modeling (He et al., 2022; Zhang & Jiang, 2024; Mazher et al., 2025). Other approaches support multimodal inputs, but often combine modalities after modality-specific encoding, through channel stacking, or via shared encoders (Zhao et al., 2024; Tak et al., 2026). SwiFT extends shifted-window attention to 4D fMRI, but targets temporal dynamics rather than multimodal integration (Liu et al., 2021; Kim et al., 2023). In contrast, MICViT is designed to jointly model full 3D spatial structure and explicitly separate intra-modal local/global processing from cross-modal local/global interaction.

Multimodal transformer and mixture-of-experts baselines Flex Mixture of Experts (Flex MoE) (Yun et al., 2024) and multimodal Medical Transformer (mmFormer) (Zhang et al., 2022) represent two strong multimodal baselines: Flex MoE uses expert specialization to handle heterogeneous modality contributions, while mmFormer uses transformer-based fusion to integrate information across modalities. These designs are well suited for multimodal learning and are stronger references than simple early- or late-fusion models. However, they do not explicitly disentangle modality-specific and cross-modal processing across both local and global 3D contexts.

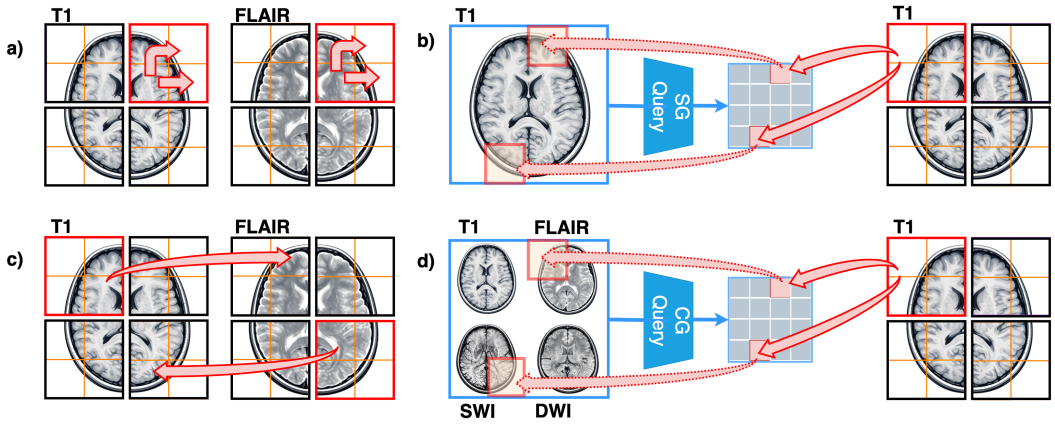

Figure 3: Illustration of the four attention types operating on latent 3D representations: (a) SL performs local self-attention within spatial windows independently for each modality; (b) SG aggregates global context within each modality via shared global queries; (c) CL applies local cross-attention across modalities within corresponding spatial windows; (d) CG models global cross-modal interactions using shared global queries across modalities.

3 METHODS

3.1 IMPLEMENTATION OF MICVIT

Let $x \in \mathbb{R}^{B\times M\times H\times W\times D}$ denote a batch of volumetric inputs with B samples, M MRI modalities, and spatial resolution $H \times W \times D$. Our MICViT extends the original GC ViT to jointly model modality-specific structure and cross-modal interactions in 3D neuroimaging (Figure 2). Specifically, each modality is first processed independently by a modality-specific 3D patch embedding stem layer, producing an initial token tensor $z^{(0)} \in \mathbb{R}^{B\times M\times H_0\times W_0\times D_0\times C_0}$, where $C_0 = \dim$ and (H_0, W_0, D_0) denote the patch-level spatial resolution (see Section 7.4 for details). The resulting embeddings are then partitioned into non-overlapping 3D windows of size $W_d \times W_h \times W_w$. The MICViT alternates between four attention mechanisms designed to disentangle intra-modality and cross-modality dependencies at both local and global scales: SL, SG, CL, and CG attention.To enable a controlled and systematic analysis, we therefore adopt a single-block architecture with four attention layers. Our multimodal design employs two complementary query generators: (i) a separated global query generator (SG query) that extracts modality-specific global context independently per modality, and (ii) a cross-modal global query generator (CG query) that fuses modalities along the channel dimension and projects the resulting shared representation back into the per-modality attention space. Formally, these modules produce q_{sep} and q_{cross}, which parameterize the SG and CG blocks, respectively. This enables the model to retain modality-specialized representations while explicitly modeling cross-modal dependencies. Each transformer block follows a standard residual structure consisting of normalization, window-based attention, residual connection, and multilayer perceptron (MLP) update, with stochastic depth applied across layers. After the final block, the representation $z^{(1)} \in \mathbb{R}^{B\times M\times H_0\times W_0\times D_0\times C_0}$ is normalized and reshaped by merging the modality and channel dimensions to obtain $\tilde{z} \in \mathbb{R}^{B\times(MC_0)\times H_0\times W_0\times D_0}$. Global average pooling yields a compact subject-level embedding $h \in \mathbb{R}^{B\times MC_0}$, which is fed into a multitask prediction head. Across model scales, we vary the base embedding width and number of attention heads while keeping the architectural structure fixed (Table 7).

3.2 ATTENTION TYPES OF MICVIT

Separated Local (SL) Attention. Local attention is applied independently within each modality, restricting interactions to spatially neighboring patches of size $W_d \times W_h \times W_w$ and preventing any cross-modal information exchange (Figure 3a). For each modality $k \in \{1, \dots, M\}$ and window $w \in \mathcal{W}$, let $Z_w^{(k)} \in \mathbb{R}^{N_w\times C}$ denote the corresponding tokens. Queries, keys, and values are obtained via linear projections: $Q_w^{(k)}, K_w^{(k)}, V_w^{(k)} \in \mathbb{R}^{N_w\times C}$. Here, C denotes the token dimension, and d_h

denotes the per-head channel dimension. Self-attention is then computed independently for each modality, window, and attention head h:

$$\text{Attention}_w(Q_w^{(k),h}, K_w^{(k),h}, V_w^{(k),h}) = \text{softmax}\left(\frac{Q_w^{(k),h}(K_w^{(k),h})^\top}{\sqrt{d_h}} + B_h^{\text{spatial}}\right) V_w^{(k),h}. \quad (1)$$

We incorporate a 3D relative positional encoding, where the bias B_h^{spatial} is derived from a learnable table $\hat{B} \in \mathbb{R}^{(2W_d-1)\times(2W_h-1)\times(2W_w-1)}$, indexed by relative offsets along depth, height, and width.

Separated Global (SG) Attention. To incorporate global contextual information, we extend the global attention mechanism of GC ViT to the multimodal 3D setting (Figure 3b). While local window attention restricts interactions to spatial neighborhoods, SG enables each window to access modality-specific global context via a shared set of global query tokens. For each modality $k \in \{1, \ldots, M\}$, we generate a modality-specific global query tensor $Q_g^{(k)} \in \mathbb{R}^{N_q \times C}$ from the corresponding modality tokens. These queries are reshaped per head as $Q_g^{(k),h} \in \mathbb{R}^{N_q \times d_h}$ and shared across all local windows of the same modality. The global queries are broadcast to match the number of windows, yielding an effective batch size $B^* = B \times |\mathcal{W}|$. Importantly, the global queries are computed once per modality and sample, and then reused across all windows: $Q_{g,w}^{(k)} = Q_g^{(k)}, \quad \forall w \in \mathcal{W}$. Keys and values are computed independently within each local window: $K_w^{(k),h}, V_w^{(k),h} \in \mathbb{R}^{N_w \times d_h}$. The global queries attend to local window tokens within the same modality:

$$\text{Attention}_w(Q_g^{(k),h}, K_w^{(k),h}, V_w^{(k),h}) = \text{softmax}\left(\frac{Q_g^{(k),h}(K_w^{(k),h})^\top}{\sqrt{d_h}} + B_h^{\text{spatial}}\right) V_w^{(k),h}. \quad (2)$$

Cross Local (CL) Attention. To explicitly model and fuse interactions between multiple 3D data modalities, we extend local window attention to jointly process tokens from all modalities within each spatial region (Figure 3c). Given modality-specific tokens $Z^{(k)}$, $k \in \{1, \ldots, M\}$, we construct a multimodal representation for each window w by concatenating tokens across modalities: $Z_w = [Z_w^{(1)}, \ldots, Z_w^{(M)}]$, resulting in a sequence of MN_w tokens. Attention is computed over this joint representation, enabling both intra-modality (self-attention within $Z_w^{(k)}$) and cross-modality (cross-attention across $Z_w^{(k)}$ and $Z_w^{(k')}$ for $k \neq k'$) interactions within each local window. Queries, keys, and values are obtained via linear projections: $Q_w, K_w, V_w \in \mathbb{R}^{(MN_w)\times C}$. To model both spatial and modality-dependent relationships, we employ a factorized relative position bias:

$$(B_h^{\text{multi}})_{i,j} = (B_h^{\text{spatial}})_{p_i,p_j} + (B_h^{\text{modality}})_{m_i,m_j}, \quad (3)$$

where p_i and m_i denote the spatial position within the window and the modality index of token i, respectively. The spatial component $B_h^{\text{spatial}} \in \mathbb{R}^{N_w \times N_w}$ is shared across modalities, while $B_h^{\text{modality}} \in \mathbb{R}^{M \times M}$ captures modality-specific interactions.

The attention operation for a window w and head h is given by:

$$\text{Attention}_w(Q_w^h, K_w^h, V_w^h) = \text{softmax}\left(\frac{Q_w^h(K_w^h)^\top}{\sqrt{d_h}} + B_h^{\text{multi}}\right) V_w^h. \quad (4)$$

Cross Global (CG) Attention. Analogous to cross-local attention, we construct a multimodal window $Z_w = [Z_w^{(1)}, \ldots, Z_w^{(M)}] \in \mathbb{R}^{(MN_w)\times C}$. In cross-global attention, queries are not modality-specific. Instead, modality features are concatenated along the channel dimension to generate a single shared global query tensor, which is projected per head and replicated across modalities, yielding identical modality slices. Thus, queries are shared, while keys and values are computed from the multimodal window tokens.

For each head h, let $Q_g^h \in \mathbb{R}^{(MN_w)\times d_h}, \quad K_w^h, V_w^h \in \mathbb{R}^{(MN_w)\times d_h}$. Attention is computed as

$$\text{Attention}_w(Q_g^h, K_w^h, V_w^h) = \text{softmax}\left(\frac{Q_g^h(K_w^h)^\top}{\sqrt{d_h}} + B_h^{\text{multi}}\right) V_w^h.$$

Table 1: Dataset characteristics across Cam-CAN, ADNI, SOOP, and UKBB. For an overview of MRI modalities and their clinical relevance, see Appendix Table 17

	Cam-CAN	**ADNI**	**SOOP**	**UKBB**
n	613	2,384	1,062	41,404
Women (%)	49.4	51.3	47.5	52.9
Age (yr)	54.7 ± 18.6	74.5 ± 7.6	63.8 ± 13.7	63.9 ± 7.7
Modalities	T1, T2, DWI, MTI	T1, FLAIR, STAR, DTI	T1, FLAIR, TRACE, ADC	T1, FLAIR, SWI, DWI

4 EXPERIMENTS

Datasets We evaluate our method on four complementary multimodal brain MRI cohorts (Table 1): the population-based UKBB (Sudlow et al., 2015), the clinical stroke cohort Stroke Outcome Optimization Project (SOOP) (Absher et al., 2024), the lifespan cohort Cambridge Centre for Ageing and Neuroscience (Cam-CAN) (Shafto et al., 2014; Taylor et al., 2017), and the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort (Petersen et al., 2010). Together, these datasets cover diverse populations, age distributions, clinical conditions, and modality configurations. MRI data were preprocessed using a pipeline: MRI volumes were denoised and then underwent bias field correction, skull stripping, and intensity normalization. All scans were co-registered to MNI space to ensure spatial consistency. Due to the high computational cost of volumetric data, volumes were downsampled from 224^3 to 64^3 resolution unless stated otherwise (see Appendix 7.1 for details).

Prediction tasks Our first evaluation task is brain age prediction, which provides a continuous, well-established benchmark for representation learning in neuroimaging (Dular & Špiclin, 2024) and enables comparable evaluation across all four cohorts. To assess whether the learned multimodal representations transfer beyond age regression, we additionally evaluate two clinically oriented downstream tasks: disease-onset prediction in UKBB using survival models (Steinfeldt et al., 2025), and binary Alzheimer's disease classification (CN vs. AD) in ADNI (Cuingnet et al., 2011). For the ADNI classification task, we report results both on the complete four-modality subset and across participants with heterogeneous modality availability.

Table 2: Ablation study of MRI modality combinations in the UKBB.

Modalities	**MAE ↓ ± SD**	**Params (M) ↓**	**Time (min/ep.) ↓**
GCViT			
SWI	4.31 ± 0.06	2.28	30.42
T1	3.81 ± 0.10	2.28	28.02
FLAIR	3.44 ± 0.05	2.28	30.39
DWI	3.62 ± 0.07	2.28	30.28
FLAIR+DWI	3.34 ± 0.05	3.56	39.27
T1+FLAIR+DWI	3.38 ± 0.02	5.01	61.46
all 4 modalities	3.34 ± 0.06	6.96	86.33
ConvNeXt			
SWI	4.51 ± 0.03	2.34	24.27
T1	3.97 ± 0.06	2.34	24.21
FLAIR	3.72 ± 0.03	2.34	24.41
DWI	3.92 ± 0.04	2.34	24.29
FLAIR+DWI	3.56 ± 0.02	3.58	35.64
T1+FLAIR+DWI	3.48 ± 0.07	5.00	50.04
all 4 modalities	3.53 ± 0.03	6.77	67.67
MICViT			
SWI	4.12 ± 0.20	2.28	30.73
T1	3.78 ± 0.11	2.28	30.93
FLAIR	3.28 ± 0.10	2.28	31.77
DWI	3.17 ± 0.02	2.28	28.34
FLAIR+DWI	3.08 ± 0.09	3.55	63.61
T1+FLAIR+DWI	2.87 ± 0.14	5.04	75.74
all 4 modalities	$\mathbf{2.82} \pm 0.06$	6.97	89.45

Optimization For training and evaluation, we follow best practices for multimodal learning (Rheude et al., 2025). All models are trained with Schedule-Free Learning (Defazio et al., 2024), Gap-Guided Sharpness-Aware Minimization (GSAM) (Zhuang et al., 2022), bfloat16 precision, and early stopping based on validation loss. Transformer-based models use pre-layer normalization for training stability (Xiong et al., 2020), and attention-based models are initialized with mimetic initialization (Trockman & Kolter, 2023). For MICViT and all standard baseline models, we perform a grid search over learning rates 10^{-2} to 10^{-5} with a fixed weight decay of 0.05. To establish a particularly strong CNN reference baseline, we additionally perform Bayesian hyperparameter sweeps for ConvNeXt with 100 runs across multiple model scales (small, medium, base, and large; see Appendix 9 for details). The best ConvNeXt configuration from this sweep is used as the CNN reference point in the main comparisons. Model selection for all architectures is based on the lowest validation MAE. All experiments use 5-fold cross-validation. In each fold, 80% of the data is used for training, 10% for validation, and 10% for testing, ensuring subject-level separation across splits. Models are trained and evaluated independently on each dataset.

Table 3: Cross-dataset evaluation on SOOP, Cam-CAN, ADNI, and UKBB. Results are reported as mean MAE ± standard deviation across five folds. All datasets were evaluated at an image resolution of 64^3; UKBB was additionally evaluated at the full resolution of 225^3. Late-fusion variants are marked with ‡.

Model	**SOOP** ($n = 1{,}062$)	**Cam-CAN** ($n = 613$)	**ADNI** ($n = 2{,}384$)	**UKBB** ($n = 41{,}404$)	**UKBB** ($n = 41{,}404$)
	MAE ↓ 64^3 voxel	**MAE** ↓ 64^3 voxel	**MAE** ↓ 64^3 voxel	**MAE** ↓ 64^3 voxel	**MAE** ↓ 225^3 voxel
Swin-T	9.127 ± 0.233	10.575 ± 0.924	4.788 ± 0.492	3.536 ± 0.076	3.083 ± 0.209
ResNet	8.833 ± 1.485	8.718 ± 1.286	5.311 ± 0.444	3.711 ± 0.168	3.402 ± 0.471
ViT	8.807 ± 1.785	8.939 ± 1.046	4.652 ± 0.206	3.832 ± 0.146	2.920 ± 0.090
ResNet ‡	7.957 ± 0.931	8.574 ± 1.167	5.289 ± 0.408	3.443 ± 0.123	3.032 ± 0.227
Flex-MoE	7.380 ± 0.697	6.525 ± 0.926	4.703 ± 0.370	3.086 ± 0.129	2.772 ± 0.168
ConvNeXt	7.326 ± 0.548	8.221 ± 0.723	4.334 ± 0.294	3.447 ± 0.039	2.690 ± 0.159
ViT ‡	7.294 ± 0.493	9.123 ± 0.819	5.004 ± 0.519	3.677 ± 0.066	2.624 ± 0.104
ConvNeXt ‡	6.774 ± 0.125	6.971 ± 0.406	4.271 ± 0.198	3.436 ± 0.062	2.778 ± 0.142
GC ViT	6.757 ± 0.135	7.849 ± 0.778	4.454 ± 0.195	3.342 ± 0.065	2.479 ± 0.050
mmFormer	6.717 ± 0.144	6.431 ± 0.600	4.194 ± 0.334	3.179 ± 0.084	2.525 ± 0.124
MICViT	$\mathbf{6.671 \pm 0.102}$	$\mathbf{6.428 \pm 0.690}$	$\mathbf{4.165 \pm 0.373}$	$\mathbf{2.816 \pm 0.062}$	$\mathbf{2.404 \pm 0.016}$

5 RESULTS

Performance in multimodal 3D MRI depends on modeling spatial and cross-modal interactions We first study how the combination and ordering of attention mechanisms affect performance in multimodal 3D MRI for brain age prediction. We train on T1, FLAIR, SWI, and DWI, where it is not obvious which information the model should prioritize at different stages of processing. We find that the sequence of intra- and cross-modal attention is critical. Evaluating all permutations of SL, SG, CL, and CG attention on UKBB, we observe that the full configuration (SL, SG, CL, CG) consistently achieves the best performance (Appendix 5). This indicates that the gains do not arise from a single dominant component, but from their interaction.

MICViT scales effectively with an increasing number of MRI modalities Building on this, we next analyze how models benefit from increasing numbers of MRI modalities (Table 2). We evaluate single modalities, and select the best-performing pair (FLAIR + DWI), as well as extensions to three and four modalities. In the unimodal setting, MICViT also achieves the lowest MAE for every individual contrast, outperforming baselines (GC ViT and ConvNeXt) for SWI, T1, FLAIR, and DWI . We find that performance improves consistently as more modalities are added, confirming the benefit of complementary imaging information. In this setting, we demonstrate that MICViT benefits most from the incremental addition of modalities, outperforming ConvNeXt as a strong CNN baseline and GC ViT as a transformer baseline across all configurations. The performance gap increases with the number of modalities, indicating that MICViT more effectively exploits cross-modal complementarities (Appendix 6).

Robust performance across multiple datasets To evaluate robustness across heterogeneous data, all models are trained and evaluated independently on datasets that differ in sample size and acquisition characteristics (Table 3). Overall, MICViT demonstrates robust and consistent performance across both large-scale and smaller external cohorts. It achieves the lowest MAE on all datasets and significantly improves over several CNN-, transformer-, and late-fusion baselines, despite the higher variability expected in smaller cohorts such as SOOP, Cam-CAN, and ADNI (Appendix: Table 10, 11, 12). The particularly clear gains on UKBB indicate that MICViT's architectural design is especially effective when sufficient multimodal data are available to exploit both modality-specific structure and cross-modal complementarities.

Consistent scaling gains from model capacity and input resolution Finally, we evaluate scaling across model capacity and input resolution. MICViT consistently performs best from small to large configurations, with its advantage persisting as parameter count increases (Appendix Figure 7). At

Table 4: Ablation study on attention mechanisms. Individual and combined attention components are removed from MICViT.

Configuration	MAE ↓ ± SD	Params (M) ↓	Time (min/ep.) ↓
Full MICViT	3.12 ± 0.19	0.18	38.84
w/o SL	3.14 ± 0.21	0.16	43.30
w/o CL	3.30 ± 0.31	0.16	43.31
w/o SG	3.32 ± 0.46	0.16	43.30
w/o CG	3.33 ± 0.24	0.16	43.60
w/o SL–SG	3.32 ± 0.46	0.15	35.97
w/o SL–CL	3.44 ± 0.26	0.15	38.97
w/o CL–CG	3.45 ± 0.27	0.15	39.27
w/o SG–CL	3.54 ± 0.45	0.15	40.42
w/o SG–CG	4.06 ± 0.20	0.15	29.03
w/o SL–SG–CL	3.41 ± 0.49	0.14	43.62
w/o SL–SG–CG	4.28 ± 0.39	0.14	44.91
w/o SG–CL–CG	4.35 ± 0.14	0.14	45.08
w/o SL–CL–CG	4.43 ± 0.99	0.14	45.92

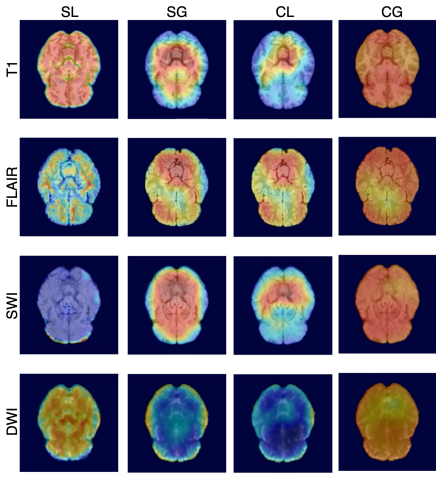

Figure 4: Grad-CAM (Selvaraju et al., 2020) visualizations for SL, SG, CL, and CG attention across T1, FLAIR, SWI, and DWI, averaged over participants in the UKBB test set.

full resolution (225^3), MICViT achieves the lowest MAE, modestly outperforming the strongest competing transformer and more clearly outperforming Flex MoE, CNN-based models, and late-fusion approaches (Table 3, Appendix 15).

Combining global and local attention across and within modalities is critical for performance Building on these observations, we perform systematic attention ablations by isolating component contributions using UKBB MRI data (Table 4). Single attention types consistently underperform, with SG yielding the highest error and variance, while CG performs substantially better, underscoring the importance of global cross-modal context. Local variants (SL, CL) provide moderate improvements but remain inferior to CG in isolation. Combining attention mechanisms consistently reduces error, with configurations including global attention (SG, CG) clearly outperforming purely local variants. Performance further improves with higher-order combinations, peaking when all four attention types are used. Adding local attention to global components yields incremental gains, indicating complementary roles in refining spatial detail and enabling cross-modal fusion. Overall, performance scales with attention diversity, with progressively smaller gains as the model approaches the full configuration.

Attribution maps reveal complementary attention across modalities The attribution patterns indicate complementary roles of the attention mechanisms across modalities and reveal patterns associated with brain aging (Figure 4). SL captures fine-grained, symmetric anatomy and differentiates gray and white matter. It emphasizes widespread white matter in T1, subcortical, brainstem, and cerebellar regions in FLAIR, and white matter and brainstem structures in DWI, while SWI contributes only marginally at the local level (Wardlaw et al., 2013; DeCarli et al., 2005; Raz et al., 2005; Cox et al., 2016; Beck et al., 2021; Farokhian et al., 2017). SG shifts the focus toward coarse anatomical organization, particularly periventricular, brainstem, and cerebellar regions associated with global atrophy and ventricular enlargement (Habes et al., 2016; Bouhrara et al., 2021; Wang et al., 2024). CL aligns spatially consistent regions across modalities, whereas CG aggregates distributed multimodal information across the brain, reflecting the multi-tissue and multi-scale nature of brain aging (Chen et al., 2026a;b).

MICViT captures relevant representations across disease endpoints Based on UKBB, we evaluated disease-onset prediction for 38 diagnostic endpoints selected from neurological, neurodegenerative, neurovascular, psychiatric, and endocrine conditions using survival modeling and compared MICViT with baseline models (further details in Appendix 16). MICViT achieved the highest mean Concordance index (C-index) (Harrell et al., 1982) for 24 of 38 diagnoses, with particularly high per-

formance for neurodegenerative endpoints, including Alzheimer's disease (C-index: 0.86 ± 0.03), dementias (C-index: 0.81 ± 0.03), and dementias and cerebral degeneration (C-index: 0.76 ± 0.05). Strong performance was also observed for neurovascular endpoints such as hypertension (C-index: 0.74 ± 0.097), hemorrhagic stroke (C-index: 0.64 ± 0.11), and transient cerebral ischemic attack (TIA; C-index: 0.63 ± 0.08), as well as psychiatric endpoints such as major depressive disorder (C-index: 0.65 ± 0.09) and mood affective disorders (C-index: 0.56 ± 0.01). These results suggest that MICViT's local-global cross-modal design captures clinically meaningful imaging variation that generalizes across neurodegenerative, neurovascular, psychiatric, and systemic disease risks.

Alzheimer's disease classification under heterogeneous modality availability Using ADNI, we evaluate CN versus AD classification under complete and heterogeneous modality availability (further details in Appendix Section 7.8 and Table 14). Overall, MICViT generalizes effectively to clinically relevant disease classification and remains robust under heterogeneous modality availability. Across all available modality configurations, it achieves the highest accuracy (0.91 ± 0.02) and Macro-F1 score (0.84 ± 0.07), while obtaining the second-highest AUC after mmFormer. On the complete four-modality subset, MICViT achieves the best performance across all metrics, including an AUC of 0.83 ± 0.03 and a Macro-F1 score of 0.71 ± 0.03. Although mmFormer provides a slightly higher AUC in the heterogeneous setting and FlexMoE identifies more AD cases, both show a less favorable balance between the CN and AD classes (Appendix 6). In contrast, MICViT combines strong discrimination with comparatively balanced class-wise predictions. Thus, MICViT appears particularly well suited for clinically realistic AD classification settings, where both complete and heterogeneous MRI acquisitions must be handled without sacrificing balanced class-wise performance.

6 Conclusion

Modeling spatial and cross-modal interactions in 3D brain MRI remains a central challenge. In this work, we introduced MICViT, a novel architecture that explicitly models intra- and cross-modality dependencies via SL, SG, CL, and CG attention, enabling structured interaction learning in both local and global contexts. We showed that MICViT consistently outperforms established 3D baselines, demonstrating that explicit interaction modeling is a key driver of performance. Moreover, the model benefits more strongly from increasing the number of input modalities than competing architectures, indicating a more effective utilization of complementary information. Across multiple neuroimaging datasets and model scales, MICViT exhibits robust and consistent performance, indicating stable behavior across varying data conditions and acquisition settings.

Beyond brain age prediction, we evaluated whether the learned representations transfer to clinically oriented tasks, including disease-onset prediction in UK Biobank and Alzheimer's disease classification in ADNI. These analyses suggest that MICViT captures information relevant beyond chronological age estimation and can support downstream clinical prediction settings.

Despite this strong empirical performance, several limitations remain. First, although we include disease-onset and AD classification analyses, broader validation across additional clinical tasks such as segmentation, diagnosis across independent hospitals, treatment-response prediction, and prospective risk modeling remains future work. Second, MICViT assumes well-aligned multimodal inputs and does not explicitly handle missing, corrupted, or imbalanced modalities. Third, the use of multiple attention mechanisms (SL, SG, CL, CG) increases architectural complexity and computational cost, particularly for high-resolution 3D data. Finally, experiments are conducted in a dataset-specific setting. Robustness under stronger domain shifts without explicit cross-dataset adaptation remains an open question. Overall, our results emphasize that explicitly modeling the structure of intra- and cross-modal interactions is critical for effective multimodal learning in 3D medical imaging.

Use of generative AI We used generative AI tools, including ChatGPT (OpenAI) and Claude (Anthropic), to drafting and revising parts of the manuscript, including improving English phrasing. We also used these tools to assist with translation and to edit plotting or analysis code. The authors reviewed the AI-assisted content included in the manuscript and checked the reported results against the corresponding analyses. The authors take responsibility for the final content of this work.

REFERENCES

John Absher, Sarah Goncher, Roger Newman-Norlund, Nicholas Perkins, Grigori Yourganov, Jan Vargas, Sanjeev Sivakumar, Naveen Parti, Shannon Sternberg, Alex Teghipco, Makayla Gibson, Sarah Wilson, Leonardo Bonilha, and Chris Rorden. The stroke outcome optimization project: Acute ischemic strokes from a comprehensive stroke center. *Scientific Data*, 11(1):839, August 2024. ISSN 2052-4463. doi: 10.1038/s41597-024-03667-5. URL https://www.nature.com/articles/s41597-024-03667-5.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal Machine Learning: A Survey and Taxonomy, August 2017. URL http://arxiv.org/abs/1705.09406. arXiv:1705.09406 [cs].

Vishnu M Bashyam, Guray Erus, Jimit Doshi, Mohamad Habes, Ilya M Nasrallah, Monica Truelove-Hill, Dhivya Srinivasan, Liz Mamourian, Raymond Pomponio, Yong Fan, Lenore J Launer, Colin L Masters, Paul Maruff, Chuanjun Zhuo, Henry Völzke, Sterling C Johnson, Jurgen Fripp, Nikolaos Koutsouleris, Theodore D Satterthwaite, Daniel Wolf, Raquel E Gur, Ruben C Gur, John Morris, Marilyn S Albert, Hans J Grabe, Susan Resnick, R Nick Bryan, David A Wolk, Haochang Shou, and Christos Davatzikos. MRI signatures of brain age and disease over the lifespan based on a deep brain network and 14 468 individuals worldwide. *Brain*, 143(7): 2312–2324, July 2020. ISSN 0006-8950, 1460-2156. doi: 10.1093/brain/awaa160. URL https://academic.oup.com/brain/article/143/7/2312/5863667.

Dani Beck, Ann-Marie G. de Lange, Ivan I. Maximov, Geneviève Richard, Ole A. Andreassen, Jan E. Nordvik, and Lars T. Westlye. White matter microstructure across the adult lifespan: A mixed longitudinal and cross-sectional study using advanced diffusion models and brain-age prediction. *NeuroImage*, 224:117441, January 2021. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2020.117441. URL https://www.sciencedirect.com/science/article/pii/S1053811920309265.

Mustapha Bouhrara, Luis E. Cortina, Nikkita Khattar, Abinand C. Rejimon, Samuel Ajamu, Defne S. Cezayirli, and Richard G. Spencer. Maturation and degeneration of the human brainstem across the adult lifespan. *Aging*, 13(11):14862–14891, June 2021. ISSN 1945-4589. doi: 10.18632/aging.203183. URL https://www.aging-us.com/lookup/doi/10.18632/aging.203183.

Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, and Yuyin Zhou. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation, February 2021. URL http://arxiv.org/abs/2102.04306. arXiv:2102.04306 [cs].

Xiang Chen, Zhiyuan Yuan, Jie Zhang, and Xiao-Yong Zhang. Multimodal quantitative MRI reveals age-related biophysical alterations in the human brain across the adult lifespan. *NeuroImage*, 327:121742, February 2026a. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2026.121742. URL https://www.sciencedirect.com/science/article/pii/S1053811926000601.

Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3D Object Detection Network for Autonomous Driving. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6526–6534, Honolulu, HI, July 2017. IEEE. ISBN 978-1-5386-0457-1. doi: 10.1109/CVPR.2017.691. URL http://ieeexplore.ieee.org/document/8100174/.

Xinjie Chen, Mario Ocampo-Pineda, Po-Jui Lu, Michelle G Jansen, Kwok-Shing Chan, Marcel Zwiers, Joukje M Oosterman, David G Norris, Andre F Marquand, Lester Melie-Garcia, Cristina

Granziera, and José P Marques. Mapping heterogeneous region- and tissue-specific brain ageing patterns using quantitative MRI. *Brain Communications*, 8(1):fcag010, February 2026b. ISSN 2632-1297. doi: 10.1093/braincomms/fcag010. URL https://doi.org/10.1093/braincomms/fcag010.

Simon R. Cox, Stuart J. Ritchie, Elliot M. Tucker-Drob, David C. Liewald, Saskia P. Hagenaars, Gail Davies, Joanna M. Wardlaw, Catharine R. Gale, Mark E. Bastin, and Ian J. Deary. Ageing and brain white matter structure in 3,513 UK Biobank participants. *Nature Communications*, 7(1):13629, December 2016. ISSN 2041-1723. doi: 10.1038/ncomms13629. URL https://www.nature.com/articles/ncomms13629.

Rémi Cuingnet, Emilie Gerardin, Jérôme Tessieras, Guillaume Auzias, Stéphane Lehéricy, Marie-Odile Habert, Marie Chupin, Habib Benali, and Olivier Colliot. Automatic classification of patients with Alzheimer's disease from structural MRI: A comparison of ten methods using the ADNI database. *NeuroImage*, 56(2):766–781, May 2011. ISSN 10538119. doi: 10.1016/j.neuroimage.2010.06.013. URL https://linkinghub.elsevier.com/retrieve/pii/S1053811910008578.

Charles DeCarli, Joseph Massaro, Danielle Harvey, John Hald, Mats Tullberg, Rhoda Au, Alexa Beiser, Ralph D'Agostino, and Philip A. Wolf. Measures of brain morphology and infarction in the framingham heart study: establishing what is normal. *Neurobiology of Aging*, 26(4):491–510, April 2005. ISSN 0197-4580. doi: 10.1016/j.neurobiolaging.2004.05.004. URL https://www.sciencedirect.com/science/article/pii/S0197458004001988.

Aaron Defazio, Xingyu Alice Yang, Harsh Mehta, Konstantin Mishchenko, Ahmed Khaled, and Ashok Cutkosky. The Road Less Scheduled, October 2024. URL http://arxiv.org/abs/2405.15682. arXiv:2405.15682 [cs].

Nicola K. Dinsdale, Emma Bluemke, Stephen M. Smith, Zobair Arya, Diego Vidaurre, Mark Jenkinson, and Ana I. L. Namburete. Learning patterns of the ageing brain in MRI using deep convolutional networks. *NeuroImage*, 224:117401, January 2021. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2020.117401. URL https://www.sciencedirect.com/science/article/pii/S1053811920308867.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, June 2021. URL http://arxiv.org/abs/2010.11929. arXiv:2010.11929 [cs].

Lara Dular and Žiga Špiclin. BASE: Brain Age Standardized Evaluation. *NeuroImage*, 285:120469, January 2024. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2023.120469. URL https://www.sciencedirect.com/science/article/pii/S1053811923006183.

Farnaz Farokhian, Chunlan Yang, Iman Beheshti, Hiroshi Matsuda, and Shuicai Wu. Age-Related Gray and White Matter Changes in Normal Adult Brains. *Aging and Disease*, 8(6):899–909, December 2017. ISSN 2152-5250. doi: 10.14336/AD.2017.0502. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC5758357/.

Robert J. Gillies, Paul E. Kinahan, and Hedvig Hricak. Radiomics: Images Are More than Pictures, They Are Data. *Radiology*, 278(2):563–577, February 2016. ISSN 0033-8419. doi: 10.1148/radiol.2015151169. URL https://pubs.rsna.org/doi/10.1148/radiol.2015151169.

Umang Gupta, Pradeep K. Lam, Greg Ver Steeg, and Paul M. Thompson. Improved Brain Age Estimation with Slice-based Set Networks, February 2021. URL http://arxiv.org/abs/2102.04438. arXiv:2102.04438 [eess].

E. Mark Haacke, Robert W. Brown, Michael R. Thompson, and Ramesh Venkatesan (eds.). *Magnetic resonance imaging: physical principles and sequence design*. Wiley-Liss, New York, NY, 1999. ISBN 978-0-471-35128-3.

Mohamad Habes, Guray Erus, Jon B. Toledo, Tianhao Zhang, Nick Bryan, Lenore J. Launer, Yves Rosseel, Deborah Janowitz, Jimit Doshi, Sandra Van der Auwera, Bettina von Sarnowski, Katrin Hegenscheid, Norbert Hosten, Georg Homuth, Henry Völzke, Ulf Schminke, Wolfgang Hoffmann, Hans J. Grabe, and Christos Davatzikos. White matter hyperintensities and imaging patterns of brain ageing in the general population. *Brain*, 139(4):1164–1179, April 2016. ISSN 0006-8950. doi: 10.1093/brain/aww008. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC5006227/.

F. E. Harrell, R. M. Califf, D. B. Pryor, K. L. Lee, and R. A. Rosati. Evaluating the yield of medical tests. *JAMA*, 247(18):2543–2546, May 1982. ISSN 0098-7484.

Ali Hatamizadeh, Hongxu Yin, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. Global Context Vision Transformers, June 2023. URL http://arxiv.org/abs/2206.09959. arXiv:2206.09959 [cs].

Sheng He, Diana Pereira, Juan David Perez, Randy L. Gollub, Shawn N. Murphy, Sanjay Prabhu, Rudolph Pienaar, Richard L. Robertson, P. Ellen Grant, and Yangming Ou. Multi-channel attention-fusion neural network for brain age estimation: Accuracy, generality, and interpretation with 16,705 healthy MRIs across lifespan. *Medical Image Analysis*, 72:102091, August 2021. ISSN 1361-8415. doi: 10.1016/j.media.2021.102091. URL https://www.sciencedirect.com/science/article/pii/S1361841521001377.

Sheng He, P. Ellen Grant, and Yangming Ou. Global-Local Transformer for Brain Age Estimation. *IEEE Transactions on Medical Imaging*, 41(1):213–224, January 2022. ISSN 0278-0062, 1558-254X. doi: 10.1109/TMI.2021.3108910. URL https://ieeexplore.ieee.org/document/9525077/.

Shih-Cheng Huang, Anuj Pareek, Saeed Seyyedi, Imon Banerjee, and Matthew P. Lungren. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *npj Digital Medicine*, 3(1):136, October 2020. ISSN 2398-6352. doi: 10.1038/s41746-020-00341-z. URL https://www.nature.com/articles/s41746-020-00341-z.

Robert J. Jirsaraie, Aaron J. Gorelik, Martins M. Gatavins, Denis A. Engemann, Ryan Bogdan, Deanna M. Barch, and Aristeidis Sotiras. A systematic review of multimodal brain age studies: Uncovering a divergence between model accuracy and utility. *Patterns*, 4(4):100712, April 2023. ISSN 2666-3899. doi: 10.1016/j.patter.2023.100712.

Johan Jönemo, Muhammad Usman Akbar, Robin Kämpe, J. Paul Hamilton, and Anders Eklund. Efficient Brain Age Prediction from 3D MRI Volumes Using 2D Projections. *Brain Sciences*, 13 (9):1329, September 2023. ISSN 2076-3425. doi: 10.3390/brainsci13091329. URL https://www.mdpi.com/2076-3425/13/9/1329.

Osman Semih Kayhan and Jan C. van Gemert. On Translation Invariance in CNNs: Convolutional Layers can Exploit Absolute Spatial Location, May 2020. URL http://arxiv.org/abs/2003.07064. arXiv:2003.07064 [cs].

Peter Yongho Kim, Junbeom Kwon, Sunghwan Joo, Sangyoon Bae, Donggyu Lee, Yoonho Jung, Shinjae Yoo, Jiook Cha, and Taesup Moon. SwiFT: Swin 4D fMRI Transformer, October 2023. URL http://arxiv.org/abs/2307.05916. arXiv:2307.05916 [cs].

Jeyeon Lee, Brian J. Burkett, Hoon-Ki Min, Matthew L. Senjem, Emily S. Lundt, Hugo Botha, Jonathan Graff-Radford, Leland R. Barnard, Jeffrey L. Gunter, Christopher G. Schwarz, Kejal Kantarci, David S. Knopman, Bradley F. Boeve, Val J. Lowe, Ronald C. Petersen, Clifford R. Jack, and David T. Jones. Deep learning-based brain age prediction in normal aging and dementia. *Nature aging*, 2(5):412–424, May 2022. ISSN 2662-8465. doi: 10.1038/s43587-022-00219-7. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC10154042/.

Esten H. Leonardsen, Han Peng, Tobias Kaufmann, Ingrid Agartz, Ole A. Andreassen, Elisabeth Gulowsen Celius, Thomas Espeseth, Hanne F. Harbo, Einar A. Høgestøl, Ann-Marie de Lange, Andre F. Marquand, Didac Vidal-Piñeiro, James M. Roe, Geir Selbæk, Øystein Sørensen, Stephen M.

Smith, Lars T. Westlye, Thomas Wolfers, and Yunpeng Wang. Deep neural networks learn general and clinically relevant representations of the ageing brain. *NeuroImage*, 256:119210, August 2022. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2022.119210.

Jun Li, Junyu Chen, Yucheng Tang, Ce Wang, Bennett A. Landman, and S. Kevin Zhou. Transforming medical imaging with Transformers? A comparative review of key properties, current progresses, and future perspectives, June 2022. URL https://arxiv.org/abs/2206.01136v3.

Yihao Li, Mostafa El Habib Daho, Pierre-Henri Conze, Rachid Zeghlache, Hugo Le Boité, Ramin Tadayoni, Béatrice Cochener, Mathieu Lamard, and Gwenolé Quellec. A review of deep learning-based information fusion techniques for multimodal medical image classification, April 2024. URL http://arxiv.org/abs/2404.15022. arXiv:2404.15022 [cs].

Yu Liu, Yu Shi, Fuhao Mu, Juan Cheng, Chang Li, and Xun Chen. Multimodal MRI Volumetric Data Fusion With Convolutional Neural Networks. *IEEE Transactions on Instrumentation and Measurement*, 71:1–15, 2022. ISSN 1557-9662. doi: 10.1109/TIM.2022.3184360. URL https://ieeexplore.ieee.org/document/9800914.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows, August 2021. URL http://arxiv.org/abs/2103.14030. arXiv:2103.14030 [cs].

Na Luo, Weiyang Shi, Zhengyi Yang, Ming Song, and Tianzi Jiang. Multimodal Fusion of Brain Imaging Data: Methods and Applications. *Machine Intelligence Research*, 21(1):136–152, February 2024. ISSN 2731-5398. doi: 10.1007/s11633-023-1442-8. URL https://doi.org/10.1007/s11633-023-1442-8.

Moona Mazher, Geoff J. M. Parker, and Daniel C. Alexander. Towards Generalisable Foundation Models for Brain MRI, December 2025. URL http://arxiv.org/abs/2510.23415. arXiv:2510.23415 [cs].

Ibomoiye Domor Mienye, Theo G. Swart, George Obaido, Matt Jordan, and Philip Ilono. Deep Convolutional Neural Networks in Medical Image Analysis: A Review. *Information*, 16(3):195, March 2025. ISSN 2078-2489. doi: 10.3390/info16030195. URL https://www.mdpi.com/2078-2489/16/3/195.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. Multimodal deep learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pp. 689–696, Madison, WI, USA, June 2011. Omnipress. ISBN 978-1-4503-0619-5.

Han Peng, Weikang Gong, Christian F. Beckmann, Andrea Vedaldi, and Stephen M. Smith. Accurate brain age prediction with lightweight deep neural networks. *Medical Image Analysis*, 68:101871, February 2021. ISSN 13618415. doi: 10.1016/j.media.2020.101871. URL https://linkinghub.elsevier.com/retrieve/pii/S1361841520302358.

R C. Petersen, P S. Aisen, L A. Beckett, M C. Donohue, A C. Gamst, D J. Harvey, C R. Jack, W J. Jagust, L M. Shaw, A W. Toga, J Q. Trojanowski, and M W. Weiner. Alzheimer's Disease Neuroimaging Initiative (ADNI). *Neurology*, 74(3):201–209, January 2010. ISSN 0028-3878. doi: 10.1212/WNL.0b013e3181cb3e25. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC2809036/.

Naftali Raz, Ulman Lindenberger, Karen M. Rodrigue, Kristen M. Kennedy, Denise Head, Adrienne Williamson, Cheryl Dahle, Denis Gerstorf, and James D. Acker. Regional Brain Changes in Aging Healthy Adults: General Trends, Individual Differences and Modifiers. *Cerebral Cortex*, 15(11):1676–1689, November 2005. ISSN 1047-3211. doi: 10.1093/cercor/bhi044. URL https://doi.org/10.1093/cercor/bhi044.

Tillmann Rheude, Roland Eils, and Benjamin Wild. Fusion or Confusion? Multimodal Complexity Is Not All You Need, 2025. URL https://arxiv.org/abs/2512.22991. Version Number: 2.

Marc-Andre Schulz, Danilo Bzdok, Stefan Haufe, John-Dylan Haynes, and Kerstin Ritter. Performance reserves in brain-imaging-based phenotype prediction. *Cell Reports*, 43(1):113597, January 2024. ISSN 22111247. doi: 10.1016/j.celrep.2023.113597. URL https://linkinghub.elsevier.com/retrieve/pii/S2211124723016091.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *International Journal of Computer Vision*, 128(2):336–359, February 2020. ISSN 0920-5691, 1573-1405. doi: 10.1007/s11263-019-01228-7. URL http://arxiv.org/abs/1610.02391. arXiv:1610.02391 [cs].

Meredith A Shafto, Lorraine K Tyler, Marie Dixon, Jason R Taylor, James B Rowe, Rhodri Cusack, Andrew J Calder, William D Marslen-Wilson, John Duncan, Tim Dalgleish, Richard N Henson, Carol Brayne, and Fiona E Matthews. The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) study protocol: a cross-sectional, lifespan, multidisciplinary examination of healthy cognitive ageing. *BMC Neurology*, 14:204, October 2014. ISSN 1471-2377. doi: 10.1186/s12883-014-0204-1. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC4219118/.

Jakob Steinfeldt, Benjamin Wild, Thore Buergel, Maik Pietzner, Julius Upmeier zu Belzen, Andre Vauvelle, Stefan Hegselmann, Spiros Denaxas, Harry Hemingway, Claudia Langenberg, Ulf Landmesser, John Deanfield, and Roland Eils. Medical history predicts phenome-wide disease onset and enables the rapid response to emerging health threats. *Nature Communications*, 16(1):585, January 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-55879-x. URL https://www.nature.com/articles/s41467-025-55879-x.

Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, Bette Liu, Paul Matthews, Giok Ong, Jill Pell, Alan Silman, Alan Young, Tim Sprosen, Tim Peakman, and Rory Collins. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Medicine*, 12(3):e1001779, March 2015. ISSN 1549-1277. doi: 10.1371/journal.pmed.1001779. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC4380465/.

Kai Sun, Siyan Xue, Fuchun Sun, Haoran Sun, Yu Luo, Ling Wang, Siyuan Wang, Na Guo, Lei Liu, Tian Zhao, Xinzhou Wang, Lei Yang, Shuo Jin, Jun Yan, and Jiahong Dong. Medical Multimodal Foundation Models in Clinical Diagnosis and Treatment: Applications, Challenges, and Future Directions, December 2024. URL http://arxiv.org/abs/2412.02621. arXiv:2412.02621 [cs].

Divyanshu Tak, Biniam A. Garomsa, Anna Zapaishchykova, Tafadzwa L. Chaunzwa, Juan Carlos Climent Pardo, Zezhong Ye, John Zielke, Yashwanth Ravipati, Suraj Pai, Sri Vajapeyam, Maryam Mahootiha, Mitchell Parker, Luke R. G. Pike, Ceilidh Smith, Ariana M. Familiar, Kevin X. Liu, Sanjay Prabhu, Omar Arnaout, Pratiti Bandopadhayay, Ali Nabavizadeh, Sabine Mueller, Hugo Jwl Aerts, Raymond Y. Huang, Tina Y. Poussaint, and Benjamin H. Kann. A generalizable foundation model for analysis of human brain MRI. *Nature Neuroscience*, February 2026. ISSN 1097-6256, 1546-1726. doi: 10.1038/s41593-026-02202-6. URL https://www.nature.com/articles/s41593-026-02202-6.

Satoshi Takahashi, Yusuke Sakaguchi, Nobuji Kouno, Ken Takasawa, Kenichi Ishizu, Yu Akagi, Rina Aoyama, Naoki Teraya, Amina Bolatkan, Norio Shinkai, Hidenori Machino, Kazuma Kobayashi, Ken Asada, Masaaki Komatsu, Syuzo Kaneko, Masashi Sugiyama, and Ryuji Hamamoto. Comparison of Vision Transformers and Convolutional Neural Networks in Medical Image Analysis: A Systematic Review. *Journal of Medical Systems*, 48(1):84, September 2024. ISSN 1573-689X. doi: 10.1007/s10916-024-02105-8. URL https://doi.org/10.1007/s10916-024-02105-8.

Jason R. Taylor, Nitin Williams, Rhodri Cusack, Tibor Auer, Meredith A. Shafto, Marie Dixon, Lorraine K. Tyler, null Cam-Can, and Richard N. Henson. The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) data repository: Structural and functional MRI, MEG, and cognitive data from a cross-sectional adult lifespan sample. *NeuroImage*, 144(Pt B):262–269, January 2017. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2015.09.018.

Asher Trockman and J. Zico Kolter. Mimetic Initialization of Self-Attention Layers, May 2023. URL http://arxiv.org/abs/2305.09828. arXiv:2305.09828 [cs].

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need, August 2023. URL http://arxiv.org/abs/1706.03762. arXiv:1706.03762 [cs].

Yu Wang, Ye Teng, Tianci Liu, Yuchun Tang, Wenjia Liang, Wenjun Wang, Zhuoran Li, Qing Xia, Feifei Xu, and Shuwei Liu. Morphological changes in the cerebellum during aging: evidence from convolutional neural networks and shape analysis. *Frontiers in Aging Neuroscience*, 16, April 2024. ISSN 1663-4365. doi: 10.3389/fnagi.2024.1359320. URL https://www.frontiersin.org/journals/aging-neuroscience/articles/10.3389/fnagi.2024.1359320/full.

Joanna M Wardlaw, Eric E Smith, Geert J Biessels, Charlotte Cordonnier, Franz Fazekas, Richard Frayne, Richard I Lindley, John T O'Brien, Frederik Barkhof, Oscar R Benavente, Sandra E Black, Carol Brayne, Monique Breteler, Hugues Chabriat, Charles DeCarli, Frank-Erik de Leeuw, Fergus Doubal, Marco Duering, Nick C Fox, Steven Greenberg, Vladimir Hachinski, Ingo Kilimann, Vincent Mok, Robert van Oostenbrugge, Leonardo Pantoni, Oliver Speck, Blossom C M Stephan, Stefan Teipel, Anand Viswanathan, David Werring, Christopher Chen, Colin Smith, Mark van Buchem, Bo Norrving, Philip B Gorelick, and Martin Dichgans. Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration. *Lancet Neurology*, 12(8):822–838, August 2013. ISSN 1474-4422. doi: 10.1016/S1474-4422(13)70124-8. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC3714437/.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tie-Yan Liu. On Layer Normalization in the Transformer Architecture, June 2020. URL http://arxiv.org/abs/2002.04745. arXiv:2002.04745 [cs].

Sukwon Yun, Inyoung Choi, Jie Peng, Yangfan Wu, Jingxuan Bao, Qiyiwen Zhang, Jiayi Xin, Qi Long, and Tianlong Chen. Flex-MoE: Modeling Arbitrary Modality Combination via the Flexible Mixture-of-Experts, October 2024. URL http://arxiv.org/abs/2410.08245. arXiv:2410.08245 [cs.LG].

Yao Zhang, Nanjun He, Jiawei Yang, Yuexiang Li, Dong Wei, Yawen Huang, Yang Zhang, Zhiqiang He, and Yefeng Zheng. mmFormer: Multimodal Medical Transformer for Incomplete Multimodal Learning of Brain Tumor Segmentation, August 2022. URL http://arxiv.org/abs/2206.02425. arXiv:2206.02425 [eess.IV].

Zhaonian Zhang and Richard Jiang. Triamese-ViT: A 3D-Aware Method for Robust Brain Age Estimation from MRIs, January 2024. URL http://arxiv.org/abs/2401.09475. arXiv:2401.09475 [cs] version: 1.

Haiyan Zhao, Hongjie Cai, and Manhua Liu. Transformer based multi-modal MRI fusion for prediction of post-menstrual age and neonatal brain development analysis. *Medical Image Analysis*, 94:103140, May 2024. ISSN 13618415. doi: 10.1016/j.media.2024.103140. URL https://linkinghub.elsevier.com/retrieve/pii/S1361841524000653.

Juntang Zhuang, Boqing Gong, Liangzhe Yuan, Yin Cui, Hartwig Adam, Nicha Dvornek, Sekhar Tatikonda, James Duncan, and Ting Liu. Surrogate Gap Minimization Improves Sharpness-Aware Training, March 2022. URL http://arxiv.org/abs/2203.08065. arXiv:2203.08065 [cs].

7 APPENDIX

Table 5: Ablation over single-layer attention permutations on UKBB. We evaluate all combinations of separated local (SL), separated global (SG), cross local (CL), and cross global (CG) attention. The full configuration (SL-SG-CL-CG) achieves the best performance (MAE ↓), consistently outperforming all other permutations. This indicates that gains arise from the interaction of complementary mechanisms rather than any single component. Separated attention (SL, SG) strengthens intra-modality representations, while cross attention (CL, CG) enables inter-modality exchange. Orthogonally, local attention (SL, CL) captures fine-grained structure, whereas global attention (SG, CG) models long-range dependencies.

Attention	**MAE ↓ ± SD**	**Parameters**
CL-CL-CL-CL	4.12 ± 0.04	176837
SL-SL-SL-SL	4.05 ± 0.06	176837
CL-SG-CL-SG	3.80 ± 0.62	174739
CL-CG-CL-CG	3.67 ± 0.61	174725
SL-SG-SL-SG	3.48 ± 0.22	174725
CL-CG-CL-SG	3.47 ± 0.34	174746
SL-CG-SL-SG	3.47 ± 0.20	174732
SL-SG-CL-SG	3.44 ± 0.10	174725
SL-SG-SL-CG	3.41 ± 0.27	174732
SG-CL-SG-SL	3.39 ± 0.15	174732
SL-CG-SL-CG	3.38 ± 0.16	174739
SG-SL-SG-SL	3.38 ± 0.13	174725
CG-CL-CG-SL	3.37 ± 0.15	174746
CL-SG-SL-SG	3.36 ± 0.36	174725
CL-SG-SL-CG	3.36 ± 0.26	174725
SL-CG-CL-CG	3.36 ± 0.17	174725
CL-CG-SL-SG	3.36 ± 0.36	174725
SG-SG-SG-SG	3.36 ± 0.13	172613
CL-SG-CL-CG	3.35 ± 0.13	174746
CL-CG-SL-CG	3.34 ± 0.26	174725
SL-CG-CL-SG	3.33 ± 0.18	174725
CG-CL-CG-CL	3.31 ± 0.14	174753
CG-SL-SG-CL	3.31 ± 0.17	174725
CG-CL-SG-SL	3.31 ± 0.15	174739
SG-SL-SG-CL	3.29 ± 0.20	174732
SG-CL-SG-CL	3.29 ± 0.10	174739
CG-CL-SG-CL	3.26 ± 0.12	174725
SG-CL-CG-SL	3.26 ± 0.09	174739
SG-SL-CG-SL	3.25 ± 0.04	174732
SG-SL-CG-CL	3.24 ± 0.05	174725
CG-CG-CG-CG	3.24 ± 0.07	172613
CG-SL-CG-CL	3.22 ± 0.04	174746
CG-SL-SG-SL	3.21 ± 0.10	174725
CG-SL-CG-SL	3.21 ± 0.08	174739
SG-CL-CG-CL	3.20 ± 0.09	174746
SL-SG-CL-CG	$\mathbf{3.12} \pm 0.19$	174725

Table 6: Model configurations across modality settings. Parameters are reported in millions (M).

Model	Parameter	T1/FLAIR/SWI/DWI modalities=1	FLAIR+DWI modalities=2	T1+FLAIR+DWI modalities=3	T1+FLAIR+DWI+SWI modalities=4
MICViT (ours)					
	dim	224	224	224	224
	num_heads	14	14	14	14
	attn_types	[SL, SG, SL, SG]	[SL, SG, CL, CG]	[SL, SG, CL, CG]	[SL, SG, CL, CG]
	window_size	[4,4,4]	[4,4,4]	[4,4,4]	[4,4,4]
	Params (M)	2.28	3.55	5.04	6.97
ConvNeXt					
	depths	[2,3,4,2]	[2,4,10,2]	[4,8,10,4]	[8,10,12,6]
	dims	[128,128,128,128]	[128,128,128,128]	[128,128,128,128]	[128,128,128,128]
	Params (M)	2.34	3.58	5.00	6.77
GCViT					
	dim	32	32	32	32
	num_heads	[2,8,8,4]	[2,8,8,4]	[2,8,8,4]	[2,8,8,4]
	depth	[1,1,1,1]	[2,4,4,2]	[2,4,5,4]	[4,10,14,4]
	window_size	[[4,4,4],[4,4,4], [8,8,8],[4,4,4]]	[[4,4,4],[4,4,4], [8,8,8],[4,4,4]]	[[4,4,4],[4,4,4], [8,8,8],[4,4,4]]	[[4,4,4],[4,4,4], [8,8,8],[4,4,4]]
	Params (M)	2.28	3.56	5.01	6.96

Table 7: Model configurations across different scales.

Model	Parameter	Small	Medium	Base	Large
MICViT (ours)					
	dim	32	64	128	224
	num_heads	2	4	8	14
	attn_types	[SL, SG, CL, CG]	[SL, SG, CL, CG]	[SL, SG, CL, CG]	[SL, SG, CL, CG]
	window_size	[4,4,4]	[4,4,4]	[4,4,4]	[4,4,4]
	Params (M)	0.17	0.62	2.34	6.97
ConvNeXt					
	depths	[1,2,3,2]	[2,2,3,2]	[2,2,4,3]	[8,10,12,6]
	dims	[32,32,32,32]	[64,64,64,64]	[128,128,128,128]	[128,128,128,128]
	Params (M)	0.19	0.61	2.40	6.77
Swin Transformer					
	embed_dim	32	64	128	224
	depths	[3,3]	[2,3]	[2,3]	[1,3]
	num_heads	[2,2]	[4,4]	[4,4]	[4,4]
	window_size	[4,4,4]	[4,4,4]	[4,4,4]	[4,4,4]
	Params (M)	0.17	0.62	2.45	6.79
ViT					
	depth	8	12	11	34
	dim	32	64	128	128
	heads	4	4	8	8
	mlp_dim	256	256	512	512
	Params (M)	0.18	0.74	2.71	6.87
GCViT					
	depth	[1,2,2,1]	[2,2,6,2]	[2,6,6,2]	[4,10,14,4]
	num_heads	[2,2,2,2]	[2,2,4,2]	[2,4,6,2]	[2,8,8,4]
	dim	8	12	24	32
	window_size	[[4,4,4],[4,4,4], [8,8,8],[4,4,4]]	[[4,4,4],[4,4,4], [8,8,8],[4,4,4]]	[[4,4,4],[4,4,4], [8,8,8],[4,4,4]]	[[4,4,4],[4,4,4], [8,8,8],[4,4,4]]
	Params (M)	0.16	0.62	2.31	6.96
ResNet					
	layers	[8,16,32,64]	[16,32,64,64]	[16,32,64,128]	[16,32,64,128]
	dim	[1,1,1,1]	[2,2,2,2]	[2,4,6,3]	[12,20,32,4]
	strides	[1,2,2,1]	[1,2,2,1]	[1,2,2,1]	[1,2,2,1]
	Params (M)	0.18	0.60	2.46	6.88
ViT (Late Fusion)					
	depth	2	2	2	4
	dim	32	64	128	176
	heads	2	4	8	8
	dim_fusion	32	64	128	176
	heads_fusion	2	4	8	16
	Params (M)	0.19	0.54	2.12	6.98
ResNet (Late Fusion)					
	hidden_dims	[8,12,14,16]	[8,16,20,32]	[8,16,32,64]	[16,32,64,88]
	resnet_layers	[2,2,2,2]	[2,3,4,2]	[2,4,8,2]	[4,5,8,2]
	strides	[1,2,2,1]	[1,2,2,1]	[1,2,2,1]	[1,2,2,1]
	dim_fusion	32	64	128	176
	heads_fusion	2	4	8	16
	Params (M)	0.18	0.60	2.28	7.00
FlexMoE					
	number of Experts	16	16	16	16
	SMoE-layer	1	1	1	1
	Top-k	4	4	4	4
	Attention Heads	2	4	8	8
	Hidden Dimension	32	64	128	256
	Encoder Base Channels	8	16	16	16
mmFormer					
Note: mmFormer is adapted by removing the segmentation decoder and using an image-level prediction head.					
	basic_dims	1	2	5	8
	transformer_dim	32	64	96	128
	num_heads	4	8	8	8
	mlp_dim	64	128	192	256

Table 8: Scaling comparison on UKBB. Mean Absolute Error (MAE) in years (±SD). Late-fusion variants are marked with ‡. Best results are shown in **bold** and second-best results are underlined.

Model	Config	Params (M)	Training time (min/epoch)	MAE ↓ ± SD
ConvNeXt	Small	0.19	21.57	3.526 ± 0.082
	Medium	0.61	60.48	3.484 ± 0.037
	Base	2.40	67.07	3.464 ± 0.040
	Large	6.77	67.67	3.447 ± 0.039
ConvNeXt ‡	Small	0.19	110.80	3.601 ± 0.047
	Medium	0.73	114.31	3.509 ± 0.047
	Base	2.42	124.26	3.450 ± 0.044
	Large	6.88	129.74	3.436 ± 0.062
ResNet	Small	0.18	26.64	4.613 ± 0.687
	Medium	0.60	38.35	4.060 ± 0.068
	Base	2.46	38.43	3.935 ± 0.068
	Large	6.88	34.50	3.711 ± 0.168
ResNet ‡	Small	0.18	22.52	5.176 ± 2.364
	Medium	0.60	28.05	3.605 ± 0.377
	Base	2.28	34.26	3.570 ± 0.105
	Large	7.00	42.07	3.443 ± 0.123
Swin-T	Small	0.17	55.18	4.005 ± 0.203
	Medium	0.62	57.41	3.742 ± 0.159
	Base	2.45	45.12	3.665 ± 0.137
	Large	6.79	64.76	3.536 ± 0.076
GCViT	Small	0.16	43.81	4.094 ± 0.617
	Medium	0.62	58.54	3.453 ± 0.101
	Base	2.31	68.20	3.410 ± 0.034
	Large	6.96	89.45	3.342 ± 0.065
ViT	Small	0.18	48.92	4.511 ± 0.781
	Medium	0.74	90.11	4.097 ± 0.183
	Base	2.71	91.53	4.006 ± 0.205
	Large	6.87	96.82	3.832 ± 0.146
ViT ‡	Small	0.19	64.56	4.878 ± 1.158
	Medium	0.54	52.05	4.376 ± 0.487
	Base	2.12	53.02	3.976 ± 0.227
	Large	6.98	84.56	3.677 ± 0.066
mmFormer	Small	0.15	24.25	4.032 ± 0.032
	Medium	0.56	41.43	3.996 ± 0.062
	Base	2.74	52.76	3.253 ± 0.060
	Large	6.66	70.47	3.179 ± 0.084
Flex-MoE	Small	0.17	11.99	3.398 ± 0.198
	Medium	0.60	12.25	3.180 ± 0.080
	Base	2.28	22.16	3.161 ± 0.191
	Large	6.81	23.56	3.086 ± 0.129
MICViT	Small	0.17	38.84	3.119 ± 0.191
	Medium	0.62	60.37	3.000 ± 0.076
	Base	2.34	69.54	$\underline{2.892 \pm 0.035}$
	Large	6.97	86.33	$\mathbf{2.816 \pm 0.062}$

Table 9: Hyperparameter search space for ConvNeXt across model scales.

Section	Parameter	Small	Medium	Base	Large
ConvNeXt					
	depths	[1,2,3,2]	[2,2,3,2]	[2,2,14,2]	[6,8,10,12]
	dims	[32,32,32,32]	[64,64,64,64]	[96,96,96,96]	[128,128,128,128]
	droppath	[0.0, 0.1, 0.2, 0.3]			
Optimizer					
	learning_rate	min: 1e−5, max: 5e−3 log-uniform			
	weight_decay	min: 1e−3, max: 3e−1 log-uniform			

Table 10: External evaluation on the SOOP cohort. Mean Absolute Error (MAE) is reported in years as mean ± standard deviation across five folds. ΔMAE denotes the absolute MAE reduction achieved by MICViT relative to each reference model. Late-fusion variants are marked with ‡. Statistically significant results ($p < 0.05$) are shown in **bold**. Statistical significance was assessed using two-sided paired t-tests across the five folds.

Model	MAE ↓ Mean ± SD	ΔMAE ↑	p-value
MICViT	$\mathbf{6.671 \pm 0.102}$	–	–
mmFormer	6.717 ± 0.144	0.046	0.201
GCViT	6.757 ± 0.135	0.086	0.123
ConvNeXt ‡	6.774 ± 0.125	0.103	0.106
ViT ‡	7.294 ± 0.493	0.623	**0.019**
ConvNeXt	7.326 ± 0.548	0.655	**0.027**
Flex-MoE	7.380 ± 0.697	0.709	**0.038**
ResNet ‡	7.957 ± 0.931	1.287	**0.018**
ViT	8.807 ± 1.785	2.136	**0.026**
ResNet	8.833 ± 1.485	2.162	**0.018**
Swin-T	9.127 ± 0.233	2.456	$< \mathbf{0.001}$

Table 11: External evaluation on the CAM-Can cohort. Mean Absolute Error (MAE) is reported in years as mean ± standard deviation across five folds. ΔMAE denotes the absolute MAE reduction achieved by MICViT relative to each reference model. Late-fusion variants are marked with ‡. Statistically significant results ($p < 0.05$) are shown in **bold**.

Model	MAE ↓ Mean ± SD	ΔMAE ↑	p-value
MICViT	$\mathbf{6.428 \pm 0.690}$	–	–
mmFormer	6.431 ± 0.600	0.004	0.495
Flex-MoE	6.525 ± 0.926	0.097	0.388
ConvNeXt ‡	6.971 ± 0.406	0.543	**0.042**
GCViT	7.849 ± 0.778	1.421	**0.001**
ConvNeXt	8.221 ± 0.723	1.794	**0.001**
ResNet ‡	8.574 ± 1.167	2.147	$< \mathbf{0.001}$
ResNet	8.718 ± 1.286	2.291	**0.007**
ViT	8.939 ± 1.046	2.511	**0.002**
ViT ‡	9.123 ± 0.819	2.695	$< \mathbf{0.001}$
Swin-T	10.575 ± 0.924	4.148	$< \mathbf{0.001}$

Table 12: External evaluation on the ADNI cohort. Mean Absolute Error (MAE) is reported in years as mean ± standard deviation across five folds. ΔMAE denotes the absolute MAE reduction achieved by MICViT relative to each reference model. Late-fusion variants are marked with ‡. Statistically significant results ($p < 0.05$) are shown in **bold**.

Model	MAE ↓ Mean ± SD	ΔMAE ↑	p-value
MICViT	$\mathbf{4.165 \pm 0.373}$	–	–
mmFormer	4.194 ± 0.334	0.029	0.328
ConvNeXt ‡	4.271 ± 0.198	0.106	0.293
ConvNeXt	4.334 ± 0.294	0.169	0.232
GCViT	4.454 ± 0.195	0.289	0.057
ViT	4.652 ± 0.206	0.487	$\mathbf{0.042}$
Flex-MoE	4.703 ± 0.370	0.538	$\mathbf{0.048}$
Swin-T	4.788 ± 0.492	0.623	$\mathbf{0.018}$
ViT ‡	5.004 ± 0.519	0.839	$\mathbf{0.024}$
ResNet ‡	5.289 ± 0.408	1.125	$\mathbf{0.005}$
ResNet	5.311 ± 0.444	1.146	$\mathbf{< 0.001}$

Table 13: Ablation of MICViT attention modules. MAE is reported in years. ΔMAE denotes the increase relative to the full model. Significant p-values ($p < 0.05$) are shown in **bold**.

Removed modules	MAE ↓	ΔMAE ↑	p-value
None (full MICViT)	3.119	–	–
w/o SL	3.136	+0.016	4.37×10^{-1}
w/o CL	3.300	+0.180	1.31×10^{-1}
w/o SG	3.316	+0.197	2.51×10^{-1}
w/o CG	3.333	+0.214	$\mathbf{2.91 \times 10^{-2}}$
w/o SL–SG	3.321	+0.201	2.15×10^{-1}
w/o SL–CG	3.392	+0.272	1.44×10^{-1}
w/o SL–CL	3.440	+0.321	$\mathbf{4.76 \times 10^{-2}}$
w/o CL–CG	3.453	+0.334	$\mathbf{2.91 \times 10^{-2}}$
w/o SG–CL	3.536	+0.417	7.79×10^{-2}
w/o SG–CG	4.057	+0.938	$\mathbf{5.94 \times 10^{-4}}$
w/o SL–SG–CL	3.414	+0.294	1.36×10^{-1}
w/o SL–SG–CG	4.275	+1.156	$\mathbf{1.98 \times 10^{-3}}$
w/o SG–CL–CG	4.350	+1.231	$\mathbf{3.11 \times 10^{-4}}$
w/o SL–CL–CG	4.433	+1.314	$\mathbf{1.48 \times 10^{-2}}$

Table 14: ADNI CN versus AD classification under different modality-availability settings. Results are reported as mean ± standard deviation across five cross-validation folds. **Bold indicates the best result**, while underlining indicates the second-best result within each setting.

Model	Accuracy ↑	AUC ↑	Macro-F1 ↑
All available modality configurations			
ConvNeXt	0.856 ± 0.042	0.894 ± 0.046	0.748 ± 0.077
FlexMoE	0.869 ± 0.064	0.892 ± 0.032	$\underline{0.810 \pm 0.047}$
GCViT	0.862 ± 0.042	0.853 ± 0.035	0.775 ± 0.050
MICViT	$\mathbf{0.914 \pm 0.017}$	$\underline{0.913 \pm 0.038}$	$\mathbf{0.836 \pm 0.065}$
mmFormer	$\underline{0.872 \pm 0.042}$	$\mathbf{0.930 \pm 0.013}$	0.739 ± 0.086
Complete four-modality subset			
ConvNeXt	$\underline{0.819 \pm 0.044}$	0.821 ± 0.041	$\underline{0.691 \pm 0.030}$
FlexMoE	0.809 ± 0.050	0.743 ± 0.073	0.568 ± 0.101
GCViT	0.816 ± 0.049	0.761 ± 0.030	0.687 ± 0.017
MICViT	$\mathbf{0.823 \pm 0.027}$	$\mathbf{0.833 \pm 0.026}$	$\mathbf{0.710 \pm 0.031}$
mmFormer	0.795 ± 0.078	$\underline{0.831 \pm 0.021}$	0.647 ± 0.058

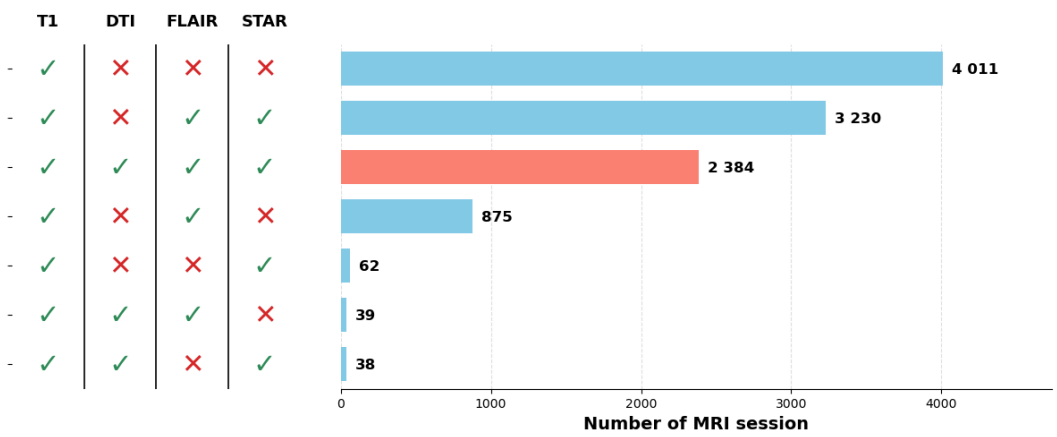

Figure 5: Distribution of available MRI modality combinations in ADNI. Bars indicate the number of MRI sessions for each combination of T1, Diffusion Tensor Imaging (DTI), FLAIR, and STAR; check marks and crosses denote available and missing modalities, respectively. While T1 is available in all shown configurations, only 2,384 sessions contain the complete four-modality set.

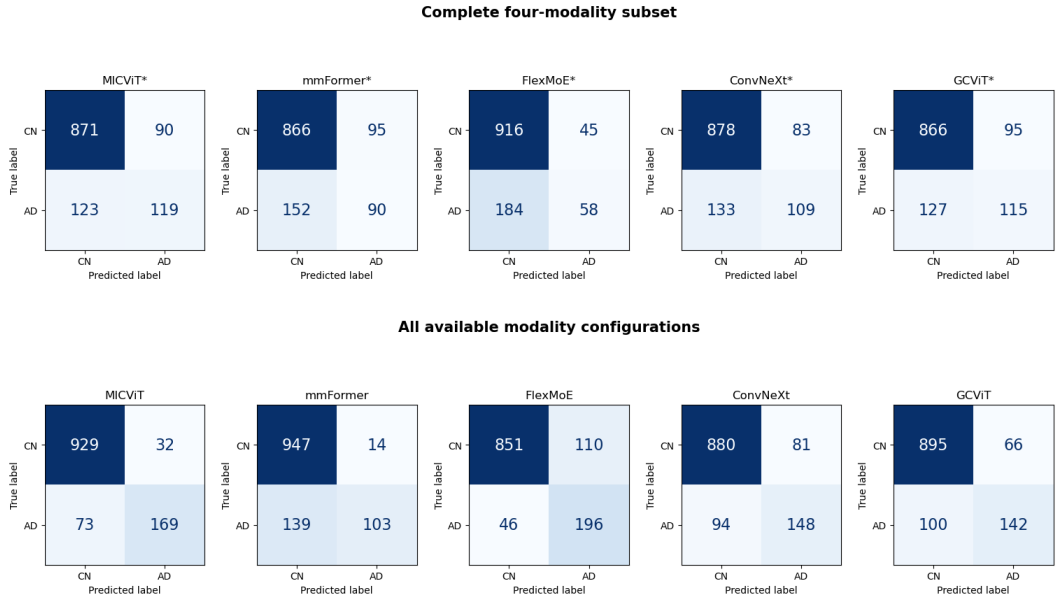

Figure 6: Aggregated confusion matrices for binary classification of cognitively normal (CN) participants and patients with Alzheimer's disease (AD) on ADNI across five cross-validation folds. The top row shows results on the complete four-modality subset, whereas the bottom row includes all available modality configurations, including incomplete cases. Rows denote true labels and columns predicted labels. Asterisks indicate models evaluated on the complete four-modality subset.

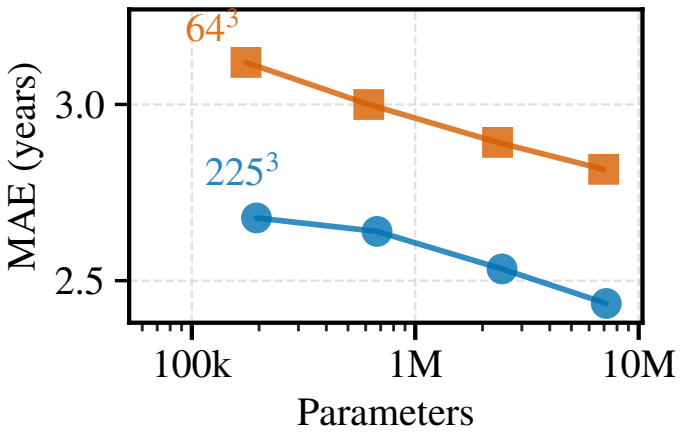

Figure 7: Scaling with model capacity and input resolution. Full-resolution MRI (225^3) reduces MAE across model sizes relative to downsampled inputs (64^3).

Table 15: Evaluation on the UKBB cohort using full-resolution MRI inputs (225^3 voxels). Mean Absolute Error (MAE) is reported in years as mean $\pm$ standard deviation across five folds. ΔMAE denotes the absolute MAE reduction achieved by MICViT relative to each reference model. Late-fusion variants are marked with ‡. Statistically significant results ($p < 0.05$) are shown in **bold**.

Model	**MAE** $\downarrow$ **Mean** $\pm$ **SD**	Δ**MAE** $\uparrow$	p**-value**
MICViT	$\mathbf{2.404 \pm 0.016}$	–	–
GCViT	2.479 ± 0.050	0.074	$\mathbf{0.009}$
mmFormer	2.525 ± 0.124	0.120	$\mathbf{0.044}$
ViT ‡	2.624 ± 0.104	0.220	$\mathbf{0.003}$
ConvNeXt	2.690 ± 0.159	0.285	$\mathbf{0.008}$
Flex-MoE	2.772 ± 0.168	0.367	$\mathbf{0.005}$
ConvNeXt ‡	2.778 ± 0.142	0.374	$\mathbf{0.002}$
ViT	2.920 ± 0.090	0.516	$\mathbf{< 0.001}$
ResNet ‡	3.032 ± 0.227	0.627	$\mathbf{0.002}$
Swin-T	3.083 ± 0.209	0.679	$\mathbf{0.001}$
ResNet	3.402 ± 0.471	0.998	$\mathbf{0.005}$

Table 16: Disease-onset prediction performance across diagnoses. Values represent the mean concordance index (C-index) ± standard deviation over five cross-validation folds. The best mean C-index for each diagnosis is highlighted in bold.

Diagnosis	n	MICViT	GCViT	FlexMoE	ConvNeXt	Swin	ConvNeXt-LF	mmFormer	ViT	ResNet	ResNet-LF	ViT-LF
Essential hypertension	3037	**0.590 ± 0.012**	0.533 ± 0.020	0.589 ± 0.012	0.551 ± 0.011	0.509 ± 0.008	0.548 ± 0.011	0.565 ± 0.010	0.565 ± 0.010	0.462 ± 0.011	0.561 ± 0.018	0.563 ± 0.015
Hypertension	1163	**0.737 ± 0.097**	0.516 ± 0.005	0.537 ± 0.010	0.529 ± 0.006	0.514 ± 0.009	0.519 ± 0.009	0.525 ± 0.010	0.525 ± 0.010	0.624 ± 0.050	0.522 ± 0.011	0.523 ± 0.007
Atherosclerosis (ASCVD)	891	0.612 ± 0.021	0.559 ± 0.033	0.622 ± 0.057	0.601 ± 0.029	0.524 ± 0.009	0.598 ± 0.017	0.591 ± 0.025	0.591 ± 0.025	0.322 ± 0.068	**0.630 ± 0.023**	0.625 ± 0.028
Chronic kidney disease	708	**0.656 ± 0.026**	0.547 ± 0.042	0.638 ± 0.051	0.577 ± 0.022	0.501 ± 0.040	0.566 ± 0.025	0.611 ± 0.043	0.611 ± 0.043	0.342 ± 0.093	0.642 ± 0.008	0.621 ± 0.016
Anxiety and anxiety disorders	655	0.511 ± 0.029	0.511 ± 0.019	**0.537 ± 0.061**	0.520 ± 0.019	0.519 ± 0.008	0.508 ± 0.049	0.521 ± 0.029	0.521 ± 0.029	0.419 ± 0.033	0.493 ± 0.014	0.492 ± 0.019
Type 2 diabetes	567	0.603 ± 0.029	0.548 ± 0.031	**0.638 ± 0.066**	0.569 ± 0.025	0.501 ± 0.017	0.554 ± 0.027	0.592 ± 0.034	0.592 ± 0.034	0.334 ± 0.076	0.595 ± 0.034	0.594 ± 0.038
Hypotension	537	**0.602 ± 0.011**	0.519 ± 0.037	0.562 ± 0.099	0.559 ± 0.039	0.493 ± 0.023	0.508 ± 0.036	0.567 ± 0.016	0.567 ± 0.016	0.324 ± 0.070	0.597 ± 0.020	0.581 ± 0.026
Headache	518	**0.538 ± 0.017**	0.511 ± 0.018	0.518 ± 0.047	0.520 ± 0.019	0.501 ± 0.040	0.493 ± 0.021	0.501 ± 0.042	0.501 ± 0.042	0.417 ± 0.029	0.492 ± 0.031	0.498 ± 0.023
Mood (affective) disorders	484	**0.557 ± 0.010**	0.502 ± 0.017	0.537 ± 0.041	0.544 ± 0.027	0.507 ± 0.027	0.485 ± 0.018	0.513 ± 0.033	0.513 ± 0.033	0.505 ± 0.030	0.501 ± 0.005	0.495 ± 0.028
Dizziness and giddiness	372	0.547 ± 0.031	0.504 ± 0.038	0.508 ± 0.046	**0.562 ± 0.052**	0.493 ± 0.038	0.507 ± 0.032	0.540 ± 0.036	0.540 ± 0.036	0.344 ± 0.055	0.537 ± 0.029	0.478 ± 0.021
Abnormality of gait and mobility	363	0.657 ± 0.052	0.546 ± 0.037	**0.670 ± 0.041**	0.608 ± 0.033	0.523 ± 0.040	0.590 ± 0.017	0.599 ± 0.025	0.599 ± 0.025	0.314 ± 0.104	0.644 ± 0.049	0.645 ± 0.038
Major depressive disorder	361	**0.652 ± 0.091**	0.501 ± 0.014	0.524 ± 0.022	0.510 ± 0.014	0.511 ± 0.009	0.503 ± 0.013	0.495 ± 0.030	0.495 ± 0.030	0.593 ± 0.053	0.505 ± 0.027	0.513 ± 0.029
Diabetes mellitus	341	0.614 ± 0.019	0.542 ± 0.047	**0.626 ± 0.031**	0.588 ± 0.039	0.499 ± 0.031	0.573 ± 0.019	0.584 ± 0.025	0.584 ± 0.025	0.412 ± 0.038	0.587 ± 0.040	0.584 ± 0.037
Repeated falls	311	0.715 ± 0.029	0.572 ± 0.061	**0.719 ± 0.037**	0.655 ± 0.024	0.512 ± 0.052	0.645 ± 0.025	0.688 ± 0.010	0.688 ± 0.010	0.379 ± 0.086	0.691 ± 0.047	0.712 ± 0.037
Stroke	307	**0.666 ± 0.046**	0.562 ± 0.026	0.649 ± 0.041	0.600 ± 0.039	0.504 ± 0.039	0.602 ± 0.045	0.632 ± 0.037	0.632 ± 0.037	0.329 ± 0.043	0.649 ± 0.024	0.646 ± 0.031
Mononeuropathies	263	**0.605 ± 0.029**	0.486 ± 0.033	0.511 ± 0.104	0.567 ± 0.033	0.520 ± 0.034	0.500 ± 0.030	0.506 ± 0.033	0.506 ± 0.033	0.376 ± 0.039	0.492 ± 0.027	0.551 ± 0.057
Nerve root and plexus disorders	246	0.529 ± 0.056	0.489 ± 0.055	0.491 ± 0.054	**0.539 ± 0.031**	0.451 ± 0.052	0.514 ± 0.068	0.520 ± 0.089	0.520 ± 0.089	0.432 ± 0.104	0.505 ± 0.051	0.503 ± 0.072
Cognitive functions and awareness symptoms	233	**0.682 ± 0.045**	0.554 ± 0.044	0.642 ± 0.049	0.651 ± 0.008	0.476 ± 0.046	0.534 ± 0.062	0.611 ± 0.059	0.611 ± 0.059	0.347 ± 0.071	0.639 ± 0.051	0.629 ± 0.024
Sleep disorders	215	**0.556 ± 0.030**	0.520 ± 0.026	0.502 ± 0.058	0.543 ± 0.056	0.538 ± 0.053	0.476 ± 0.015	0.535 ± 0.039	0.535 ± 0.039	0.429 ± 0.064	0.490 ± 0.053	0.501 ± 0.047
Carpal tunnel syndrome	210	**0.603 ± 0.042**	0.482 ± 0.067	0.549 ± 0.103	0.540 ± 0.051	0.525 ± 0.030	0.489 ± 0.050	0.523 ± 0.027	0.523 ± 0.027	0.372 ± 0.048	0.498 ± 0.030	0.503 ± 0.066
Mononeuritis of upper limb	210	**0.615 ± 0.028**	0.498 ± 0.068	0.522 ± 0.094	0.578 ± 0.025	0.535 ± 0.060	0.493 ± 0.048	0.509 ± 0.037	0.509 ± 0.037	0.385 ± 0.083	0.499 ± 0.040	0.528 ± 0.059
Other nervous-system symptoms	199	0.709 ± 0.066	0.522 ± 0.097	**0.719 ± 0.075**	0.631 ± 0.058	0.490 ± 0.063	0.623 ± 0.059	0.693 ± 0.026	0.693 ± 0.026	0.346 ± 0.059	0.667 ± 0.076	0.694 ± 0.066
Sleep apnea	194	0.574 ± 0.055	0.500 ± 0.056	**0.587 ± 0.067**	0.568 ± 0.077	0.483 ± 0.046	0.527 ± 0.101	0.559 ± 0.059	0.559 ± 0.059	0.361 ± 0.070	0.552 ± 0.043	0.551 ± 0.065
Extrapyramidal and movement disorders	175	0.629 ± 0.039	0.541 ± 0.077	**0.640 ± 0.022**	0.569 ± 0.030	0.475 ± 0.059	0.553 ± 0.039	0.634 ± 0.054	0.634 ± 0.054	0.350 ± 0.077	0.616 ± 0.015	0.607 ± 0.059
Altered mental status, unspecified	166	**0.666 ± 0.072**	0.540 ± 0.062	0.567 ± 0.148	0.609 ± 0.077	0.499 ± 0.036	0.535 ± 0.048	0.609 ± 0.102	0.609 ± 0.102	0.352 ± 0.091	0.651 ± 0.060	0.657 ± 0.052
Polyneuropathies	164	**0.648 ± 0.039**	0.572 ± 0.052	0.618 ± 0.090	0.549 ± 0.066	0.524 ± 0.122	0.593 ± 0.058	0.572 ± 0.070	0.572 ± 0.070	0.335 ± 0.210	0.641 ± 0.061	0.628 ± 0.068
Migraine	153	**0.588 ± 0.063**	0.489 ± 0.012	0.467 ± 0.029	0.494 ± 0.028	0.501 ± 0.019	0.484 ± 0.014	0.502 ± 0.029	0.502 ± 0.029	0.568 ± 0.045	0.471 ± 0.023	0.483 ± 0.028
Epilepsy, recurrent seizures, convulsions	132	0.604 ± 0.053	0.501 ± 0.062	0.591 ± 0.087	0.585 ± 0.052	0.574 ± 0.056	0.594 ± 0.091	0.562 ± 0.063	0.562 ± 0.063	0.352 ± 0.131	0.600 ± 0.034	**0.615 ± 0.055**
Dementias and cerebral degeneration	130	**0.761 ± 0.050**	0.604 ± 0.101	0.725 ± 0.043	0.672 ± 0.047	0.512 ± 0.081	0.659 ± 0.118	0.682 ± 0.075	0.682 ± 0.075	0.352 ± 0.091	0.734 ± 0.070	0.722 ± 0.075
Hemiplegia and hemiparesis	130	**0.663 ± 0.054**	0.485 ± 0.090	0.619 ± 0.072	0.542 ± 0.041	0.485 ± 0.128	0.548 ± 0.134	0.627 ± 0.031	0.627 ± 0.031	0.265 ± 0.070	0.627 ± 0.059	0.628 ± 0.068
Occlusion and stenosis of cerebral arteries	117	**0.670 ± 0.102**	0.594 ± 0.069	0.608 ± 0.069	0.637 ± 0.047	0.466 ± 0.070	0.561 ± 0.054	0.565 ± 0.080	0.565 ± 0.080	0.466 ± 0.082	0.659 ± 0.054	0.617 ± 0.049
Disturbances of skin sensation	117	0.521 ± 0.041	0.489 ± 0.014	0.530 ± 0.058	**0.556 ± 0.050**	0.484 ± 0.054	0.542 ± 0.084	0.491 ± 0.039	0.491 ± 0.039	0.432 ± 0.091	0.522 ± 0.042	0.484 ± 0.041
Dementias	116	**0.812 ± 0.025**	0.591 ± 0.104	0.769 ± 0.026	0.686 ± 0.067	0.483 ± 0.081	0.686 ± 0.030	0.706 ± 0.025	0.706 ± 0.025	0.316 ± 0.098	0.770 ± 0.051	0.749 ± 0.060
Cerebral palsy and other paralytic syndromes	111	**0.657 ± 0.055**	0.564 ± 0.145	0.450 ± 0.079	0.558 ± 0.061	0.501 ± 0.183	0.583 ± 0.160	0.652 ± 0.055	0.652 ± 0.055	0.232 ± 0.108	0.602 ± 0.078	0.656 ± 0.075
Transient cerebral ischemic attack (TIA)	103	**0.627 ± 0.077**	0.565 ± 0.054	0.577 ± 0.068	0.542 ± 0.086	0.533 ± 0.089	0.526 ± 0.026	0.475 ± 0.075	0.475 ± 0.075	0.348 ± 0.114	0.601 ± 0.054	0.582 ± 0.084
Hemorrhagic stroke	102	**0.641 ± 0.110**	0.561 ± 0.046	0.547 ± 0.118	0.568 ± 0.078	0.503 ± 0.108	0.578 ± 0.141	0.584 ± 0.075	0.584 ± 0.075	0.251 ± 0.051	0.594 ± 0.105	0.633 ± 0.103
Speech disturbance	101	0.668 ± 0.097	0.626 ± 0.044	0.577 ± 0.120	0.575 ± 0.115	0.480 ± 0.043	0.628 ± 0.074	0.618 ± 0.070	0.618 ± 0.070	0.423 ± 0.064	0.631 ± 0.068	**0.682 ± 0.079**
Alzheimer's disease	74	**0.859 ± 0.028**	0.601 ± 0.099	0.819 ± 0.034	0.715 ± 0.063	0.503 ± 0.066	0.711 ± 0.117	0.761 ± 0.093	0.761 ± 0.093	0.237 ± 0.080	0.837 ± 0.045	0.843 ± 0.049

Values are reported as mean C-index ± standard deviation across five cross-validation folds. LF denotes late fusion.
The sample count n indicates the number of observed events for the corresponding diagnosis.

Table 17: Overview of MRI contrasts, imaging techniques, and derived maps and their clinical relevance in neuroimaging.

Modality	**Full name**	**Clinical relevance**
T1	T_1-weighted imaging	Provides detailed brain anatomy for assessing brain volume and atrophy.
T2	T_2-weighted imaging	Sensitive to increased tissue water content, supporting the detection of edema and lesions.
FLAIR	Fluid-attenuated inversion recovery	Suppresses cerebrospinal fluid signal to improve the visibility of white matter lesions, including those associated with multiple sclerosis and cerebral small vessel disease.
DWI	Diffusion-weighted imaging	Detects diffusion abnormalities, particularly in acute ischemia; interpreted together with ADC maps.
DTI	Diffusion tensor imaging	Characterizes white matter microstructure and enables reconstruction of fiber pathways using tractography.
MTI	Magnetization transfer imaging	Probes macromolecular tissue properties and myelin-related changes; primarily used in research.
STAR*	T_2^*-weighted gradient-echo imaging	Sensitive to susceptibility effects from blood products and iron, supporting the detection of hemorrhage and cerebral microbleeds.
TRACE	Trace-weighted diffusion imaging	Combines diffusion-weighted images to reduce directional dependence; supports detection of diffusion abnormalities, particularly in acute ischemia.
ADC	Apparent diffusion coefficient map	Quantifies apparent water diffusivity and helps distinguish true diffusion restriction from T_2 shine-through.
SWI	Susceptibility-weighted imaging	Provides sensitive visualization of veins, blood products, and iron-related changes, particularly cerebral microbleeds.

Table 18: Selection of recent deep learning methods for MRI-based brain age prediction. n/a: not reported; ∼: inferred but not explicitly stated. (no access): full text unavailable. MAEs are not directly comparable due to differences in clinical cohorts, model complexity, and training splits.

Author	Year	Database	Architecture	MRI Modalities	2D/ 3D	MAE
Gao et al.	2023	ADNI	graph neural network	rs-fMRI	n/a	5.92
Ding et al.	2023	Southwest University Longitudinal Imaging Multimodal Brain Dataset	siamese convolutional neural network	rs-fMRI	3D	5.34
Hu et al.	2023	∼ proprietary (Guangzhou Women and Children's Medical Center)	convolutional neural network (ResNet-18)	T1-weighted MRI	3D	67.66 (days; resampled 3D volume)
Zhao et al.	2024	developing Human Connectome Project	transformers	T2-weighted structural MRI, diffusion MRI	∼ 3D	0.5 (weeks)
Sowmya et al.	2025	UK Biobank	hybrid deep learning framework (CNNs, transformers, and graph neural networks)	∼ T1-weighted structural MRI	∼ 3D	2.30
Kim et al.	2025	proprietary (Samsung Medical Center)	convolutional neural networks (DenseNet)	T1-weighted MRI	2D	2.73 (after age bias correction)
Kurian et al.	2025	(no access)	transformer	T1-weighted MRI, quantitative maps (QSM, R2*, and CSF fraction) derived from multi-echo GRE	(no access)	(no access)
Lee et al.	2026	ABIDE, CoRR, IXI, OASIS-3	backbone inspired by ConvNeXt, transformer	T1-weighted MRI, gray and white matter structural maps (derived from T1-weighted MRI)	3D	2.28; 3.58, 4.46; 2.59 (on each dataset)
Chang et al.	2026	ABIDE, ADNI, BGSP, HBN, CORR, fcon_1000, ICBM, OASIS-3	convolutional neural network	T1-weighted sMRI	2D (projection)	2.7–2.8; 2.50 (ensemble); 2.54 (after bias correction)
Jomsky et al.	2026	ADNI, BGSP, NIFD, SLIM, DLBS, SALD, AIBL, CoRR, SchizConnect, IXI, OAS1, OAS2, PPMI	convolutional neural networks (VGG8)	non-contrast T1w MRI (for deep learning-derived cerebral blood volume), structural MRI	∼ 3D	3.95

7.1 MRI PREPROCESSING

The MRI data were preprocessed using a standardized pipeline implemented in Python 3.11 with Advanced Normalization Tools (ANTs). First, raw Digital Imaging and Communications in Medicine (DICOM) images were converted to Neuroimaging Informatics Technology Initiative (NIfTI) format to ensure compatibility with downstream processing. Noise reduction was applied to improve the signal-to-noise ratio, followed by bias field correction to compensate for intensity inhomogeneities. Skull stripping was then performed to remove non-brain tissue and isolate intracranial structures. Subsequently, intensity normalization was applied to harmonize voxel intensity distributions across subjects. Finally, all images were spatially aligned to a common anatomical reference space via co-registration to the Montreal Neurological Institute (MNI) space, enabling voxel-wise comparability across the cohort. Given the high computational complexity of volumetric MRI data, all scans were initially downsampled to a resolution of $64 \times 64 \times 64$ voxels from their original resolution of 224^3.

7.2 ATTENTION COMPUTATION

Given an input sequence of 3D patch embeddings $Z \in \mathbb{R}^{N \times C}$, where N represents the number of tokens and C is the embedding dimension, embeddings are projected into query (Q), key (K), and value (V) matrices. This is done using learned linear transformations $W_Q^h, W_K^h, W_V^h \in \mathbb{R}^{C \times d_h}$ for each of the H attention heads. For a single head h, the matrices are:

$$Q_h = ZW_Q^h, \quad K_h = ZW_K^h, \quad V_h = ZW_V^h \tag{5}$$

where $Q_h, K_h, V_h \in \mathbb{R}^{N \times d_h}$ and $d_h = C/H$ is the dimension per head.

The attention scores for each head are computed using scaled dot-product attention. However, to make the model aware of the 3D spatial relationships, we integrate relative positional encodings (RPE) by adding a learnable bias term $B_h \in \mathbb{R}^{N \times N}$ to the attention scores before the softmax normalization. The attention output for a single head is thus calculated as:

$$\text{Attention}(Q_h, K_h, V_h) = \text{softmax}\left(\frac{Q_h K_h^\top}{\sqrt{d_h}} + B_h\right) V_h \tag{6}$$

The outputs from all H attention heads are concatenated and passed through a final linear projection layer ($W_O \in \mathbb{R}^{C \times C}$) to produce the final output of the MHSA module:

$$\text{MHSA}(Z) = \text{Concat}(\text{head}_1, \ldots, \text{head}_H) W_O \tag{7}$$

This mechanism allows a 3D attention-based architecture to effectively model both local and global spatial interactions, informed by the relative 3D positions of the tokens.

7.3 COMPUTATIONAL COMPLEXITY

The computational complexity of MICViT is governed by its hierarchical 3D tokenization and window-based attention. At stage i, let M denote the number of modalities, N_i the number of tokens per modality, C_i the feature dimension, and $W_i = w_i^{(d)} w_i^{(h)} w_i^{(w)}$ the 3D window volume. Projection and feed-forward layers scale as $\mathcal{O}(MN_iC_i^2)$.

Modality-separated attention operates on intra-modality windows with complexity $\mathcal{O}(MN_iW_iC_i)$, which is linear in N_i for fixed window size. Cross-modal attention jointly attends across modalities within each window and increases the cost to $\mathcal{O}(M^2N_iW_iC_i)$ through pairwise modality interactions. Global attention uses stage-wise global queries with window-local keys and values, avoiding quadratic scaling in the total number of tokens.

Aggregating over all stages yields

$$\mathcal{O}\left(\sum_i \left[(S_i + X_i)MN_iC_i^2 + S_iMN_iW_iC_i + X_iM^2N_iW_iC_i\right]\right), \tag{8}$$

where S_i and X_i denote the numbers of modality-separated and cross-modal attention blocks at stage i, respectively. Overall, MICViT scales linearly in the number of tokens N_i (for fixed window size), in contrast to full volumetric self-attention, which scales quadratically as $\mathcal{O}(M^2 N_i^2 C_i)$. This favorable scaling is critical for enabling high-resolution multimodal 3D neuroimaging.

7.4 Multi-modal patch embedding.

Let $\mathbf{x} \in \mathbb{R}^{B\times M\times H\times W\times D}$ denote a batch of M co-registered MRI modalities. We first map each modality to a latent patch representation using a modality-specific 3D patch embedding operator. Specifically, for modality m, we extract $\mathbf{x}^{(m)} \in \mathbb{R}^{B\times 1\times H\times W\times D}$ and compute

$$\mathbf{z}^{(m)} = f_m\left(\mathbf{x}^{(m)}\right) \in \mathbb{R}^{B\times H'\times W'\times D'\times C},$$

where f_m consists of a strided 3D convolution with kernel size P and stride S, followed by a local convolutional refinement block. The strided convolution performs patchification and linear projection into a C-dimensional embedding space, while the refinement block adds an inductive bias for local anatomical continuity through depthwise 3D convolution, nonlinearity, squeeze-and-excitation, and pointwise mixing. Importantly, this refinement block does not further reduce spatial resolution at the embedding stage.

The modality-specific embeddings are then stacked as

$$\mathbf{z} = \text{stack}\left(\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(M)}\right) \in \mathbb{R}^{B\times M\times H'\times W'\times D'\times C}.$$

This design preserves an explicit modality axis, rather than collapsing modalities into channels at the input. As a result, subsequent transformer blocks can model intra-modality structure and inter-modality interactions separately, which is particularly desirable for multi-contrast MRI where complementary information is distributed across modalities. Note that the stem layers are not identical across all models and baselines. While MICViT employs modality-specific patch embeddings with convolutional refinement, baseline architectures use their respective standard input stems, reflecting their underlying design assumptions.

7.5 Scaling Floating Point Operations (FLOPs) and Parameters

Establishing a fair comparison between 3D ViT and 3D CNN presents significant challenges not typically encountered in 2D benchmarks. While 2D models can often be aligned in both parameter count and FLOPs, 3D architectures exhibit disparate scaling behaviors, particularly concerning the input's linear dimension L. For 3D ViT, with N tokens derived from input patches ($N \propto L^3$), FLOPs are dominated by the self-attention mechanism, scaling as $O(N^2 D) = O(L^6 D)$, where D is the embedding dimension. In contrast, their parameter count, primarily determined by projection and MLP layers, remains largely independent of L, scaling as $O(D^2)$. Conversely, 3D CNN FLOPs scale more moderately with input size for a fixed architecture (kernel size K, channels C, depth D_{layer}), typically as $O(L^3 K^3 C^2 D_{layer})$. However, achieving FLOPs comparable to 3D ViT often necessitates substantial increases in network depth or width (D_{layer} or C), leading to a dramatic inflation of parameters, which scale as $O(D_{layer} K^3 C^2)$. Given this inherent difficulty in simultaneously matching both FLOPs and model parameters, we advocate for a two-staged comparison strategy: one benchmark matching models based on FLOPs to assess computational efficiency, and a second matching models based on parameter count to evaluate model capacity. We posit that for research focusing on scaling laws, the parameter-matched comparison holds particular significance.

7.6 Baseline adaptations.

For mmFormer, we retain the multimodal encoder and fusion modules but remove the segmentation decoder, replacing it with the same image-level prediction head used for the other baselines. This adapts the architecture to regression and classification tasks rather than voxel-level segmentation.

7.7 Disease onset prediction

To assess the clinical and prognostic relevance of the learned representations, we evaluated disease-onset prediction across 38 diagnostic endpoints using survival modeling, which accounts for both the

time to diagnosis and right-censored observations (Table 16). MICViT was compared with a broad set of CNN-, transformer-, mixture-of-experts-, and fusion-based architectures, including GCViT, FlexMoE, ConvNeXt, Swin, mmFormer, ViT, ResNet, and late-fusion variants. MICViT achieved the highest mean C-index for 24 of the 38 endpoints, demonstrating strong performance across a heterogeneous range of neurological, cerebrovascular, psychiatric, and systemic conditions. Particularly strong results were observed for Alzheimer's disease (0.859 ± 0.028), dementia (0.812 ± 0.025), dementias and cerebral degeneration (0.761 ± 0.050), cognitive symptoms (0.682 ± 0.045), stroke (0.666 ± 0.046), and hemiplegia or hemiparesis (0.663 ± 0.054). MICViT also obtained the best mean performance for several endpoints with larger event counts, including essential hypertension, hypertension, chronic kidney disease, and hypotension. For repeated falls and several movement- or nervous-system-related outcomes, its performance remained close to the strongest competing model even when it did not rank first. Overall, these results indicate that explicitly integrating local anatomical patterns, global spatial context, and cross-modal interactions enables MICViT to learn broadly informative MRI representations that generalize beyond a single clinical outcome. Nevertheless, results for endpoints with few observed events or high variation across folds should be interpreted cautiously and require validation in independent cohorts.

7.8 Evaluation under Complete and Heterogeneous Modality Availability

To evaluate model performance under realistic variations in MRI acquisition, we consider two modality-availability settings. The heterogeneous modality-availability setting includes all seven modality combinations observed in ADNI, thereby retaining sessions with both complete and incomplete imaging data (Appendix 5). T1-weighted MRI is available for every session and serves as the common modality, whereas DTI, FLAIR, and STAR are available only for subsets of the data. The most frequent configuration contains only T1 ($n = 4{,}011$), followed by T1, FLAIR, and STAR without DTI ($n = 3{,}230$), while $2{,}384$ sessions contain all four modalities. Less frequent configurations include T1 with FLAIR ($n = 875$), T1 with STAR ($n = 62$), T1 with DTI and FLAIR ($n = 39$), and T1 with DTI and STAR ($n = 38$). This heterogeneous setting reflects the variability encountered in clinical and longitudinal imaging datasets, where not every modality is acquired at each examination. In contrast, the complete four-modality subset is restricted to sessions for which T1, DTI, FLAIR, and STAR are all available. This setting provides standardized multimodal inputs and enables a controlled comparison of the models when no modality is missing. Evaluating both settings therefore distinguishes a model's ability to integrate a fixed set of complete multimodal inputs from its robustness to heterogeneous and incomplete modality configurations.

7.9 Alzheimer's Disease Classification under Heterogeneous Modality Availability

To assess whether the learned multimodal representations generalize to a clinically relevant downstream task, we evaluate all models on the classification of cognitively normal (CN) participants and patients with Alzheimer's disease (AD) in ADNI (Appendix, Fig.14). Across all available modality configurations, MICViT achieves the highest accuracy (0.914 ± 0.017) and Macro-F1 score (0.836 ± 0.065), exceeding the respective second-best results by 4.2 and 2.6 percentage points. Its AUC of 0.913 ± 0.038 is the second highest, following mmFormer at 0.930 ± 0.013. The aggregated confusion matrices show that MICViT correctly classifies 929 CN and 169 AD cases, corresponding to class-wise recalls of approximately 96.7% and 69.8%, respectively. It produces 32 false-positive and 73 false-negative predictions (Appendix, Fig.6). In comparison, mmFormer correctly identifies 947 CN cases but detects only 103 AD cases, resulting in 139 missed AD cases and a lower Macro-F1 score (0.739 ± 0.086). FlexMoE identifies the largest number of AD cases (196; approximately 81.0% AD recall) but misclassifies 110 CN participants as AD, illustrating a sensitivity–specificity trade-off. ConvNeXt and GCViT also miss more AD cases than MICViT and achieve lower overall accuracy and Macro-F1 scores. On the complete four-modality subset, MICViT obtains the best performance across all three metrics, reaching an accuracy of 0.823 ± 0.027, an AUC of 0.833 ± 0.026, and a Macro-F1 score of 0.710 ± 0.031. It correctly identifies 119 of 242 AD cases (approximately 49.2% recall) while retaining a CN recall of approximately 90.6%. The competing models correctly classify between 58 and 115 AD cases. Performance on the complete subset is generally lower than when all available modality configurations are included, particularly for the minority AD class. Overall, MICViT achieves the strongest Macro-F1 score in both evaluation settings and maintains a comparatively balanced recognition of CN and AD cases.

7.10 DETAILS OF RESOURCES USED

Our experiments are conducted on a High-Performance Cluster (HPC) with the following environment. We use 80GB NVIDIA H100 and NVIDIA H200 GPUs for training and evaluation. The cluster nodes are equipped with Intel Xeon Platinum (8468 or 8568Y+) or AMD EPYC (9534 or 9355) CPUs and 754GB–2TB of RAM.