STRUCTURE BEFORE TIME: ANATOMY-PRESERVING REPRESENTATION LEARNING FOR LONGITUDINAL 3D BRAIN MRIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Longitudinal modeling of high-resolution 3D brain MRI requires representations that preserve localized disease information while remaining tractable as patient histories grow. This is particularly challenging in Alzheimer's Disease (AD), where structural changes are spatially heterogeneous, subtle at early disease stages such as Mild Cognitive Impairment (MCI), and observed over irregular follow-up. We introduce **Structure Before Time (SBT)**, an anatomy-preserving representation learning principle for longitudinal brain MRI. SBT uses neuroanatomical structure to constrain representation formation before global and temporal aggregation, yielding compact, anatomically informed representations for longitudinal modeling. Controlled experiments show that the resulting gains arise from meaningful anatomical organization and its early introduction into the framework. Across AD diagnosis, MCI progression, and survival prediction, SBT consistently improves over single-scan and longitudinal baselines, while benefiting increasingly from accumulated history. The learned representation further transfers across cohorts, remains effective under alternative anatomical parcellation, extends to Frontotemporal dementia (FTD), and localizes controlled abnormalities with known regional ground truth in a synthetic dataset. These results support anatomy-preserving representation formation as an effective inductive bias for high-dimensional longitudinal neuroimaging.

1 INTRODUCTION

Neurodegenerative disease unfolds over time, yet its structural manifestations are neither spatially uniform nor temporally synchronous (Apostolova et al. (2012); Ferreira et al. (2017)). In Alzheimer's disease (AD), different anatomical regions may remain relatively preserved, undergo gradual structural decline, or become affected at different stages of disease progression (Braak & Braak (1991); Braak et al. (2006); Jack et al. (2013); Ferreira et al. (2017); Planche et al. (2022); Jellinger (2022); Rauchmann et al. (2025)). These differences are particularly important in mild cognitive impairment (MCI), where disease-related structural deviations can be subtle and difficult to distinguish from cognitively normal (CN) individuals (Zhang et al. (2017)). Longitudinal T1-weighted structural MRI (sMRI) therefore provides more than repeated observations of the same brain: it captures how spatially localized anatomy changes over time (Vemuri & Jack Jr (2010); Whitwell et al. (2007); Frisoni et al. (2010)). Exploiting this information requires representations that preserve where disease-related evidence arises while remaining suitable for irregular and variable-length patient histories. Longitudinal neuroimaging is therefore fundamentally a *representation learning problem before it becomes a temporal prediction problem*.

Learning such representations from high-resolution 3D MRI creates a fundamental trade-off. Fine-grained spatial information is especially valuable when disease effects are localized or subtle, as in MCI (Frisoni et al. (2010)). Retaining dense volumetric representations across examinations, however, makes spatiotemporal interaction increasingly expensive as both image resolution and longitudinal history grow. Conversely, early compression into global scan representations makes temporal modeling tractable, but can mix evidence originating from distinct anatomical structures and obscure its regional organization. Existing approaches therefore often trade spatial richness for temporal flexibility through downsampling, feature compression, or restricted longitudinal histories. As

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.

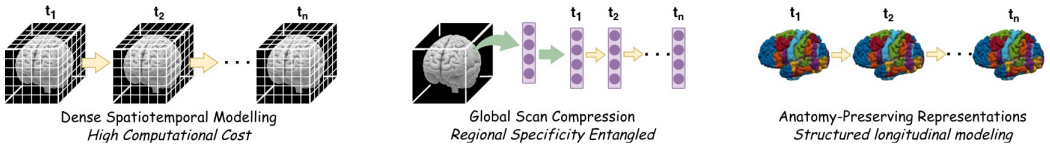

Figure 1: **Longitudinal neurodegeneration poses a structured representation problem.** Dense volumetric modeling preserves spatial detail at a rapidly growing computational cost as patient history increases, whereas global representations make temporal modeling tractable but can entangle localized evidence. An effective longitudinal representation should therefore retain localized information while supporting efficient modeling of irregular, variable-length histories.

illustrated in Fig. 1, the central challenge is to retain meaningful local information while remaining tractable as longitudinal context increases.

We present and study the hypothesis that *structure should shape representation formation before time is modeled*. Neuroanatomy provides a persistent spatial reference across examinations: regional appearance may change with aging or disease, while anatomical organization remains meaningful across visits. This motivates an anatomy-preserving strategy in which high-resolution observations are first contextualized within anatomically defined regions before broader spatial and temporal aggregation. Anatomy therefore acts as a structured inductive bias on *how* representations are formed. The resulting factorization retains fine-grained information within meaningful regional contexts while allowing longitudinal modeling to operate on compact, anatomically informed representations rather than dense volumetric tokens from every examination.

We instantiate this principle for longitudinal T1-weighted brain MRI, with AD as the primary clinical setting, and evaluate it through controlled experiments that probe the representation principle from complementary perspectives. Our results support a broader view: when predictive evidence evolves within persistent substructures of high-dimensional observations, preserving those structures during representation formation provides a useful inductive bias for longitudinal learning. Our contributions are three-fold:

- **Structure-before-time (SBT) learning principle.** We formulate longitudinal neuroimaging as a representation learning problem in which anatomically meaningful local structure guides representation formation before broader contextual and temporal aggregation.
- **Anatomy-preserving longitudinal framework.** We instantiate this principle for high-resolution 3D MRI through a spatially factorized model that forms locally contextualized anatomical representations before temporal modeling. It supports irregular, variable-length longitudinal histories while preserving regional information over time.
- **Controlled and comprehensive validation.** We systematically investigate *what*, *when*, and *how* anatomical structure contributes to representation learning. We then evaluate the resulting representations across diagnosis, progression, prognosis, cohorts, and alternative anatomical parcellation. Using a synthetic 3D dataset with known regional ground truth and longitudinal progression, we further show that SBT localizes controlled abnormalities and captures variations in their underlying severity as they evolve over time.

2 Related Works

Longitudinal modeling of structural MRI. Deep learning methods for longitudinal neuroimaging differ primarily in how they balance spatial representation richness against the amount of temporal context that can be modeled. One family first compresses each 3D examination into a compact scan-level representation and subsequently integrates observations using recurrent or attention-based temporal models, often under fixed-interval assumptions (Cui et al. (2019)) or capped histories (Jomeiri et al. (2024); Aghajanian et al. (2025)). To reduce the computational burden of longitudinal learning, several approaches (Ouyang et al. (2023); Zhao et al. (2021); Ouyang et al. (2021; 2022)) further rely on aggressive input downsampling (64^3 volumes) or curated 2D slice-selection (Hu et al. (2023)), compromising fine-grained anatomical detail. A complementary direction retains richer cross-visit spatial information while restricting the temporal context. LongFormer (Chen et al. (2024)), for example, models a current–prior scan pair using optical flow and deformable attention. Other ap-

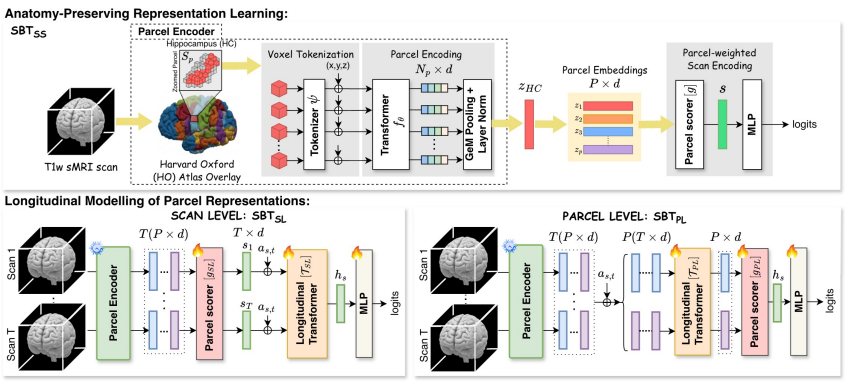

Figure 2: **Structure Before Time (SBT) Framework.** Each high-resolution 3D MRI is first represented through anatomically structured local contexts, producing a compact set of region-level representations before longitudinal modeling. Parcel Encoder (PE) is shared across anatomical regions, while atlas correspondence determines the context in which representations are formed. Longitudinal information is then integrated either after scan-level aggregation (SBT_{SL}) or while retaining regional correspondence through time (SBT_{PL}).

proaches similarly operate on fixed scan pairs or predetermined longitudinal histories (Liu et al. (2024)), thus truncating the dataset. Such formulations preserve richer cross-scan information, but do not directly extend to arbitrary-length, irregularly sampled histories. Across these approaches, a recurring trade-off therefore emerges: reducing spatial dimensionality facilitates temporal modeling over longer histories, whereas retaining high-dimensional spatial information typically constrains the number or configuration of longitudinal observations. (See Appendix Sec C.3 for further discussion on broader literature)

3 Method

3.1 Structure-Before-Time Formulation

Consider a subject s with a longitudinal history $\mathcal{H}_s = \{(X_{s,t},\, a_{s,t})\}_{t=1}^{T_s}$, where $X_{s,t} \in \mathbb{R}^{H\times W\times D}$ denotes a 3D T1-weighted MRI acquired at age $a_{s,t}$, and T_s varies across subjects. $\mathcal{H}_s$ contains scans only available up to the prediction point. Given $\mathcal{H}_s$, the objective is to predict a subject-level outcome y_s while retaining spatially localized information that may evolve across examinations. We follow the *Structure Before Time (SBT)* principle (Fig 2): each scan is first represented as a set of anatomically organized features $Z_t = \{z_{1,t}, \ldots, z_{P,t}\}$, where $z_{p,t} \in \mathbb{R}^d$ denotes the representation of parcel p. These structured parcel representations then form the basis for longitudinal modeling.

3.2 Anatomy-Preserving Representation Learning

Volumetric tokenization. Each 3D scan volume is partitioned into N non-overlapping v^3 patches. A shared tokenizer ψ maps patch $X_{i,t}$ to $x_{i,t} \in \mathbb{R}^d$, and spatial 3D coordinates c_i are encoded by an MLP (ϕ): $\bar{x}_{i,t} = \psi(X_{i,t}) + \phi(c_i)$.

Anatomical correspondence. A probabilistic brain atlas defines P regions in the common image space. Let $M \in \mathbb{R}_{\geq 0}^{N\times P}$ denote patch-to-region membership, where M_{ip} measures the fraction of patch i occupied by parcel p. The context of region p is $\mathcal{I}_p = \{i : M_{ip} > 0\}$. The resulting soft patch-to-region assignments allow a boundary token to contribute to more than one anatomical region, avoiding an artificially hard partition at uncertain regional boundaries.

Structure before contextualization. The central design choice in SBT is to introduce anatomical structure *before* attention-based contextualization. Tokens associated with each parcel region are processed independently using a shared Transformer f_θ (Dosovitskiy et al. (2021)) followed by GeM pooling (Radenović et al. (2018)) to obtain a fixed-dimensional anatomical representation despite differences in regional shape or volume:

$$H_{p,t} = f_\theta\left(\{\bar{x}_{i,t}\}_{i\in\mathcal{I}_p}\right), \qquad z_{p,t} = \text{LN}(\text{GeM}(H_{p,t})) \in \mathbb{R}^d,$$

The same parameters θ are shared across all parcels; thus, SBT does not learn a separate encoder for each anatomical region. Instead, anatomy determines the context within which the shared representation function operates. This construction imposes an anatomical locality constraint. For a parcel representation $z_{p,t}$, tokens outside its atlas-defined support do not participate in its contextualization: for $j \notin \mathcal{I}_p$, $\partial z_{p,t}/\partial \bar{x}_{j,t} = 0$. In contrast, if unrestricted global interaction precedes anatomical grouping, a subsequently pooled regional representation cannot guarantee that the feature was formed solely from its corresponding anatomical context (discussed in detail in Appendix Sec C.5). SBT therefore uses anatomy to constrain *how representations are formed*.

Anatomically informed scan representation. For scan-level aggregation, a shared Parcel scorer g (an MLP) assigns an importance score to each parcel, quantifying its relative regional contribution to the model's prediction,

$$\alpha_{p,t} = \frac{\exp(g(z_{p,t})/\tau)}{\sum_q \exp(g(z_{q,t})/\tau)}, \qquad s_t = \sum_{p=1}^{P} \alpha_{p,t} z_{p,t}.$$

Thus, global aggregation occurs only after each local representation has been contextualized within its anatomical support. During single-scan training (SBT_{SS}), s_t is passed to an MLP prediction head for binary classification task.

3.3 Longitudinal Modeling over Structured Representations

After learning the anatomy-preserving representation, the Parcel encoder is frozen, and the parcel representations for each scan are cached for longitudinal learning. Besides reducing memory requirements, this preserves a common anatomy-conditioned representation space while enabling efficient modeling of variable-length longitudinal histories. Visit ages $a_{s,t}$ of the subject are encoded as additive temporal embeddings via an MLP: $e_{s,t} = \phi_t(a_{s,t})$. The structured representation permits longitudinal evidence to be aggregated at different levels of granularity.

Scan-level Longitudinal Modeling (SBT_{SL}). Parcel features are first aggregated into the anatomy-informed scan representation s_t using the parcel weights ($\alpha_{p,t}$) from the parcel scorer (g_{SL}). The resulting sequence of scan representations is then modeled by a Longitudinal Transformer $\mathcal{T}_{SL}$:

$$h_s = \text{Pool}[\mathcal{T}_{SL}(s_1 + e_{s,1}, \ldots, s_{T_s} + e_{s,T_s})]$$

This formulation retains anatomical information in the construction of every scan representation while keeping the temporal sequence compact.

Parcel-level Longitudinal Modeling (SBT_{PL}). Alternatively, anatomical correspondence can be retained explicitly through the temporal stage. For each region p, the corresponding representations across visits are modeled as:

$$v_{p,s} = \text{Pool}[\mathcal{T}_{PL}(z_{p,1} + e_{s,1}, \ldots, z_{p,T_s} + e_{s,T_s})]$$

where the same Longitudinal Transformer ($\mathcal{T}_{PL}$) is shared across all parcels. Trajectory representations are then aggregated using parcel weights ($\alpha_{p,t}$) from the parcel scorer (g_{PL}), $\beta_{p,s} = \text{softmax}_p(g_{PL}(v_{p,s}))$, giving $h_s = \sum_p \beta_{p,s} v_{p,s}$.

The two formulations differ principally in the stage at which anatomical representations are globally aggregated: SBT_{SL} aggregates them before temporal interaction, and SBT_{PL} retains parcel correspondence throughout longitudinal modeling. Both support irregular, variable-length histories. The final representation h_s is passed to an MLP head for binary classification task.

3.4 Computational Factorization

Let N denote the number of spatial tokens per scan and N_p the number associated with anatomical region p. A hypothetical joint self-attention model operating over all N spatial tokens from T visits would require $\mathcal{O}(T^2N^2)$ pairwise interactions. SBT instead performs anatomical contextualization independently at each visit, with cost $\mathcal{O}(T\sum_{p=1}^{P} N_p^2)$, followed by longitudinal interaction over compact representations: $\mathcal{O}(T^2)$ for SBT-SL and $\mathcal{O}(PT^2)$ for SBT-PL. Hence,

$$\text{SBT}_{\text{SL}}\text{: } \mathcal{O}\left(T\sum_p N_p^2 + T^2\right), \qquad \text{SBT}_{\text{PL}}\text{: } \mathcal{O}\left(T\sum_p N_p^2 + PT^2\right).$$

The quadratic temporal term, therefore, acts on parcel-level/ scan-level representations rather than on the complete set of high-resolution volumetric tokens. Thus, fine-grained spatial computation grows linearly with the number of visits.

4 EXPERIMENTS

We evaluate SBT from four complementary perspectives: predictive performance, the role of anatomical structure, the use of longitudinal information, and generalization beyond the development setting. We further test whether the resulting representations preserve localized disease evidence under controlled ground truth.

4.1 SETUP

Public Datasets. We use three longitudinal T1w sMRI cohorts spanning the AD spectrum: ADNI (Jack Jr et al. (2008)), AIBL (Fowler et al. (2021)), and MIRIAD (Malone et al. (2013)). ADNI (phases 1/GO/2/3/4) is used as primary development cohort. We additionally use NIFD (Rosen et al. (2010)) for Frontotemporal Dementia (FTD) classification to examine applicability beyond the AD spectrum. All scans are processed through a common pipeline and registered to MNI152 1 mm isotropic (Mazziotta et al. (1995)) template space. We evaluate AD/CN classification across ADNI, AIBL, and MIRIAD, and stable or progressive MCI (sMCI/pMCI) progression prediction on ADNI. We use the Harvard-Oxford (HO) atlas (FMRIB Software Library) with 69 parcels to define the anatomical context. It is selected for its volumetric macro-anatomical coverage of structures relevant to neurodegeneration. Cohort statistics, atlas, and preprocessing details are provided in Appendix Sec. B.

Synthetic dataset. To directly test whether SBT preserves the spatial source and severity of an abnormal signal, we additionally construct a controlled 3D synthetic dataset analogous to a longitudinal brain setting with known parcel ground truth. We generated 62,458 3D samples (Fig 3) of 5000 subjects with 6 fixed parcels (spheres & cubes). The dataset captures two trajectories across subjects: CN→CN, CN→ATR (atrophy). Dataset design and generation details are provided in Appendix Sec. B.4.

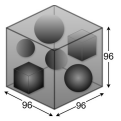

Figure 3: 3D Sample.

Evaluation protocol. We use diagnosis-stratified 5-fold cross-validation with subject-wise splits, ensuring that all visits from a subject remain within the same fold. Further, 10% of each training fold is used for validation. All methods within a controlled comparison use identical subject splits and longitudinal observations. All values reported in the tables are $mean_{(\pm std)}$ over 5-folds. Training and implementation details are provided in Appendix Sec. C.2.

Baselines. We compare SBT with complementary baselines that isolate the effects of backbone architecture, anatomical structuring, and temporal architecture from SBT's representation principle. Standard 3D CNNs and ViTs (Folego et al. (2020); Korolev et al. (2017); Ruiz et al. (2020); Zhang et al. (2021); Lu et al. (2025)) provide conventional volumetric representations widely used for AD and MCI tasks, while AAGN (Jiang & Miao (2024)) provides an anatomy-aware CNN comparison for single scan. LongFormer (Chen et al. (2024)) presents a dedicated paired-scan longitudinal model. We further construct matched DenseNet-121- and ViT-based longitudinal baselines, using either a GRU (Cho et al. (2014)) or the same Longitudinal Transformer (LT) employed by SBT, and train them with the same staged supervision. Further details are discussed in Appendix Sec. C.3.

4.2 PERFORMANCE ON ALZHEIMER'S DISEASE DIAGNOSIS

Table 1 compares SBT across single-scan, paired-scan, and variable-length longitudinal settings. A consistent pattern emerges across input regimes. At the single-scan level, SBT improves over conventional CNN and ViT representations and remains stronger than the anatomy-aware AAGN baseline on the primary ADNI tasks. AAGN introduces anatomical guidance after convolutional feature extraction, whereas SBT uses anatomical support to constrain the context in which representations are formed. The results, therefore, favor using anatomy to shape representation formation early using a transformer backbone. The longitudinal comparisons strengthen this observation. SBT consistently improves over paired-scan LongFormer and, more importantly, over matched DenseNet+LT and ViT+LT baselines. Because these controlled comparisons preserve the temporal architecture

Table 1: Prior-work and baseline comparison across ADNI, AIBL, MIRIAD and NIFD. Values are mean$_{(\text{std})}$ over 5 folds. Best results are **bold** for single & paired and best & second best for longitudinal inputs. (LT.: Longitudinal Transformer; Inf.: Inferencing with ADNI ckpt; FT: Finetuning with ADNI ckpt; 2T: paired scans, current & prior; e2e: end-to-end run)

Model	ADNI [AD/CN]		ADNI [sMCI/pMCI]		AIBL [AD/CN]		MIRIAD [AD/CN]		NIFD [FTD/CN]	
	BAcc(%)	AUC(%)	BAcc(%)	AUC(%)	BAcc(%)	AUC(%)	BAcc(%)	AUC(%)	BAcc(%)	AUC(%)
Single scan										
3D ResNet-152 [He et al. (2016)]	77.12$_{(3.13)}$	85.98$_{(4.33)}$	59.10$_{(2.61)}$	64.26$_{(1.32)}$	64.21$_{(2.20)}$	79.42$_{(3.06)}$	83.73$_{(4.24)}$	90.26$_{(4.12)}$	64.71$_{(3.43)}$	75.14$_{(2.27)}$
3D DenseNet-121 [Huang et al. (2017a)]	77.37$_{(2.68)}$	85.80$_{(3.54)}$	59.23$_{(3.96)}$	65.02$_{(4.86)}$	64.89$_{(3.74)}$	79.34$_{(3.09)}$	85.13$_{(4.43)}$	90.54$_{(3.50)}$	76.86$_{(2.80)}$	83.51$_{(3.86)}$
3D ViT [Dosovitskiy et al. (2021)]	78.06$_{(3.56)}$	83.63$_{(2.61)}$	60.82$_{(2.80)}$	64.70$_{(4.93)}$	63.26$_{(2.58)}$	78.26$_{(3.00)}$	82.29$_{(4.84)}$	84.52$_{(4.29)}$	52.96$_{(1.98)}$	60.68$_{(7.12)}$
M3T [Jang & Hwang (2022)]	80.72$_{(1.16)}$	87.57$_{(3.85)}$	61.15$_{(3.20)}$	67.13$_{(2.45)}$	66.50$_{(3.86)}$	83.06$_{(3.35)}$	83.02$_{(2.28)}$	91.03$_{(5.35)}$	78.94$_{(3.22)}$	84.09$_{(3.25)}$
AAGN [Jiang & Miao (2024)]	81.84$_{(1.71)}$	89.80$_{(1.20)}$	63.75$_{(2.64)}$	72.75$_{(2.67)}$	65.34$_{(2.50)}$	79.49$_{(3.84)}$	84.06$_{(4.92)}$	86.52$_{(4.50)}$	80.81$_{(3.28)}$	85.94$_{(1.74)}$
SBT$_{\text{SS}}$	**84.21**$_{(3.84)}$	**91.18**$_{(2.86)}$	66.47$_{(3.14)}$	78.89$_{(2.06)}$	68.18$_{(0.90)}$	79.57$_{(4.95)}$	87.26$_{(4.31)}$	86.96$_{(4.29)}$	80.98$_{(3.12)}$	86.86$_{(3.45)}$
SBT$_{\text{SS}}$ (Inf.)	–	–	–	–	82.62$_{(4.46)}$	91.22$_{(3.86)}$	89.66$_{(4.18)}$	97.08$_{(3.35)}$	–	–
SBT$_{\text{SS}}$ (FT)	–	–	**70.67**$_{(2.13)}$	**80.19**$_{(2.09)}$	**84.03**$_{(2.21)}$	**93.55**$_{(1.60)}$	**92.90**$_{(4.76)}$	**97.53**$_{(4.89)}$	**81.95**$_{(2.83)}$	**87.16**$_{(3.24)}$
FT gains over SBT$_{\text{SS}}$	–	–	↑ 4.20	↑ 1.30	↑ 1.41	↑ 3.24	↑ 3.24	↑ 0.45	↑ 0.07	↑ 0.30
Paired scans										
Longformer [Chen et al. (2024)]	82.84$_{(2.61)}$	91.20$_{(3.32)}$	64.91$_{(3.06)}$	77.13$_{(2.54)}$	67.47$_{(3.19)}$	84.71$_{(3.46)}$	85.57$_{(3.32)}$	90.07$_{(3.38)}$	80.17$_{(3.36)}$	86.14$_{(3.17)}$
SBT$_{\text{SL}}$ (2T)	**86.12**$_{(3.44)}$	**93.06**$_{(2.58)}$	**73.50**$_{(2.29)}$	**81.53**$_{(1.55)}$	**85.69**$_{(3.99)}$	**88.29**$_{(4.24)}$	**95.50**$_{(3.54)}$	**96.64**$_{(4.19)}$	**84.98**$_{(2.34)}$	**88.73**$_{(2.47)}$
Any length histories with any interval										
3D DenseNet+GRU	79.37$_{(3.64)}$	87.64$_{(2.03)}$	62.93$_{(1.51)}$	75.57$_{(1.06)}$	65.04$_{(3.87)}$	80.26$_{(2.35)}$	83.82$_{(3.71)}$	89.86$_{(2.20)}$	78.19$_{(2.54)}$	85.62$_{(3.03)}$
3D DenseNet+LT.	79.12$_{(3.55)}$	87.62$_{(2.17)}$	64.75$_{(2.43)}$	73.04$_{(1.82)}$	65.77$_{(4.03)}$	81.86$_{(2.86)}$	84.77$_{(3.66)}$	91.64$_{(3.01)}$	79.19$_{(3.14)}$	84.91$_{(2.93)}$
3D ViT+LT.	80.97$_{(2.49)}$	87.98$_{(1.55)}$	63.63$_{(2.26)}$	73.31$_{(1.09)}$	64.26$_{(2.57)}$	81.66$_{(2.16)}$	85.55$_{(5.75)}$	92.58$_{(2.04)}$	78.01$_{(3.45)}$	82.60$_{(3.83)}$
SBT$_{\text{SL}}$ (Inf.)	–	–	–	–	82.75$_{(3.07)}$	90.58$_{(2.18)}$	90.83$_{(4.19)}$	93.42$_{(3.62)}$	–	–
SBT$_{\text{SL}}$ (e2e)	88.78$_{(1.98)}$	**95.47**$_{(1.94)}$	72.66$_{(2.28)}$	82.31$_{(2.72)}$	85.60$_{(2.98)}$	91.62$_{(3.21)}$	96.02$_{(2.60)}$	97.53$_{(1.74)}$	84.13$_{(3.04)}$	89.70$_{(3.01)}$
SBT$_{\text{SL}}$	**89.20**$_{(2.21)}$	94.84$_{(1.66)}$	**78.26**$_{(0.91)}$	**85.96**$_{(1.18)}$	86.86$_{(3.47)}$	90.89$_{(4.69)}$	**96.90**$_{(3.55)}$	98.05$_{(3.22)}$	**85.30**$_{(2.58)}$	**90.01**$_{(2.02)}$
SBT$_{\text{PL}}$	88.27$_{(3.80)}$	94.58$_{(1.57)}$	76.65$_{(1.90)}$	85.86$_{(0.60)}$	**88.38**$_{(3.09)}$	**92.21**$_{(3.06)}$	96.54$_{(3.58)}$	**98.33**$_{(3.00)}$	84.50$_{(3.31)}$	89.63$_{(2.57)}$
Best longitudinal gains	↑ 4.99	↑ 4.29	↑ 7.59	↑ 5.77	↑ 4.35	↓ 1.34	↑ 4.00	↑ 0.80	↑ 3.35	↑ 2.85

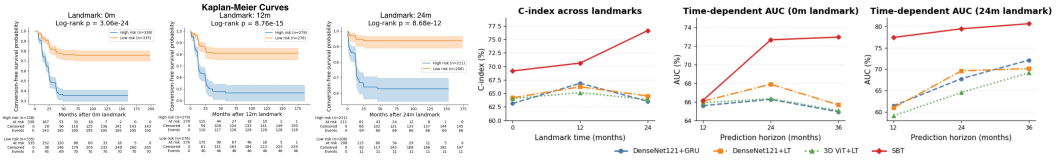

Figure 4: **Longitudinal history with SBT$_{\text{SL}}$ improves prognostic discrimination.** Landmark-based survival evaluation uses scans observed up to each landmark. The figure shows Kaplan-Meier risk stratification (Kaplan & Meier (1958)), C-index (Harrell et al. (1982)) changes with landmark months, and time-dependent AUCs across varying landmarks for increasing prediction horizons.

while changing the representation presented to it, the resulting gains isolate the importance of the representation available *before* temporal modeling. The advantage is particularly pronounced for sMCI/pMCI, where disease-related structural differences are subtler and longitudinal information is more consequential. SBT$_{\text{SL}}$ and SBT$_{\text{PL}}$ both perform strongly across cohorts, suggesting that their shared anatomy-preserving representation is effective for longitudinal modeling. We additionally train SBT$_{\text{SL}}$ end-to-end, jointly optimizing the Parcel Encoder and longitudinal model. End-to-end training remains competitive, confirming that the SBT advantage is not dependent on freezing the representation; however, the staged formulation is more consistent across tasks while enabling efficient longitudinal training over cached parcel representations. Further results, including the synthetic data experiments, are presented in Appendix Sec D.

4.3 From Diagnosis to Progression and Prognosis

The sMCI/pMCI task evaluates whether the representation can distinguish individuals who subsequently progress to AD. To explicitly model *when* progression occurs, we perform landmark-based survival analysis using only examinations observed up to each landmark (Fig. 4). At each landmark, subjects already converted are excluded; Cox proportional hazards model is fitted on frozen representations, with later conversions as events and non-converters censored at last follow-up. The Kaplan–Meier analyses show clear separation between predicted risk groups, while the time-dependent AUC results demonstrate discrimination of future conversion across future prediction horizons. More importantly, SBT's C-index improves as the landmark advances and additional patient history becomes available, while competing baselines derive substantially less benefit from the added history. The increasing margin at later landmarks provides evidence that the value of the SBT representation grows as longitudinal disease information accumulates, extending the benefit from diagnosis and conversion classification to prognosis. Further details are in Appendix Sec. D.3.

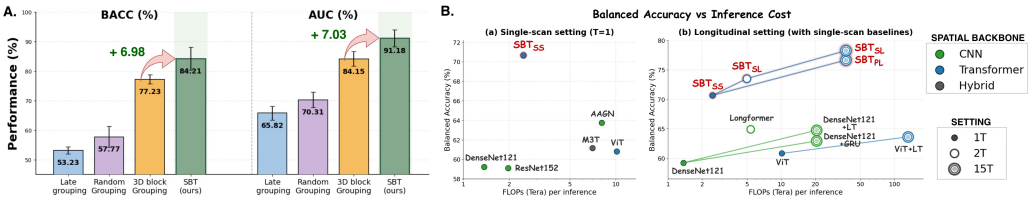

Figure 5: **A.** SBT_{SS} performance against random grouping, regular spatial blocks, and anatomical grouping, together with late introduction of anatomical structure. **B.** Balanced accuracy versus inference cost under single and longitudinal settings for the difficult task of sMCI vs pMCI.

4.4 Generalization Beyond the Development Setting

We next examine whether the representation principle persists under changes in cohort shifts.

Cross-cohort generalization. ADNI-trained checkpoints are directly evaluated on the independent AIBL and MIRIAD cohorts without target-cohort optimization. As shown in Table 1, SBT retains strong performance under this shift, indicating that the learned representation transfers beyond the acquisition population used for development. The corresponding fine-tuned results further show that the same representation can be effectively adapted when target-cohort data are available.

Alternative anatomical parcellation. We repeat our representation learning with the Neuromorphometrics atlas to test whether SBT depends on a specific parcel configuration. Despite changing the anatomical parcellation, the longitudinal advantage of SBT is preserved across both AD/CN and sMCI/pMCI tasks (Table 2). The framework, therefore, benefits from anatomically structured representation formation without relying on a specific atlas.

Table 2: Performance using the Neuromorphometrics atlas with 136 parcels.

Model	ADNI [AD/CN]		ADNI [sMCI/pMCI]	
	BAcc(%)	AUC(%)	BAcc(%)	AUC(%)
3D DenseNet+GRU	$79.37_{(3.64)}$	$87.64_{(2.03)}$	$62.93_{(1.51)}$	$75.57_{(1.06)}$
3D DenseNet+LT	$79.12_{(3.55)}$	$87.62_{(2.17)}$	$64.75_{(2.43)}$	$73.04_{(1.82)}$
3D ViT+LT	$80.97_{(3.49)}$	$87.98_{(1.55)}$	$63.63_{(2.26)}$	$73.31_{(1.09)}$
SBT_{SS}	$80.23_{(3.38)}$	$84.14_{(3.17)}$	$62.11_{(1.45)}$	$75.67_{(1.45)}$
SBT_{SL}	$\mathbf{84.56}_{(3.83)}$	$\mathbf{89.35}_{(3.24)}$	$\mathbf{73.02}_{(3.07)}$	$\mathbf{81.04}_{(3.66)}$
SBT_{PL}	$83.98_{(3.04)}$	$89.04_{(3.12)}$	$72.24_{(3.33)}$	$80.08_{(3.51)}$
Min gain over SBT_{SS}	$+\mathbf{4.33}$	$+\mathbf{5.21}$	$+\mathbf{10.13}$	$+\mathbf{5.37}$

Beyond the AD spectrum. Finally, NIFD provides a disease setting with a structural degeneration pattern distinct from AD. SBT remains effective for FTD/CN classification (Table 1), supporting the applicability of the representation principle beyond AD-specific anatomical patterns.

4.5 Architectural Ablations

Additional architectural ablations are summarized in Table 3. In the single-scan setting, learned parcel scoring provides a clear advantage over uniform parcel aggregation. To prevent over-reliance on a small subset of regions, parcel-dropout is introduced as a regularizer during training. A moderate parcel dropout further improves the resulting representation. In longitudinal setting, replacing Transformer with GRU preserves most of performance, and variations in pooling and Transformer depth produce comparatively smaller changes. These results suggest that the principal gains are not tied to one specific temporal module; the anatomically structured representation formed before temporal modeling remains the dominant design factor.

Table 3: Ablation of SBT framework on ADNI AD/CN task.

Model Configuration	BAcc (%)	AUC (%)
Single-scan setting		
3D ViT with CLS	$78.06_{(3.56)}$	$83.63_{(2.61)}$
3D ViT with mean token pool	$78.54_{(1.51)}$	$84.62_{(2.58)}$
PE with parcel CLS + Mean parcel	$73.54_{(2.89)}$	$78.70_{(5.63)}$
PE with GeM pool ($p = 1$) + Mean parcel	$74.41_{(3.01)}$	$77.34_{(5.25)}$
PE with GeM pool ($p = 2$) + Mean parcel	$76.01_{(2.56)}$	$81.43_{(3.27)}$
PE with GeM pool ($p = 1$) + Parcel scorer	$81.42_{(4.32)}$	$89.85_{(3.00)}$
PE with GeM pool ($p = 2$) + Parcel scorer	$83.15_{(2.14)}$	$90.16_{(2.15)}$
PE with GeM pool ($p = 2$) + Parcel scorer with Parcel dropout ($\rho = 0.10$)	$\mathbf{84.21}_{(3.84)}$	$\mathbf{91.18}_{(2.86)}$
Longitudinal modeling		
3D ViT + Longitudinal Transformer	$80.97_{(2.49)}$	$87.98_{(1.55)}$
SBT-SL + GRU	$88.65_{(2.17)}$	$94.19_{(1.90)}$
SBT-SL + Transformer with CLS token	$88.31_{(2.93)}$	$94.09_{(2.34)}$
SBT-SL + Transformer ($b = 1$) + mean-pool	$87.95_{(2.23)}$	$93.61_{(2.44)}$
SBT-SL + Transformer ($b = 2$) + mean-pool	$\mathbf{89.20}_{(2.21)}$	$\mathbf{94.84}_{(1.66)}$
SBT-SL + Transformer ($b = 3$) + mean-pool	$88.05_{(2.49)}$	$94.04_{(2.05)}$
SBT-PL + GRU	$87.10_{(2.94)}$	$93.42_{(2.77)}$
SBT-PL + Longitudinal Transformer	$88.27_{(3.80)}$	$94.58_{(1.57)}$

4.6 What Does Anatomical Structure Contribute?

We next test whether SBT benefits specifically from anatomically meaningful organization, rather than simply decomposing the volumetric token sequence into smaller spatial contexts.

Is grouping alone sufficient? We replace anatomical supports with matched random groups and regular 3D spatial blocks while keeping the representation architecture unchanged. As shown in Fig. 5A, random grouping performs substantially worse, while regular spatial blocks recover much of the lost performance. Anatomical grouping provides a further clear improvement. Organizing contextualization around semantically meaningful anatomical structure provides a stronger inductive bias for representation learning.

When should anatomical structure be introduced? We next vary the stage at which anatomy enters representation formation. In the late-grouping alternative, tokens first undergo unrestricted global contextualization and are grouped anatomically only afterward. Despite using the same anatomical information for aggregation, late grouping performs markedly worse than introducing anatomy before contextualization (Fig. 5A). This distinction is central to SBT: anatomical structure is most effective when it constrains *how* representations are formed, preserving meaningful regional context before broader feature interaction.

Do anatomical boundaries matter? Finally, we examine the probabilistic atlas memberships used to construct local contexts. Removing the weakest memberships has little effect, whereas stronger pruning or hard argmax membership substantially reduces performance (Table 4). The result indicates that the benefit does not arise merely from increasing token overlap: very weak associations are largely redundant, while retaining meaningful correspondence near anatomical boundaries preserves useful regional context.

Table 4: Effect of progressively removing low-probability or using hard argmax parcel membership on ADNI AD/CN task.

Membership	Removed (%)	BAcc (%)	AUC (%)
Probabilistic	0	$84.21_{(3.84)}$	$91.18_{(2.86)}$
Probabilistic	10	$84.21_{(3.84)}$	$91.18_{(2.86)}$
Probabilistic	20	$84.21_{(3.84)}$	$91.18_{(2.86)}$
Probabilistic	30	$75.82_{(3.99)}$	$84.49_{(2.98)}$
Probabilistic	50	$68.85_{(2.96)}$	$78.51_{(2.99)}$
Probabilistic	90	$52.00_{(1.41)}$	$60.88_{(2.98)}$
Hard argmax	–	$72.76_{(3.66)}$	$85.10_{(3.22)}$

Together, these experiments show that SBT depends on *which* contexts are formed, *when* structure is introduced, and *how* uncertain anatomical support is represented.

4.7 How Does Structure Interact with Longitudinal History?

Having established the role of anatomical structure within an examination, we next ask whether the resulting representation changes *how effectively a model uses longitudinal information*.

Accumulating longitudinal history. Figure 6 compares SBT_{SL} with the matched ViT3D+LT baseline as the available history window increases. The baseline representation exhibits comparatively limited improvement as longer histories become available, whereas SBT progressively benefits from additional longitudinal context before eventually saturating. Consequently, the performance gap widens with increasing history, indicating that anatomically structured representations provide a more effective substrate for integrating information across time.

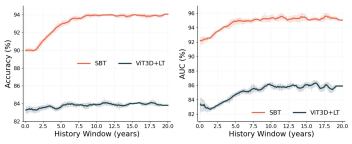

Figure 6: Performance improvements with longitudinal data on ADNI [AD/CN] as a function of the history window (0→20y).

Accuracy-cost trade-off. The inference-cost analysis in Fig. 5B shows that the predictive advantage of SBT is not obtained just by increasing scans. SBT occupies a favorable accuracy–cost regime in both single-scan and longitudinal settings, reflecting the benefit of performing high-resolution contextualization locally while restricting longitudinal interaction to compact structured representations.

Role of temporal information. Table 5 separates the contribution of image representation from acquisition timing. Age alone is insufficient to explain SBT performance, while MRI representation remains strong even without an explicit temporal encoding. Adding either relative elapsed time or absolute acquisition age further improves performance. Thus, primary predictive signal arises from image representation, while acquisition timing provides complementary longitudinal information.

Table 5: Ablation of temporal encoding on ADNI, AD/CN task.

Model	Age encoding	BAcc (%)	AUC (%)
Age-alone	Age features only	$57.01_{(0.93)}$	$75.99_{(2.30)}$
SBT_{SL}	None	$86.46_{(2.61)}$	$90.15_{(2.55)}$
SBT_{SL}	Relative time	$88.98_{(3.42)}$	$94.20_{(2.36)}$
SBT_{SL}	Absolute age	$89.20_{(2.21)}$	$94.84_{(1.66)}$
SBT_{PL}	None	$87.18_{(3.41)}$	$92.41_{(3.41)}$
SBT_{PL}	Relative time	$88.16_{(3.19)}$	$93.98_{(2.86)}$
SBT_{PL}	Absolute age	$88.27_{(3.80)}$	$94.58_{(1.57)}$

4.8 Does SBT Preserve Localized Disease Evidence?

A central premise of SBT is that preserving anatomical organization during representation formation should allow localized structural abnormalities to remain identifiable after subsequent spatial and temporal aggregation. We therefore first evaluate parcel-level evidence under controlled ground truth, and then examine longitudinal patterns in clinical MRI. Additional subjects & cohort-level analysis, and discussion on synthetic dataset for SBT_{SL} vs SBT_{PL} are provided in Appendix Sec. E.

Controlled regional abnormalities. The synthetic setting provides explicit ground truth for both the affected parcel and abnormality severity. SBT_{SL} reliably identifies the atrophied regions, while the learned parcel weights closely track the underlying severity (Fig. 8).

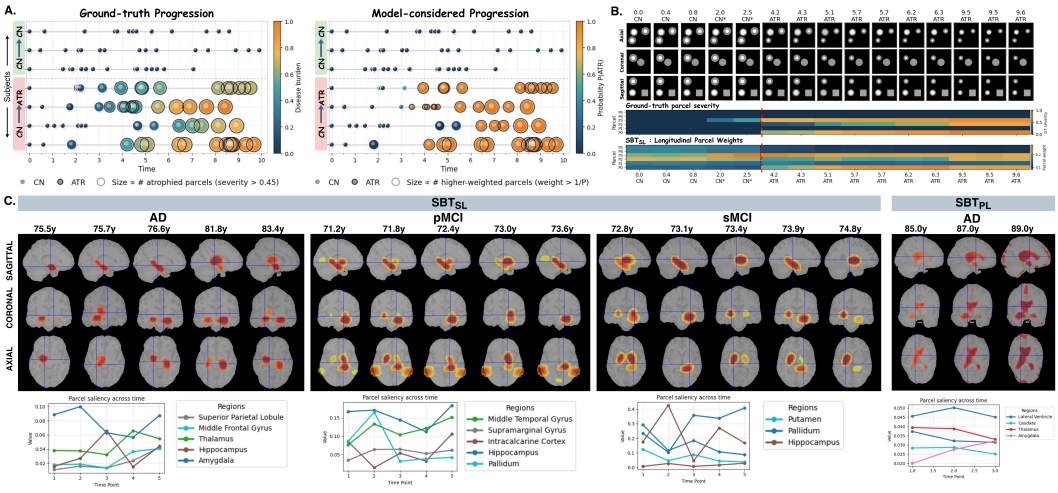

Figure 7: **SBT preserves localized evidence as abnormalities evolve longitudinally. A.** Controlled synthetic trajectories, with each row denoting one subject across visits. In the ground-truth panel, color indicates disease burden, and blob size indicates #atrophied parcels. In the model panel, color denotes $P(\mathrm{ATR} \mid x_{1:t})$ using all scans observed up to visit t, while blob size indicates #higher-weighted parcels. **B.** A representative 3D synthetic sequence compares ground-truth parcel severity with $\mathrm{SBT}_{\mathrm{SL}}$ parcel weights across visits. **C.** Clinical examples visualize attention rollouts and corresponding longitudinal parcel-saliencies for AD, pMCI, and sMCI subjects.

Figure 7 A,B further compares known synthetic progression with the regions emphasized by $\mathrm{SBT}_{\mathrm{SL}}$ over time. Across controlled trajectories, the model-considered disease progression follows increasing ground-truth abnormality burden, while the longitudinal parcel weights shift toward the parcels in which abnormalities emerge and accumulate. This shows that the learned parcel weighting is aligned with the prediction as well as the spatial source and magnitude of the underlying abnormality. SBT therefore retains localized evidence throughout longitudinal aggregation.

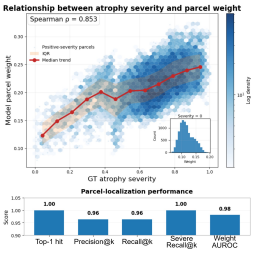

Figure 8: Stats of parcel weight assignment on synthetic test data.

Disease-related interpretations (Fig. 7C). We finally examine attention rollouts (Abnar & Zuidema (2020)) and parcel saliencies for sample AD, pMCI, and sMCI histories (details in Appendix Sec E). The attention maps remain spatially localized across visits, while parcel saliencies expose the most contributing parcels across time. Several highly emphasized regions overlap with well-established structural patterns in AD and MCI (Planche et al. (2022); Braak et al. (2006); Whitwell et al. (2008); Jack Jr et al. (2000); Poulin et al. (2011); Byun et al. (2015); Fan et al. (2008)), including hippocampal & amygdalar involvement, temporal & parietal regions, and ventricular changes associated with progressive neurodegeneration. The clinical examples, therefore, provide qualitative anatomical concordance with disease-pattern findings in the clinical literature.

5 Discussion and Conclusion

We introduced **Structure Before Time (SBT)**, an anatomy-preserving representation principle for longitudinal 3D brain MRI in which anatomical structure shapes representation formation before temporal aggregation. Across AD & FTD diagnosis, MCI progression, and survival prediction, SBT consistently benefits from meaningful anatomical organization, exploits increasing longitudinal history more effectively than global representations, and retains localized disease evidence. These findings suggest that the representation formed *before* temporal modeling is an important determinant of effective longitudinal modeling. The present study is limited to T1w structural MRI. Future work can extend SBT to complementary modalities such as T2w MRI, FLAIR, or multimodal imaging and evaluate whether the same representation principle generalizes to other longitudinal neurological conditions. Broadly, SBT motivates structure-preserving representation learning when predictive signals evolve within persistent substructures of high-dimensional observations.

AI USE STATEMENT

In this work, we used generative AI tools solely to assist with grammatical correction, language refinement, and improving the readability of the manuscript. We did not use generative AI tools to develop the research methodology, design experiments, implement methods, process or analyze data, interpret results, formulate scientific claims, generate figures, or conduct literature review. All AI-assisted edits were reviewed and verified by the authors. We take full responsibility for the final content of this work, including all text, claims, results, and conclusions.

REPRODUCIBILITY STATEMENT

To support reproducibility, comprehensive descriptions of model architectures, training configurations, and evaluation protocols are provided in the main text and appendix. The dataset details, evaluation protocol, and preprocessing pipeline are discussed in Sec 4.1 and Appendix Sec B. The synthetic dataset generation details are provided in Appendix Sec B.4. The model architecture and training details are provided in Appendix Sec C.2.

REFERENCES

Neuromorphometrics. http://neuromorphometrics.com. RRID:SCR_005656.

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4190–4197, 2020.

Sepehr Aghajanian, Fateme Mohammadifard, Ida Mohammadi, Shahryar Rajai Firouzabadi, Ali Baradaran Bagheri, Elham Moases Ghaffary, and Omid Mirmosayyeb. Longitudinal structural mri-based deep learning and radiomics features for predicting alzheimer's disease progression. *Alzheimer's Research & Therapy*, 17(1), 2025.

Fatih Altay, Guillermo Ramón Sánchez, Yanli James, Stephen V Faraone, Senem Velipasalar, and Asif Salekin. Preclinical stage alzheimer's disease detection using magnetic resonance image scans. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 15088–15097, 2021.

Liana G Apostolova, Amity E Green, Sona Babakchanian, Kristy S Hwang, Yi-Yu Chou, Arthur W Toga, and Paul M Thompson. Hippocampal atrophy and ventricular enlargement in normal aging, mild cognitive impairment and alzheimer's disease. *Alzheimer disease and associated disorders*, 26(1):17, 2012.

Josephine Barnes, Rachael I Scahill, Jonathan M Schott, Chris Frost, Martin N Rossor, and Nick C Fox. Does alzheimer's disease affect hippocampal asymmetry? evidence from a cross-sectional and longitudinal volumetric mri study. *Dementia and geriatric cognitive disorders*, 19(5-6):338–344, 2005.

Randall J Bateman, Chengjie Xiong, Tammie LS Benzinger, et al. Clinical and biomarker changes in dominantly inherited alzheimer's disease. *New England Journal of Medicine*, 367(9):795–804, 2012.

Yana Blinkouskaya and Johannes Weickenmeier. Brain shape changes associated with cerebral atrophy in healthy aging and alzheimer's disease. *Frontiers in Mechanical Engineering*, 7, 2021.

Heiko Braak and Eva Braak. Neuropathological stageing of alzheimer-related changes. *Acta neuropathologica*, 82(4):239–259, 1991.

Heiko Braak, Irina Alafuzoff, Thomas Arzberger, Hans Kretzschmar, and Kelly Del Tredici. Staging of alzheimer disease-associated neurofibrillary pathology using paraffin sections and immunocytochemistry. *Acta neuropathologica*, 112(4):389–404, 2006.

Randy L Buckner, Jessica R Andrews-Hanna, and Daniel L Schacter. The brain's default network: anatomy, function, and relevance to disease. *Annals of the New York Academy of Sciences*, 1124 (1):1–38, 2008.

Min Soo Byun, Song E Kim, Jinsick Park, Dahyun Yi, et al. Heterogeneity of regional brain atrophy patterns associated with distinct progression rates in alzheimer's disease. *PLoS One*, 10(11), 2015.

Qiuhui Chen, Qiang Fu, Hao Bai, and Yi Hong. Longformer: Longitudinal transformer for alzheimer's disease classification with structural mris. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3563–3572. IEEE, 2024.

Gaël Chételat, Brigitte Landeau, Francis Eustache, Florence Mézenge, Fausto Viader, Vincent de La Sayette, Béatrice Desgranges, and J-C Baron. Using voxel-based morphometry to map the structural changes associated with rapid conversion in mci: a longitudinal mri study. *Neuroimage*, 27(4):934–946, 2005.

Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, eighth workshop on syntax, semantics and structure in statistical translation*, pp. 103–111, 2014.

David R Cox. Regression models and life-tables. *Journal of the royal statistical society: Series B (methodological)*, 34(2):187–202, 1972.

Ruoxuan Cui, Manhua Liu, Alzheimer's Disease Neuroimaging Initiative, et al. Rnn-based longitudinal analysis for diagnosis of alzheimer's disease. *Computerized Medical Imaging and Graphics*, 73:1–10, 2019.

Christos Davatzikos, Feng Xu, Yang An, Yong Fan, and Susan M Resnick. Longitudinal progression of alzheimer's-like patterns of atrophy in normal older adults: the spare-ad index. *Brain*, 132(8): 2026–2035, 2009.

Laura W De Jong, Karin van der Hiele, Ilya M Veer, JJ Houwing, RGJ Westendorp, et al. Strongly reduced volumes of putamen and thalamus in alzheimer's disease: an mri study. *Brain*, 131(12): 3277–3285, 2008.

Michael A DeTure and Dennis W Dickson. The neuropathological diagnosis of alzheimer's disease. *Molecular neurodegeneration*, 14(1), 2019.

Bradford C Dickerson, Akram Bakkour, David H Salat, et al. The cortical signature of alzheimer's disease: regionally specific cortical thinning relates to symptom severity in very mild to mild ad dementia and is detectable in asymptomatic amyloid-positive individuals. *Cerebral cortex*, 19(3): 497–510, 2009.

Rachelle S Doody, Valory Pavlik, Paul Massman, Susan Rountree, Eveleen Darby, and Wenyaw Chan. Predicting progression of alzheimer's disease. *Alzheimer's research & therapy*, 2(1), 2010.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

Yong Fan, Nematollah Batmanghelich, Chris M Clark, Christos Davatzikos, Alzheimer's Disease Neuroimaging Initiative, et al. Spatial patterns of brain atrophy in mci patients, identified via high-dimensional pattern classification, predict subsequent cognitive decline. *Neuroimage*, 39(4): 1731–1743, 2008.

Ali Farzan, Syansiah Mashohor, Abd Rahman Ramli, and Rozi Mahmud. Boosting diagnosis accuracy of alzheimer's disease using high dimensional recognition of longitudinal brain atrophy patterns. *Behavioural brain research*, 290:124–130, 2015.

Daniel Ferreira, Chloë Verhagen, Juan Andrés Hernández-Cabrera, Lena Cavallin, Chun-Jie Guo, Urban Ekman, J-Sebastian Muehlboeck, Andrew Simmons, José Barroso, Lars-Olof Wahlund, et al. Distinct subtypes of alzheimer's disease based on patterns of brain atrophy: longitudinal trajectories and clinical applications. *Scientific reports*, 7(1), 2017.

FMRIB Software Library. Harvard-oxford cortical structural atlas. `http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/Atlases`. RRID:SCR_001476.

Guilherme Folego, Marina Weiler, Raphael F Casseb, Ramon Pires, and Anderson Rocha. Alzheimer's disease detection through whole-brain 3d-cnn mri. *Frontiers in bioengineering and biotechnology*, 8, 2020.

Christopher Fowler, Stephanie R Rainey-Smith, Sabine Bird, et al. Fifteen years of the australian imaging, biomarkers and lifestyle (aibl) study: progress and observations from 2,359 older adults spanning the spectrum from cognitive normality to alzheimer's disease. *Journal of Alzheimer's disease reports*, 5(1):443–468, 2021.

Giovanni B Frisoni, Nick C Fox, Clifford R Jack Jr, Philip Scheltens, and Paul M Thompson. The clinical use of structural mri in alzheimer disease. *Nature reviews neurology*, 6(2):67–77, 2010.

Xingyu Gao, Hongjie Cai, and Manhua Liu. A hybrid multi-scale attention convolution and aging transformer network for alzheimer's disease diagnosis. *IEEE Journal of Biomedical and Health Informatics*, 27(7):3292–3301, 2023.

Hao Guan, Ling Yue, Pew-Thian Yap, Shifu Xiao, Andrea Bozoki, and Mingxia Liu. Attention-guided autoencoder for automated progression prediction of subjective cognitive decline with structural mri. *IEEE journal of biomedical and health informatics*, 27(6):2980–2989, 2023.

Frank E Harrell, Robert M Califf, David B Pryor, Kerry L Lee, and Robert A Rosati. Evaluating the yield of medical tests. *Jama*, 247(18):2543–2546, 1982.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

WJP Henneman, JD Sluimer, J Barnes, WM Van Der Flier, IC Sluimer, NC Fox, Ph Scheltens, H Vrenken, and F Barkhof. Hippocampal atrophy rates in alzheimer disease: added value over whole brain volume measures. *Neurology*, 72(11):999–1007, 2009.

Andrew Hoopes, Jocelyn S Mora, et al. Synthstrip: skull-stripping for any brain image. *NeuroImage*, 260:119474, 2022.

Zhentao Hu, Zheng Wang, Yong Jin, and Wei Hou. Vgg-tswinformer: Transformer-based deep learning model for early alzheimer's disease prediction. *Computer Methods and Programs in Biomedicine*, 229, 2023.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 2261–2269. Ieee, 2017a.

Meiyan Huang, Wei Yang, Qianjin Feng, and Wufan Chen. Longitudinal measurement and hierarchical classification framework for the prediction of alzheimer's disease. *Scientific reports*, 7(1), 2017b.

Clifford R Jack, David S Knopman, William J Jagust, Ronald C Petersen, Michael W Weiner, Paul S Aisen, Leslie M Shaw, Prashanthi Vemuri, Heather J Wiste, Stephen D Weigand, et al. Tracking pathophysiological processes in alzheimer's disease: an updated hypothetical model of dynamic biomarkers. *The lancet neurology*, 12(2):207–216, 2013.

Clifford R Jack Jr, Ronald C Petersen, Yue Cheng Xu, Peter C O'Brien, Glenn E Smith, Robert J Ivnik, Bradley F Boeve, Stephen C Waring, Eric G Tangalos, and Emre Kokmen. Prediction of ad with mri-based hippocampal volume in mild cognitive impairment. *Neurology*, 52(7):1397–1397, 1999.

Clifford R Jack Jr, Ronald Carl Petersen, Y Xu, PC O'brien, Glenn E Smith, Robert J Ivnik, Bradley F Boeve, Eric George Tangalos, and E Kokmen. Rates of hippocampal atrophy correlate with change in clinical status in aging and ad. *Neurology*, 55(4):484–490, 2000.

Clifford R Jack Jr, Matt A Bernstein, Nick C Fox, Paul Thompson, et al. The alzheimer's disease neuroimaging initiative (adni): Mri methods. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 27(4):685–691, 2008.

Clifford R Jack Jr, David A Bennett, Kaj Blennow, Maria C Carrillo, Billy Dunn, Samantha Budd Haeberlein, David M Holtzman, William Jagust, Frank Jessen, Jason Karlawish, et al. Nia-aa research framework: toward a biological definition of alzheimer's disease. *Alzheimer's & dementia*, 14(4):535–562, 2018.

Jinseong Jang and Dosik Hwang. M3t: three-dimensional medical image classifier using multi-plane and multi-slice transformer. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20686–20697. IEEE, 2022.

Kurt A Jellinger. Recent update on the heterogeneity of the alzheimer's disease spectrum. *Journal of Neural Transmission*, 129(1):1–24, 2022.

Hongchao Jiang and Chunyan Miao. Anatomy-aware gating network for explainable alzheimer's disease diagnosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 90–100. Springer, 2024.

Alireza Jomeiri, Ahmad Habibizad Navin, and Mahboubeh Shamsi. Longitudinal mri analysis using a hybrid densenet-bilstm method for alzheimer's disease prediction. *Behavioural Brain Research*, 463, 2024.

Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.

Uttam Khatri, Jun-Hyung Kim, and Goo-Rak Kwon. Alzheimer's disease and mild cognitive impairment detection using smri with efficient receptive field and enhanced multi-axis attention fusion. *IEEE Access*, 12:100848–100861, 2024.

Sergey Korolev, Amir Safiullin, Mikhail Belyaev, and Yulia Dodonova. Residual and plain convolutional neural networks for 3d brain mri classification. In *2017 IEEE 14th international symposium on biomedical imaging (ISBI 2017)*, pp. 835–838. IEEE, 2017.

William H Kruskal and W Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621, 1952.

Kassymzhomart Kunanbayev, Vyacheslav Shen, and Dae-Shik Kim. Training vit with limited data for alzheimer's disease classification: An empirical study. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 334–343. Springer, 2024.

Benjamin Lam, Mario Masellis, Morris Freedman, Donald T Stuss, and Sandra E Black. Clinical, imaging, and pathological heterogeneity of the alzheimer's disease syndrome. *Alzheimer's research & therapy*, 5(1), 2013.

Chao Li, Yue Cui, Na Luo, Yong Liu, Pierrick Bourgeat, Jurgen Fripp, and Tianzi Jiang. Transresnet: Integrating transformers and cnns for alzheimer's disease classification. In *2022 IEEE 19th international symposium on biomedical imaging (ISBI)*, pp. 1–5. IEEE, 2022.

Hongming Li, Mohamad Habes, David A Wolk, Yong Fan, Alzheimer's Disease Neuroimaging Initiative, the Australian Imaging Biomarkers, and Lifestyle Study of Aging. A deep learning model for early prediction of alzheimer's disease dementia based on hippocampal magnetic resonance imaging data. *Alzheimer's & Dementia*, 15(8):1059–1070, 2019.

Xingfeng Li, Damien Coyle, Liam Maguire, David R Watson, and Thomas M McGinnity. Gray matter concentration and effective connectivity changes in alzheimer's disease: a longitudinal structural mri study. *Neuroradiology*, 53(10):733–748, 2011.

Chunfeng Lian, Mingxia Liu, Jun Zhang, and Dinggang Shen. Hierarchical fully convolutional network for joint atrophy localization and alzheimer's disease diagnosis using structural mri. *IEEE transactions on pattern analysis and machine intelligence*, 42(4):880–893, 2018.

Jinwei Liu, Yashu Xu, Yi Liu, Huating Luo, Wenxiang Huang, and Lizhong Yao. Attention-guided 3d cnn with lesion feature selection for early alzheimer's disease prediction using longitudinal smri. *IEEE Journal of Biomedical and Health Informatics*, 29(1):324–332, 2024.

Si-Yuan Lu, Yu-Dong Zhang, and Yu-Dong Yao. An efficient vision transformer for alzheimer's disease classification using magnetic resonance images. *Biomedical Signal Processing and Control*, 101, 2025.

Ian B Malone, David Cash, Gerard R Ridgway, David G MacManus, Sebastien Ourselin, Nick C Fox, and Jonathan M Schott. Miriad—public release of a multiple time point alzheimer's mr imaging dataset. *NeuroImage*, 70:33–36, 2013.

John C Mazziotta, Arthur W Toga, Alan Evans, Peter Fox, and Jack Lancaster. A probabilistic atlas of the human brain: Theory and rationale for its development: The international consortium for brain mapping (icbm). *Neuroimage*, 2(2):89–101, 1995.

Guy M McKhann, David S Knopman, Howard Chertkow, Bradley T Hyman, Clifford R Jack Jr, Claudia H Kawas, William E Klunk, Walter J Koroshetz, Jennifer J Manly, Richard Mayeux, et al. The diagnosis of dementia due to alzheimer's disease: recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease. *Alzheimer's & dementia*, 7(3):263–269, 2011.

Sean M Nestor, Raul Rupsingh, Michael Borrie, Matthew Smith, Vittorio Accomazzi, Jennie L Wells, Jennifer Fogarty, Robert Bartha, and Alzheimer's Disease Neuroimaging Initiative. Ventricular enlargement as a possible measure of alzheimer's disease progression validated using the alzheimer's disease neuroimaging initiative database. *Brain*, 131(9):2443–2454, 2008.

Kanghan Oh, Young-Chul Chung, Ko Woon Kim, Woo-Sung Kim, and Il-Seok Oh. Classification and visualization of alzheimer's disease using volumetric convolutional neural network and transfer learning. *Scientific reports*, 9(1), 2019.

Mubarak Olaoluwa, Heni Loukil, Arafet Sbei, and Hassen Drira. Geohsaf: Geometric hippocampus shape analysis framework for longitudinal alzheimer's disease classification. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2156–2167. IEEE, 2026.

Jiahong Ouyang, Qingyu Zhao, Ehsan Adeli, Edith V Sullivan, Adolf Pfefferbaum, Greg Zaharchuk, and Kilian M Pohl. Self-supervised longitudinal neighbourhood embedding. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 80–89. Springer, 2021.

Jiahong Ouyang, Qingyu Zhao, Ehsan Adeli, Greg Zaharchuk, and Kilian M Pohl. Disentangling normal aging from severity of disease via weak supervision on longitudinal mri. *IEEE transactions on medical imaging*, 41(10):2558–2569, 2022.

Jiahong Ouyang, Qingyu Zhao, Ehsan Adeli, Wei Peng, Greg Zaharchuk, and Kilian M Pohl. Lsor: Longitudinally-consistent self-organized representation learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 279–289. Springer, 2023.

Corina Pennanen, Miia Kivipelto, Susanna Tuomainen, Päivi Hartikainen, Tuomo Hänninen, Mikko P Laakso, Merja Hallikainen, Matti Vanhanen, Aulikki Nissinen, Eeva-Liisa Helkala, et al. Hippocampus and entorhinal cortex in mild cognitive impairment and early ad. *Neurobiology of aging*, 25(3):303–310, 2004.

Ronald C Petersen. Mild cognitive impairment as a diagnostic entity. *Journal of internal medicine*, 256(3):183–194, 2004.

Vincent Planche, José V Manjon, Boris Mansencal, Enrique Lanuza, Thomas Tourdias, Gwenaëlle Catheline, and Pierrick Coupé. Structural progression of alzheimer's disease over decades: the mri staging scheme. *Brain communications*, 4(3), 2022.

Stéphane P Poulin, Rebecca Dautoff, John C Morris, Lisa Feldman Barrett, Bradford C Dickerson, Alzheimer's Disease Neuroimaging Initiative, et al. Amygdala atrophy is prominent in early alzheimer's disease and relates to symptom severity. *Psychiatry Research: Neuroimaging*, 194 (1):7–13, 2011.

Anqi Qiu, Liyuan Xu, Chaoqiang Liu, Alzheimer's Disease Neuroimaging Initiative, et al. Predicting diagnosis 4 years prior to alzheimer's disease incident. *NeuroImage: Clinical*, 34, 2022.

Shangran Qiu, Prajakta S Joshi, Matthew I Miller, et al. Development and validation of an interpretable deep learning framework for alzheimer's disease classification. *Brain*, 143(6):1920–1933, 2020.

Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning cnn image retrieval with no human annotation. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1655–1668, 2018.

Boris-Stephan Rauchmann, Ersin Ersözlü, Dorothea Luedecke, Nicolai Franzmeier, and Robert Perneczky. Multimodal and longitudinal characterization of distinct tau and atrophy clusters in alzheimer's disease spectrum. *Scientific reports*, 15(1), 2025.

Howard J Rosen, AL Boxer, ML Gorno-Tempini, and MW Weiner. Frontotemporal lobar degeneration neuroimaging initiative (ftldni). *Funded by the National Institute on Aging*, 2010.

Juan Ruiz, Mufti Mahmud, Md Modasshir, M Shamim Kaiser, and for the Alzheimer's Disease Neuroimaging Initiative. 3d densenet ensemble in 4-way classification of alzheimer's disease. In *International conference on brain informatics*, pp. 85–96. Springer, 2020.

Ignacio Sarasua, Sebastian Pölsterl, and Christian Wachinger. Hippocampal representations for deep learning on alzheimer's disease. *Scientific reports*, 12(1):8619, 2022.

Yaakov Stern. Cognitive reserve in ageing and alzheimer's disease. *The Lancet Neurology*, 11(11): 1006–1012, 2012.

Prashanthi Vemuri and Clifford R Jack Jr. Role of structural mri in alzheimer's disease. *Alzheimer's research & therapy*, 2(4), 2010.

Ying Wang, Susan M Resnick, Christos Davatzikos, Baltimore Longitudinal Study of Aging, and the Alzheimer's Disease Neuroimaging Initiative. Analysis of spatio-temporal brain imaging patterns by hidden markov models and serial mri images. *Human brain mapping*, 35(9):4777–4794, 2014.

Jennifer L Whitwell, Ronald C Petersen, Selamawit Negash, Stephen D Weigand, Kejal Kantarci, Robert J Ivnik, David S Knopman, Bradley F Boeve, Glenn E Smith, and Clifford R Jack. Patterns of atrophy differ among specific subtypes of mild cognitive impairment. *Archives of neurology*, 64(8):1130–1138, 2007.

Jennifer L Whitwell, Maria M Shiung, SA Przybelski, et al. Mri patterns of atrophy associated with progression to ad in amnestic mild cognitive impairment. *Neurology*, 70(7):512–520, 2008.

Yuanchen Wu, Yuan Zhou, Weiming Zeng, Qian Qian, and Miao Song. An attention-based 3d cnn with multi-scale integration block for alzheimer's disease classification. *IEEE Journal of Biomedical and Health Informatics*, 26(11):5665–5673, 2022.

Jie Zhang, Bowen Zheng, Ang Gao, Xin Feng, Dong Liang, and Xiaojing Long. A 3d densely connected convolution neural network with connection-wise attention mechanism for alzheimer's disease classification. *Magnetic Resonance Imaging*, 78:119–126, 2021.

Jun Zhang, Mingxia Liu, Le An, Yaozong Gao, and Dinggang Shen. Alzheimer's disease diagnosis using landmark-based features from longitudinal structural mr images. *IEEE journal of biomedical and health informatics*, 21(6):1607–1616, 2017.

Qingyu Zhao, Zixuan Liu, Ehsan Adeli, and Kilian M Pohl. Longitudinal self-supervised learning. *Medical image analysis*, 71, 2021.

Wenyong Zhu, Liang Sun, Jiashuang Huang, Liangxiu Han, and Daoqiang Zhang. Dual attention multi-instance deep learning for alzheimer's disease diagnosis with structural mri. *IEEE Transactions on Medical Imaging*, 40(9):2354–2366, 2021a.

Yingying Zhu, Minjeong Kim, Xiaofeng Zhu, Daniel Kaufer, Guorong Wu, Alzheimer's Disease Neuroimaging Initiative, et al. Long range early diagnosis of alzheimer's disease using longitudinal mr imaging data. *Medical image analysis*, 67, 2021b.

Appendix

CONTENTS

A Clinical Background and Motivation **16**

B Datasets, Atlas and Preprocessing **17**

B.1 Public Datasets. 17

B.2 Atlas Selection 18

B.3 Preprocessing and Registration Quality 19

B.4 Controlled Synthetic Dataset 19

C Implementation and Baseline Details **22**

C.1 Code 22

C.2 Training setup and Model details 22

C.3 Prior works and Baselines 23

C.3.1 Broader neuroimaging literature 23

C.3.2 Baseline Implementations and Computational Comparison 24

C.4 Additional Longitudinal Variant: SBT_{PL-B} 25

C.5 Why Structure Must Precede Global Contextualization 25

D Additional Experimental Analysis **26**

D.1 Detailed Classification Results 26

D.2 Temporal Encoding under Cohort Shift 26

D.3 Survival analysis 27

D.4 Confounder Analysis 29

D.5 Robustness to preprocessing perturbations. 32

E Explainability Details and Additional Analysis **32**

E.1 Cohort-level analysis 32

E.2 Attention Rollout Construction 34

E.3 Additional Subject-Level Examples and Failure Analysis 34

E.4 SBT_{SL} vs SBT_{PL} on Controlled Progression 36

A CLINICAL BACKGROUND AND MOTIVATION

Neurodegenerative diseases (NDDs) develop progressively rather than through a single pathological event (Apostolova et al. (2012); Barnes et al. (2005); Braak & Braak (1991); Braak et al. (2006)), gradually altering neural circuits and the structural organization of the brain. In Alzheimer's disease (AD), the NIA–AA AT(N) research framework groups biomarkers into β-amyloid (A), pathological tau (T), and neurodegeneration (N) (DeTure & Dickson (2019); Bateman et al. (2012); Jack Jr et al. (2018); Planche et al. (2022)). Our study focuses on the neurodegenerative component because structural MRI (sMRI) can capture associated morphological changes, including regional atrophy, shape alteration, and tissue loss (Vemuri & Jack Jr (2010); Whitwell et al. (2007); Frisoni et al.

Table A1: Cohort statistics of all datasets used in our experiments.

Dataset	#scans	#subjects	#Max visits	Age (mean±std,y)	Gender (M / F)	Field strength (1.5T / 3.0T)	Manufacturer (Siemens/GE/Philips)
ADNI	5,627	1,703 [AD:493,CN:1210]	15	72.93 ± 7.87	734 / 969	429 / 1,411	1,063 / 476 / 289
ADNI(MCI)	5,308	687 [sMCI:305,pMCI:282]	15	74.18 ± 7.44	408 / 279	275 / 512	389 / 258 / 134
AIBL	1,084	599 [AD:104,CN:500]	5	73.36 ± 6.42	255 / 343	107 / 541	599 / 0 / 0
MIRIAD	708	69 [AD:46,CN:23]	9	69.89 ± 6.98	31 / 38	69 / 0	0 / 69 / 0
NIFD	962	332 [FTD:214,CN:118]	7	64.18 ± 8.16	176 / 156	0 / 332	317 / 16 / 0

(2010); Jack Jr et al. (1999)). We specifically study T1-weighted sMRI without complementary imaging modalities or clinical scores, allowing us to isolate the information contained in longitudinal structural imaging.

AD-related structural change follows a spatially heterogeneous progression, with early involvement of medial temporal structures, including the hippocampus, followed by broader cortical involvement (Jack et al. (2013); Braak & Braak (1991); Ferreira et al. (2017); Blinkouskaya & Weickenmeier (2021); Planche et al. (2022); Jellinger (2022); McKhann et al. (2011); Nestor et al. (2008); Rauchmann et al. (2025)). This is particularly relevant in mild cognitive impairment (MCI), where stable and progressive subjects can exhibit overlapping morphology at a single examination (Zhang et al. (2017)). Longitudinal imaging provides additional information by revealing how regional structural changes accumulate across successive examinations (Blinkouskaya & Weickenmeier (2021); Ferreira et al. (2017); Qiu et al. (2022); Apostolova et al. (2012); Planche et al. (2022); Rauchmann et al. (2025)). This motivates representations that retain anatomically localized information while modeling its evolution over time.

B Datasets, Atlas and Preprocessing

B.1 Public Datasets.

Our experimental evaluation spans four longitudinal cohorts: ADNI, AIBL, MIRIAD and NIFD, chosen to collectively stress-test the model across the full spectrum of real-world acquisition variability. The detailed longitudinal statistics for each cohort, including the number of scans per subject, follow-up duration, and their relationship across cohorts, are visualised in Fig. A1.

- **ADNI** (ADNI 1/GO/2/3/4) is a large multi-center study spanning dozens of North American sites (U.S./Canada) with contributions across Siemens, GE, and Philips scanners at 1.5T and 3T, and protocol families such as MPRAGE. It offers broad diagnostic coverage (CN/MCI/AD) with multi-year follow-ups, yielding substantial inter-site and inter-scanner variability in acquisition parameters and subject demographics. As reflected in Fig. A1, the ADNI cohort (full) exhibits a long tail extending to 15 scans, indicating a heterogeneous follow-up density. Crucially, this variability is representative of real-world longitudinal neuroimaging registry data, and SBT(our model) is designed to accommodate such irregular temporal sampling without requiring interpolation or fixed visit schedules.
- **AIBL** (Australian Imaging, Biomarkers and Lifestyle) is a prospective Australian cohort with imaging conducted primarily at multiple sites in *Melbourne* and *Perth*, predominantly on Siemens platforms (1.5T/3T). It provides serial T1 weighted scans with rich clinical and biomarker metadata and comparatively fewer longitudinal visits per subject than ADNI.
- **MIRIAD** (Minimal Interval Resonance Imaging in Alzheimer's Disease) is a single-center cohort from London (UCL), acquired on a fixed scanner/protocol with dense revisit intervals (weeks–months). It enables high sensitivity to within-subject change under no site/scanner variability.
- **NIFD** (Neuroimaging in Frontotemporal Dementia) is a multi-site NIH-funded study acquired across three academic medical centres: the University of California San Francisco (UCSF), Mayo Clinic (Rochester), and Massachusetts General Hospital (MGH). It is used to evaluate SBT beyond the AD spectrum in FTD.

AD/CN/FTD subjects were defined by their baseline diagnosis, with no diagnostic conversion in the retained cohort.

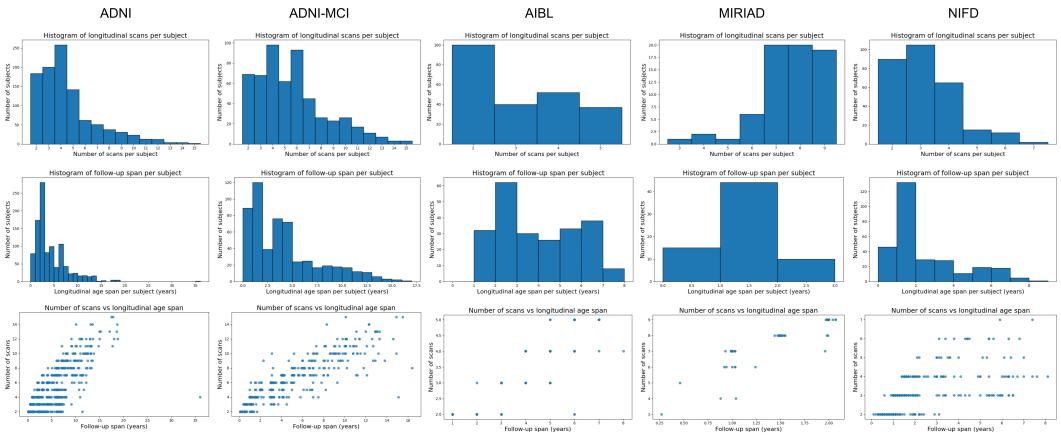

Figure A1: **Longitudinal cohort characteristics.** Distribution of scan counts, follow-up duration, and scan-frequency versus follow-up span across the evaluated cohorts.

ADNI MCI Curation (sMCI/pMCI): We define the baseline as the first visit at which a subject is diagnosed with MCI. Both **stable MCI (sMCI)** and **progressive MCI (pMCI)** subjects are initially diagnosed with MCI at baseline, but the cognitive condition of sMCI group remained stable and did not convert to AD within 36 months after their first visit. In contrast, the pMCI group progressed to a clinical diagnosis of AD during the same 36-month follow-up period (Ouyang et al. (2022)). For pMCI subjects, only scans acquired before the first AD diagnosis are retained, ensuring that all pMCI observations represent the pre-conversion MCI stage. For time-to-event evaluation, we use the full MCI cohort, treating AD conversion as the event and non-converters as right-censored at their last available follow-up.

B.2 Atlas Selection

A brain atlas provides anatomically defined regions called "parcels". A *probabilistic* atlas additionally assigns each voxel a soft membership over regions. In our pipeline, each scan is first registered to MNI space, after which the atlas is further registered to the individual scan to better handle inter-subject variability and boundary uncertainty. For our setting, the atlas must satisfy two requirements: (i) provide sufficient macro-anatomical coverage of structures relevant to neurodegeneration, and (ii) provide volumetric parcellation, unlike the surface-based atlas. Atlas sensitivity must therefore be interpreted with respect to both anatomical coverage and volumetric support; comparing atlases that differ substantially in these properties would confound atlas identity with differences in the anatomical information available to the model.

HO atlas: We use the probabilistic Harvard–Oxford (HO) cortical and subcortical atlases at 1 mm resolution, yielding 69 parcels in our setup. HO provides broad coverage of AD-relevant structures, including the hippocampus, amygdala, parahippocampal regions, posterior cingulate/precuneus, and ventricular structures. Its probabilistic labels allow patch tokens near region boundaries to contribute to multiple anatomical contexts rather than enforcing a hard partition. HO is nevertheless macro-anatomical: the entorhinal and perirhinal cortices are subsumed within the anterior Parahippocampal Gyrus, while hippocampal subfields such as CA1, subiculum, and dentate gyrus are represented within the global Hippocampus parcel. Thus, these finer structures are not individually resolved, but their signal remains incorporated within the broader regional representation.

Neuromorphometrics atlas: To test whether SBT depends on the specific HO definition, we additionally evaluate it using the Neuromorphometrics atlas (neu), a probabilistic atlas, which provides broadly comparable anatomical coverage with a different granularity of parcellation having 136 regions. This provides a more controlled atlas-sensitivity analysis while avoiding comparisons that would introduce additional confounds due to major differences in anatomical coverage or representation.

Table A2: Patch-level registration and parcel coverage QC across clinical groups. Values are reported as mean ± standard deviation, with medians shown in parentheses where relevant. All groups retained nearly complete parcel coverage, indicating that downstream disease/progression effects are unlikely to be driven by group-specific registration failure.

Group	N	Non-empty parcels	Mean parcel coverage	Low-coverage parcels
CN	4142	69.00 ± 0.00 (69)	0.9925 ± 0.0177 (0.9958)	0.018 ± 1.075 (0)
AD	1485	69.00 ± 0.00 (69)	0.9869 ± 0.0122 (0.9899)	0.010 ± 0.170 (0)
sMCI	3308	69.00 ± 0.00 (69)	0.9913 ± 0.0088 (0.9940)	0.002 ± 0.045 (0)
pMCI	1319	69.00 ± 0.00 (69)	0.9891 ± 0.0110 (0.9927)	0.006 ± 0.117 (0)

B.3 Preprocessing and Registration Quality

Preprocessing Details: All scans of shape $182 \times 218 \times 182$ are preprocessed with a common pipeline comprising: reorientation to canonical (`fslreorient2std`), bias-field correction (ANTs `N4BiasFieldCorrection`), skull stripping (SynthStrip Hoopes et al. (2022)), affine+nonlinear registration to MNI152 1 mm (FSL `FLIRT`→`FNIRT`), intensity z-scoring within the brain mask and final masking. Finally, the scans are padded to be divisible by the patch volumes. The resulting MNI-aligned, skull-stripped, z-scored volumes are saved per scan and used for tokenization in Parcel encoder.

Parcel Coverage Quality Control To assess whether the anatomical contexts available to the model differ systematically across diagnostic groups, we performed patch-level parcel-coverage quality control on the registered scans. For each registered scan, we derived a brain mask from the preprocessed image and converted it into the same patch space used for parcel tokenization. We then quantified: (i) the number of non-empty parcels, (ii) mean parcel coverage relative to the CN reference distribution, (iii) the number of low-coverage parcels, and (iv) the percentage of valid brain patches assigned to any atlas-defined parcel.

Table A2 summarizes the main QC metrics across CN, AD, sMCI, and pMCI scans. All disease/progression groups retained complete parcel coverage, with a median of 69/69 non-empty parcels and zero low-coverage parcels. Although mean parcel coverage showed some group differences, the absolute differences were small. Importantly, AD, sMCI, and pMCI did not show parcel dropout or systematic failure of atlas assignment.

The Kruskal–Wallis test (Kruskal & Wallis (1952)) showed no significant group difference in the number of non-empty parcels ($p = 0.65$), indicating no evidence of systematic parcel dropout across CN, AD, sMCI, and pMCI. The number of low-coverage parcels was also negligible in all groups, with a median of zero across all cohorts. We also observed a small reduction in valid brain-mask patches and total brain-mask voxels following the expected ordering CN $>$ sMCI $>$ pMCI $>$ AD. They therefore reduce concern that differences in predictive performance arise from missing anatomical contexts in particular diagnostic groups.

B.4 Controlled Synthetic Dataset

To directly assess whether the proposed representations preserve localized anatomical abnormalities and their longitudinal evolution, we constructed a controlled synthetic volumetric dataset in which the spatial location, onset, severity, and temporal progression of every abnormality are known exactly. The dataset was designed specifically to evaluate two properties that are difficult to establish from real neuroimaging cohorts: (i) whether highly weighted parcels correspond to the regions in which abnormalities were introduced, and (ii) whether the longitudinal model tracks the progressive recruitment and worsening of these abnormalities over time. Fig A2 illustrates the parcels and sample data generated.

Synthetic anatomy. Each scan is represented by a $96 \times 96 \times 96$ 3D volume with a zero-valued background and six non-overlapping parcels. The parcel set consists of four spherical and two cuboidal regions of varying sizes, gray-scale intensities, and internal structure. Parcel locations are fixed across subjects and visits, providing exact spatial correspondence throughout the dataset. To introduce intra-parcel structural variation, each parcel contains either one or two concentric intensity

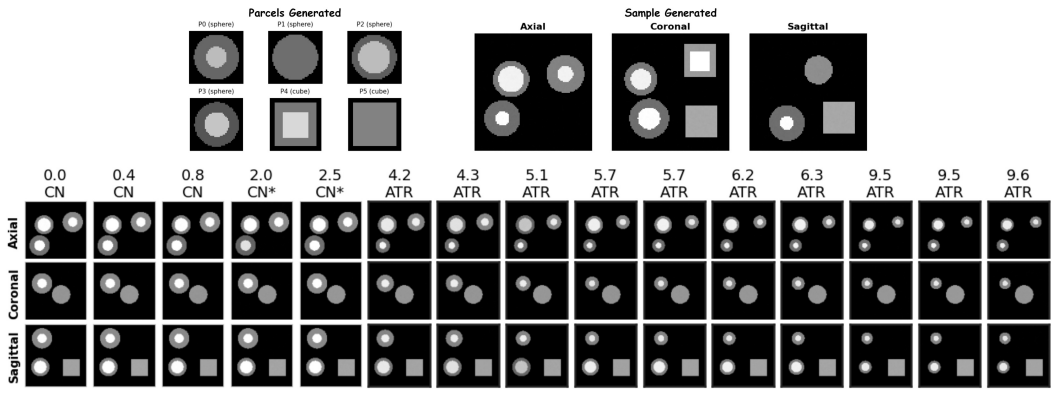

Figure A2: **Controlled synthetic data.** Example 3D volumes illustrating the fixed parcel geometry and severity-dependent structural perturbations used to generate longitudinal trajectories.

layers. The resulting volumes therefore provide spatially localized structures with heterogeneous geometry, volume, and appearance while retaining exact knowledge of parcel boundaries. This construction is not intended to simulate realistic brain anatomy; rather, it provides a controlled analogy of the structured setting in which localized regions with different morphology evolve over time.

Longitudinal trajectories. We generate two types of subject trajectories: *stable CN* and *converter*. Stable CN subjects retain unchanged parcel anatomy throughout follow-up, apart from mild scan-level nuisance variation. Converter subjects follow a cumulative $\mathrm{CN} \rightarrow \mathrm{ATR}$ trajectory in which abnormalities are progressively introduced into a randomly ordered subset of parcels. Any parcel may therefore be recruited first, and subsequent parcels are sampled without replacement.

For a selected parcel p, severity evolves continuously as:

$$s_p(t) = \begin{cases} 0, & t < \tau_p, \\ s_p^{\max}\left[1 - \exp\left(-r_p(t - \tau_p)\right)\right], & t \geq \tau_p, \end{cases} \tag{1}$$

where τ_p is the parcel-specific recruitment time, r_p is its progression rate, and $s_p^{\max}$ is its maximum severity. This produces a monotonic, saturating progression with no spontaneous recovery.

For each converter, the onset of the first affected parcel is sampled randomly. Additional parcels are then recruited sequentially using randomly sampled inter-recruitment intervals. Recruitment order, onset times, progression rates, and maximum severities are independently sampled for each subject, producing heterogeneous longitudinal trajectories despite a common generative process.

Conversion and preclinical scans. The total latent disease burden at time t is defined as $B(t) = \sum_{p=1}^{P} s_p(t)$. The synthetic conversion time is defined as the earliest observation at which $B(t)$ exceeds a fixed burden threshold, set to 0.85 in our experiments. Because regional abnormalities can begin before this threshold is reached, biological onset and observed conversion are deliberately separated. Scans acquired after the onset of regional abnormality but before conversion are retained as *preclinical CN* (CN*) observations. This enables evaluation of whether the model responds to evolving regional abnormalities before the subject-level label changes. CN* observations retain the CN target label during training.

Severity-dependent perturbations. The latent parcel severity $s_p(t) \in [0, 1]$ directly controls the magnitude and type of perturbation applied to parcel p at visit t. Rather than switching abruptly between normal and abnormal states, the synthetic progression is rendered continuously, with appearance changes introduced at different severity ranges.

For the inner region of an affected parcel, intensity is attenuated as:

$$I_p^{\mathrm{inner}}(t) = I_{p,0}^{\mathrm{inner}}\left(1 - 0.55\, s_p(t)\right), \tag{2}$$

where $I_{p,0}^{\mathrm{inner}}$ denotes the original inner-region intensity. Thus, the intensity attenuation increases linearly with parcel severity.

Table A3: Configuration of the controlled synthetic dataset. ($\mathcal{U}(a, b)$ denotes uniform sampling over the interval $[a, b]$.)

Parameter	Setting
Volume size	96^3
Patch size	8^3
Number of parcels	6
Parcel composition	4 spheres + 2 cubes
Parcel radius	$\mathcal{U}(6, 10)$ voxels
Number of subjects	5,000
Stable CN / converter	50% / 50%
Visits per subject	10–15
Maximum follow-up	10 years
Affected parcels per converter	2–5
First abnormality onset	$\mathcal{U}(0.25, 4.0)$ years
Scan interval	$\mathcal{U}(0.7, 2.0)$ years
Progression rate	$\mathcal{U}(0.35, 0.90)$
Maximum parcel severity	$\mathcal{U}(0.65, 1.0)$
Conversion burden threshold	0.85
Train / validation / test	70% / 10% / 20% subjects

The outer-shell intensity is similarly attenuated as

$$I_p^{\text{shell}}(t) = I_{p,0}^{\text{shell}} \left(1 - 0.50\, s_p(t)\right), \tag{3}$$

thereby progressively altering the intensity contrast between parcel substructures as severity increases.

Texture disruption is introduced only after the parcel reaches moderate severity. Its activation strength is defined as

$$g_{\text{tex}}(s_p) = \text{clip}\left(\frac{s_p - 0.25}{0.75}, 0, 1\right), \tag{4}$$

where $g_{\text{tex}} = 0$ for $s_p \leq 0.25$, and increases progressively thereafter. Gaussian-smoothed local noise is added to the affected region with an amplitude proportional to $\sigma_{\text{tex}} = 0.12\, g_{\text{tex}}(s_p)$. Hence, mild abnormalities are represented primarily by intensity and contrast changes, while increasing severity additionally introduces local textural disruption.

Structural atrophy is introduced at a later stage of progression. We define

$$g_{\text{atr}}(s_p) = \text{clip}\left(\frac{s_p - 0.45}{0.55}, 0, 1\right), \tag{5}$$

such that no geometric atrophy is applied for $s_p \leq 0.45$. For larger severity values, the spatial extent of the parcel is progressively reduced using

$$\gamma_p(t) = \max\left(0.35,\, 1 - 0.45\, g_{\text{atr}}(s_p(t))\right), \tag{6}$$

where $\gamma_p(t)$ denotes the multiplicative scaling of the parcel geometry. The inner and outer parcel intensities are also mildly reduced during this stage.

Overall, the rendering process follows the progression

$$\text{intensity alteration} \;\rightarrow\; \text{+ texture disruption} \;\rightarrow\; \text{+ structural atrophy},$$

with the magnitude of each perturbation controlled by the same latent severity $s_p(t)$. Consequently, parcel severity represents a continuous latent progression variable rather than a discrete abnormality label, providing an exact correspondence between known regional disease burden and the resulting synthetic structural changes.

Scan-level variability. To prevent the task from reducing to exact template matching, mild global nuisance variability is independently applied to each scan. This includes small multiplicative intensity variation, a smooth low-frequency bias field, and Gaussian-smoothed image noise. These perturbations alter scan appearance without changing the parcel-level ground-truth abnormality.

The dataset configuration details are summarized in Table A3.

Table A4: Training and hyperparameter configuration details for SBT_{SS}

Aspect	Value
Backbone	ViT-base (modified), pre-norm
Patch size	$8 \times 8 \times 8$
Tokenizer	MLP
Encoder layers	12
Encoder heads	12
Embed dim	768
FFN hidden dim	2048
Dropout	0.10
GeM pooling	p=2
Parcel drop prob	0.10
Optimizer	AdamW
Loss function	Binary cross-entropy
LR	1×10^{-4}
Weight decay	1×10^{-4}
Scheduler	Cosine , min LR: 10^{-6}
Label smoothing	0.05
Accumulation steps	5
Batch size	12
Epochs	20

Table A5: Training and hyperparameter configuration details for SBT_{SL},SBT_{PL}

Aspect	Value
Backbone	2-layer transformer
Encoder layers	2
Encoder heads	8
MLP ratio	8
Embed dim	768
Pool	Mean-pooling
Dropout	0.10
Optimizer	AdamW
Loss function	Binary cross-entropy
LR	3×10^{-4}
Weight decay	3×10^{-3}
Scheduler	Cosine , min LR: 10^{-6}
Batch size	64
Epochs	20

Table A6: Training and hyperparameter configuration details for SBT_{SS} on the controlled synthetic dataset.

Aspect	Value
Patch size	$8 \times 8 \times 8$
Embed dim	256
Encoder layers	4
Encoder heads	4
Optimizer	AdamW
LR	3×10^{-4}
Batch size	64
Epochs	5

Table A7: Training and hyperparameter configuration details for SBT_{SL},SBT_{PL} on the controlled synthetic dataset.

Aspect	Value
Backbone	2-layer transformer
Embed dim	256
Encoder layers	2
Encoder heads	4
Optimizer	AdamW
LR	3×10^{-4}
Batch size	32
Epochs	5

C Implementation and Baseline Details

C.1 Code

The complete model code, synthetic data generation scripts and training scripts will be uploaded to GitHub.

C.2 Training setup and Model details

The model and optimization configurations used for SBT_{SS} and the longitudinal variants SBT_{SL}/SBT_{PL} are summarized in Tables A4 and A5, respectively. The complete longitudinal SBT architecture contains approximately **77**M parameters. For the maximum evaluated history length of $T = 15$ visits, `ptflops` estimates a per-subject forward-pass cost of $\mathbf{3.72 \times 10^{13}}$ FLOPs. The corresponding inference-time FLOPs for SBT variants, baselines, and prior methods used in Fig. 5B are reported in Table A9.

Synthetic experiments. For the controlled synthetic dataset, we use a deliberately lower-capacity architecture given the simpler volumetric structure of the controlled setting. The synthetic SBT encoder uses a 256-dimensional embedding space with 4 Transformer layers and 4 attention heads, while retaining the same parcel-based representation and longitudinal modeling principles as the main model. The complete configurations and training details are provided in Tables A6 and A7.

Experimental Protocol. We follow diagnosis-stratified, 5-fold cross-validation with subject-wise split per fold to prevent any leakage and 10%val split. We report mean$_{(std)}$ results of Balanced Accuracy (BACC), and AUC on the 20% held-out test set within each fold. The same is followed for synthetic dataset as well. All experiments were conducted on 4 NVIDIA RTX A6000s GPUs with Intel CPU (1TB RAM). Notably, *SBT exhibits architectural flexibility*, as its parameter count

(shared weights) is invariant to number of parcels(P) and visits(T), enabling scalability to variable-length visits.

Parcel Dropout details: To prevent over-reliance on a small subset of regions, parcel-dropout is introduced as a regularizer with rate $\rho \in [0, 1)$ during training: sample $d_{p,t} \sim \text{Bernoulli}(\rho)$ and set $k_{p,t} = 1 - d_{p,t}$. Then, softmax is applied over parcels using temperature-scaled scores with $\tau > 0$:

$$\alpha_{p,t} = \frac{\exp\big(g(z_{p,t})/\tau\big)}{\sum_{q=1}^{P} \exp\big(g(z_{q,t})/\tau\big)}, \qquad g(z_{p,t}) = \begin{cases} g(z_{p,t}), & k_{p,t} = 1, \\ -\infty, & k_{p,t} = 0, \end{cases}$$

Grouping-based ablation experiments: To validate that the gains of our anatomy-based grouping are attributable to biologically meaningful parcel structure rather than merely to token partitioning itself, we introduce two non-anatomical grouping baselines. These baselines preserve the same overall framework, including the encoder, parcel-wise aggregation mechanism using the scorer, classifier, and the total number of groups, while changing only the grouping strategy.

(i) Random grouping: In the first baseline, all foreground brain tokens are reassigned into a fixed set of 69 groups using a random partition. Background tokens are excluded, and each valid brain token is assigned to exactly one random group. The random grouping is generated once and then kept fixed for all samples, ensuring that the comparison reflects a stable non-anatomical partition. This baseline tests whether the proposed framework derives its benefit simply from dividing tokens into multiple groups, irrespective of whether those groups carry anatomical meaning.

(ii) 3D block grouping: In the second baseline, all foreground brain tokens are partitioned into a fixed set of 69 spatially contiguous groups using a geometric block-based strategy. Starting from the complete set of valid brain tokens, the token set is recursively split along the spatial axis with the largest extent, using median cuts, until exactly 69 groups are obtained. This produces groups that are spatially coherent but not aligned to neuroanatomical boundaries. The block-grouping baseline therefore tests whether coarse spatial locality alone is sufficient, or whether explicit anatomical organization is necessary for the observed performance gains.

Table A8: Ablations based on grouping strategy for ADNI on AD vs. CN task (mean$_{\pm \text{std}}$ over 5 folds).

Model	Acc. (%)	BAcc. (%)	AUC (%)
Random Grouping	$75.58_{\pm 1.89}$	$57.77_{\pm 3.53}$	$70.31_{\pm 2.58}$
3D block Grouping	$78.2_{\pm 2.09}$	$77.23_{\pm 1.54}$	$84.15_{\pm 2.45}$
SBT$_{SS}$	$\mathbf{88.54_{\pm 2.12}}$	$\mathbf{84.21_{\pm 3.84}}$	$\mathbf{91.18_{\pm 2.86}}$

Together, these baselines serve as complementary controls: random grouping removes both anatomical and spatial structure, whereas block grouping preserves spatial contiguity but discards anatomical meaning. Table A8 shows that anatomy-aware grouping consistently outperforms both baselines.

C.3 Prior works and Baselines

C.3.1 Broader neuroimaging literature

Deep learning (DL) for AD diagnosis and MCI's prognostic task from structural T1w MRI has evolved through complementary paradigms, each addressing key limitations.

3D Single-scan models. Early DL methods moved from 2D slice inputs (Gao et al. (2023)) to consuming full 3D scans to preserve volumetric context and spatial continuity. 3D CNNs (Folego et al. (2020); Oh et al. (2019); Qiu et al. (2020)) were extensively used due to their strong inductive bias under limited data (Khatri et al. (2024)), but their interpretations are often clinically opaque and can be driven by shortcut cues. Recently, ViTs (Kunanbayev et al. (2024)) and CNN-ViT hybrids (Altay et al. (2021); Zhu et al. (2021a); Li et al. (2022); Wu et al. (2022)) have been explored to better model large global 3D context. Region-discovery approaches like (Lian et al. (2018); Zhu et al. (2021a)) attempt to localize discriminative patterns via patch-based aggregation, but arbitrary patching can still encourage learning shortcut cues. Multi-plane/ multi-slice tokenization works (Khatri et al. (2024)) like M3T (Jang & Hwang (2022)) adopted a hybrid backbone, improving efficiency but compromising true 3D continuity. To incorporate anatomy-aware designs in single-scan inputs, AAGN (Jiang & Miao (2024)) injects priors using an atlas. Yet, AAGN's reliance on ROI-feature extraction from overlapping CNN feature maps limits true parcel-specific representations. While anatomical grounding is especially valuable for clinically critical pMCI vs. sMCI conversion (Guan et al. (2023); Jiang & Miao (2024)), these methods remain structurally cross-sectional (single scan), leaving progression modeling and longitudinal validity largely unaddressed.

Table A9: Architectural and computational comparison of SBT with single-scan and longitudinal baselines. FLOPs correspond to the indicated number of visits T and are estimated using `ptflops`.

Model	Long.	T	FLOPs (e+12)	Spatial Backbone	Anat.-aware	High-dim.	Folds (original)	Var. length	Irreg. interval
Single-scan									
DenseNet121	✗	1	1.37	3D CNN	✗	✓	–	✗	✗
ResNet152	✗	1	1.97	3D CNN	✗	✓	–	✗	✗
ViT	✗	1	10.12	3D Transformer	✗	✓	–	✗	✗
M3T	✗	1	7.01	Hybrid (slice-based)	✗	✗	1	✗	✗
AAGN	✗	1	8.09	3D CNN	✓	✗	5	✗	✗
SBT_{SS}	✗	1	2.48	3D Transformer	✓	✓	**5**	✗	✗
Multi-scan / longitudinal									
Longformer	✓	2	5.37	3D CNN	✗	✓	1	✗	✗
DenseNet121+GRU	✓	15	20.55	3D CNN	✗	✓	–	✓	✓
DenseNet121+LT	✓	15	20.55	3D CNN	✗	✓	–	✓	✓
ViT+LT	✓	15	131.56	3D Transformer	✗	✓	–	✓	✓
SBT_{SL}	✓	2	4.96	3D Transformer	✓	✓	**5**	✓	✓
SBT_{SL}	✓	15	37.30	3D Transformer	✓	✓	**5**	✓	✓
SBT_{PL}	✓	15	37.22	3D Transformer	✓	✓	**5**	✓	✓

Note. *Long.* denotes whether the model explicitly uses longitudinal scans. *FLOPs* denotes full FLOPs reported for the respective visits considered (T) during inference using ptflops. *Anat.-aware* indicates whether the method explicitly incorporates anatomical or parcel-level structure. *High-dim.* indicates whether the method directly operates on high-dimensional scan-derived representations or downsampled inputs. "*Folds (original)*" denotes the original paper's settings for reporting results. "–" indicates that the corresponding entry is a baseline. *Var. length* denotes support for variable numbers of visits across subjects. *Irreg. interval* denotes whether irregular follow-up intervals can be explicitly incorporated. ✓: supported; ✗: not supported.

Longitudinal neuroimaging. The main paper focuses on deep longitudinal representation-learning methods most directly related to SBT. Earlier longitudinal structural MRI approaches commonly relied on predefined imaging measurements rather than end-to-end learned representations. Voxel-level methods model longitudinal change from voxel-wise morphometric measurements (Davatzikos et al. (2009); Wang et al. (2014); Huang et al. (2017b)), while ROI-based approaches characterize progression through trajectories of regional features (Zhu et al. (2021b); Li et al. (2011); Farzan et al. (2015)). Another line of work focuses on longitudinal shape or morphometric changes in specific anatomical structures, particularly the hippocampus (Li et al. (2019); Sarasua et al. (2022); Olaoluwa et al. (2026)). These methods provide explicit measurements of anatomical change, but operate on predefined voxel-, ROI-, or structure-level descriptors rather than learning spatial representations and temporal aggregation directly from longitudinal imaging data. Recent DL models increasingly aim to avoid these brittle preprocessing dependencies by learning task-relevant representations through a data-driven approach (Oh et al. (2019)).

C.3.2 Baseline Implementations and Computational Comparison

Beyond the methods discussed in the main paper, several deep longitudinal frameworks impose input or temporal constraints that make direct comparison under our full-resolution, variable-length setting difficult.

Protocol. The primary comparisons in the main paper are therefore restricted to architectures that can be evaluated on full-resolution inputs (1 mm volumetric setting used by SBT) without discarding longitudinal observations solely because of fixed-history requirements. Their architectural properties and computational costs are summarized in Table A9. Several of these methods were originally evaluated using different input resolutions, cohort sizes, or numbers of folds; for example, LongFormer (Chen et al. (2024)) reports results using a single fold in its original evaluation. For consistency, we retrain the applicable architectures using the same preprocessed inputs, subject-level splits, and five-fold evaluation protocol as SBT. For the paired-scan comparison with LongFormer, we additionally train SBT_{SL} (2T) using the same temporal input configuration, i.e., a current and a prior scan. We also report an end-to-end SBT_{SL} control in which the Parcel Encoder and longitudinal model are jointly optimized. Due to the memory constraints, this setting uses a batch size of one subject, with the number of scans per optimization step determined by the available history of that subject. Despite this, the end-to-end model remains stronger than the evaluated baselines.

Table A10: Performance of other methods across datasets and tasks on the full dataset as SBT. Values are reported as $\text{mean}_{\pm\text{std}}$ over 5-folds (in %).

	ADNI (AD/CN)		ADNI (sMCI/pMCI)		AIBL (AD/CN)		MIRIAD (AD/CN)	
Method	BACC	AUC	BACC	AUC	BACC	AUC	BACC	AUC
Jomeiri et al.Jomeiri et al. (2024)	$77.43_{\pm 3.95}$	$87.11_{\pm 1.09}$	$62.24_{\pm 1.47}$	$73.06_{\pm 4.24}$	$69.95_{\pm 3.85}$	$85.23_{\pm 4.90}$	$81.70_{\pm 3.52}$	$85.45_{\pm 4.99}$
Ouyang et al.Ouyang et al. (2022)	$74.61_{\pm 2.21}$	$82.45_{\pm 2.37}$	$60.17_{\pm 2.93}$	$69.74_{\pm 3.13}$	$65.47_{\pm 2.69}$	$82.06_{\pm 4.32}$	$86.66_{\pm 3.87}$	$88.82_{\pm 4.44}$

Methods whose original formulations with spatial or temporal constraints are evaluated separately as supplementary comparisons. For the works that use fixed histories like Jomeiri et al. (2024) or compromise the spatial resolution to $64 \times 64 \times 64$ like Ouyang et al. (2022), we have retrained these models and reported their results in Table A10. To enable Jomeiri et al. (2024) to handle variable-length inputs, we introduced age as the temporal embedding, like the other Longitudinal baselines in the main paper. Also, Ouyang et al. (2022) uses only a pair of scans for the longitudinal modeling. It is observed that SBT_{SL} with paired scans (Table 1) outperforms these methods.

C.4 Additional Longitudinal Variant: SBT_{PL-B}

We additionally consider a joint parcel–time variant, $\text{SBT}_{\text{PL}-\text{B}}$, to test whether explicitly modeling all cross-parcel and cross-time interactions provides an advantage over the factorized parcel-wise temporal modeling used by SBT_{PL}. Unlike SBT_{PL}, which models the temporal trajectory of each anatomical region using a shared Longitudinal Transformer, $\text{SBT}_{\text{PL}-\text{B}}$ forms a single sequence containing all parcel–visit representations.

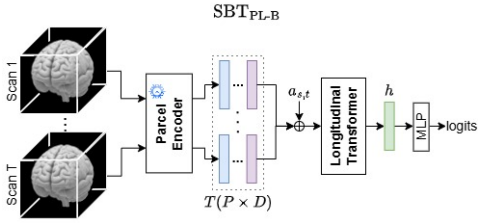

Figure A3: A parcel level variant of SBT using the same Longitudinal Transformer framework ($\text{SBT}_{\textbf{PL-B}}$).

$\text{SBT}_{\text{PL}-\text{B}}$ learns *joint* cross-parcel and cross-time interactions in a single transformer, capturing coordinated anatomical changes (Fig. A3). We build a single sequence that interleaves *all* parcel–time tokens and add their respective age-based temporal embedding:

$$\tilde{z}_{p,t} = z_{p,t} + \phi_t(A_s(t)), \quad h = \mathcal{T}_{parcelB}\big(\tilde{z}_{1:P,1:T}\big) \in \mathbb{R}^D$$

A Longitudinal transformer $\mathcal{T}_{parcelB}(\cdot)$ jointly models cross-parcel and cross-time interactions. The `[CLS]` token is fed to an MLP head for logits: $\hat{y} = \text{softmax}\big(\text{MLP}(h)\big)$.

C.5 Why Structure Must Precede Global Contextualization

The purpose of anatomical grouping in SBT is to ensure that each regional representation z_p is formed from its corresponding anatomical support before broader aggregation. This distinction becomes important when comparing *global-first* and *structure-first* representations.

Let $X = \{x_i\}_{i=1}^N$ denote the volumetric tokens and let $M \in \mathbb{R}_{\geq 0}^{N \times P}$ denote the token-to-parcel membership with $\mathcal{I}_p = \{i : M_{ip} > 0\}$.

Global-first representation. Consider a design in which unrestricted self-attention is first applied over all volumetric tokens, $Y = f_\theta(X)$, and anatomical grouping is performed only afterward:

$$u_p = \sum_{i=1}^{N} M_{ip} Y_i, \qquad s = \sum_{p=1}^{P} \alpha_p u_p.$$

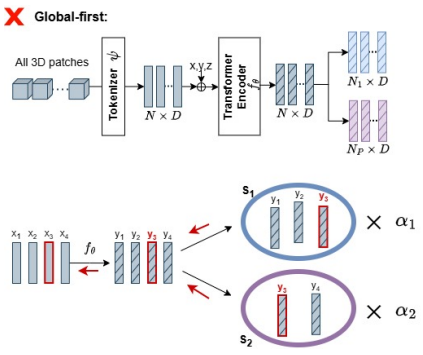

Figure A4: Late anatomical grouping after global contextualization.

Although u_p is pooled using the anatomical support of parcel p, each contextualized token Y_i may already depend on tokens throughout the scan. Consequently,

$$\frac{\partial u_p}{\partial x_j} = \sum_{i=1}^{N} M_{ip} \frac{\partial Y_i}{\partial x_j},$$

which is generally non-zero even for $j \notin \mathcal{I}_p$ under global self-attention. Thus, anatomical pooling performed after global interaction does not guarantee that the resulting parcel representation was formed solely from its corresponding anatomical context.

Structure-first representation. SBT instead restricts contextualization to the anatomical support before aggregation. For parcel p,

$$z_p = f_\theta\big(X_{\mathcal{I}_p}\big), \qquad s = \sum_{p=1}^{P} \alpha_p z_p.$$

This construction satisfies the support-locality property $\frac{\partial z_p}{\partial x_j} = 0, j \notin \mathcal{I}_p$. Probabilistic memberships may allow boundary tokens to participate in more than one parcel, but information outside the corresponding support cannot enter that parcel representation before aggregation.

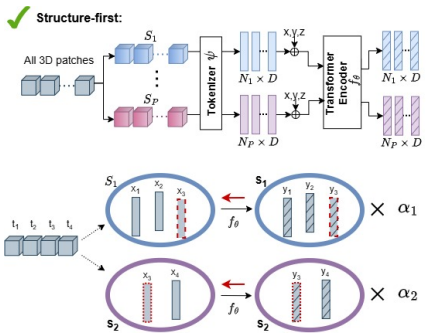

Figure A5: SBT's principle of early structure introduction.

Illustrative mixing scenario. (Fig A4,A5) Consider two overlapping parcel supports

$$\mathcal{I}_1 = \{x_1, x_2, x_3\}, \qquad \mathcal{I}_2 = \{x_3, x_4\}.$$

After global self-attention, features pooled into parcel 1 may already contain information originating from x_4, despite $4 \notin \mathcal{I}_1$. In contrast, structure-first contextualization forms $z_1 = f_\theta(x_1, x_2, x_3)$ independently of x_4. The shared token x_3 may legitimately participate in both regions through the probabilistic anatomical assignment.

Empirical implication. This analysis does not imply that global-first architectures cannot learn useful regional representations. Rather, they lack an architectural guarantee that region-level features remain local after global contextualization. Consistent with this distinction, the late-grouping control in Figure 5 performs substantially worse than structure-first SBT. Together, the analysis and controlled experiment support introducing anatomical structure during representation formation rather than only after global feature mixing.

D Additional Experimental Analysis

D.1 Detailed Classification Results

Table A11 provides the complete classification results for the longitudinal SBT variants SBT_{SL}, SBT_{PL}, and $\text{SBT}_{\text{PL-B}}$, including alternative MLP- and 3D DenseNet-based tokenization configurations. We report accuracy, balanced accuracy, F1 score, and AUC as mean$_{\pm\text{std}}$ over five subject-level folds. Note: The comparatively higher standard deviations in all tables arise from fold-to-fold variation in scan counts induced by subject-level partitioning.

Controlled synthetic classification. Before evaluating regional localization, we verify that the controlled synthetic trajectories are learnable at both single-scan and longitudinal levels. Table A12 reports performance on the held-out synthetic test set. All three variants achieve high discrimination, with longitudinal modeling providing a modest improvement over single-scan prediction.

D.2 Temporal Encoding under Cohort Shift

We additionally examine whether the choice of temporal encoding affects cross-cohort inference. Models trained on ADNI are evaluated directly on the external AIBL and

Table A11: Detailed classification results across longitudinal SBT variants and tokenization configurations. Results are reported as mean$_{\pm \mathrm{std}}$ over five subject-level folds.

Dataset & Task	Model	Tokenizer	Acc (%)	Bal. Acc (%)	F1 (%)	AUC (%)
ADNI [AD vs CN]	SBT_{PL}	MLP	$92.96_{\pm 1.61}$	$88.27_{\pm 3.80}$	$84.39_{\pm 3.55}$	$94.58_{\pm 1.57}$
	$SBT_{PL\text{-}B}$	MLP	$92.05_{\pm 2.47}$	$87.38_{\pm 2.83}$	$81.36_{\pm 4.33}$	$94.00_{\pm 2.02}$
	SBT_{SL}	MLP	$\mathbf{93.91_{\pm 1.68}}$	$89.20_{\pm 2.21}$	$84.81_{\pm 4.05}$	$\mathbf{94.84_{\pm 1.66}}$
	SBT_{PL}	3D Densenet	$92.14_{\pm 1.22}$	$87.93_{\pm 2.23}$	$82.88_{\pm 4.53}$	$94.24_{\pm 1.17}$
	$SBT_{PL\text{-}B}$	3D Densenet	$91.46_{\pm 1.25}$	$87.54_{\pm 1.47}$	$82.42_{\pm 2.76}$	$93.55_{\pm 1.22}$
	SBT_{SL}	3D Densenet	$93.01_{\pm 1.15}$	$\mathbf{89.90_{\pm 1.74}}$	$\mathbf{85.25_{\pm 3.19}}$	$94.71_{\pm 0.78}$
ADNI [sMCI vs pMCI]	SBT_{PL}	MLP	$84.61_{\pm 1.13}$	$76.65_{\pm 1.90}$	$67.02_{\pm 1.98}$	$85.86_{\pm 0.60}$
	$SBT_{PL\text{-}B}$	MLP	$82.73_{\pm 1.65}$	$76.94_{\pm 2.52}$	$67.08_{\pm 3.11}$	$82.02_{\pm 1.41}$
	SBT_{SL}	MLP	$\mathbf{84.62_{\pm 1.00}}$	$\mathbf{78.26_{\pm 0.91}}$	$\mathbf{68.55_{\pm 1.53}}$	$\mathbf{85.96_{\pm 1.18}}$
	SBT_{PL}	3D Densenet	$84.27_{\pm 1.57}$	$76.09_{\pm 1.90}$	$66.20_{\pm 2.68}$	$83.72_{\pm 3.73}$
	$SBT_{PL\text{-}B}$	3D Densenet	$81.97_{\pm 1.60}$	$73.36_{\pm 2.65}$	$62.07_{\pm 3.54}$	$83.15_{\pm 2.29}$
	SBT_{SL}	3D Densenet	$84.54_{\pm 1.36}$	$77.51_{\pm 2.01}$	$67.71_{\pm 2.06}$	$85.52_{\pm 1.88}$
AIBL [AD vs CN]	SBT_{PL}	MLP	$\mathbf{95.39_{\pm 0.57}}$	$\mathbf{88.38_{\pm 3.09}}$	$\mathbf{81.53_{\pm 4.24}}$	$92.21_{\pm 3.06}$
	$SBT_{PL\text{-}B}$	MLP	$94.96_{\pm 1.43}$	$86.57_{\pm 4.42}$	$79.66_{\pm 4.43}$	$91.14_{\pm 4.17}$
	SBT_{SL}	MLP	$95.13_{\pm 0.59}$	$86.86_{\pm 3.47}$	$80.37_{\pm 3.19}$	$90.89_{\pm 4.69}$
	SBT_{PL}	3D Densenet	$93.68_{\pm 1.78}$	$84.82_{\pm 3.80}$	$75.88_{\pm 4.38}$	$92.13_{\pm 3.36}$
	$SBT_{PL\text{-}B}$	3D Densenet	$93.36_{\pm 0.98}$	$81.84_{\pm 3.47}$	$72.46_{\pm 4.81}$	$\mathbf{93.11_{\pm 1.71}}$
	SBT_{SL}	3D Densenet	$93.82_{\pm 1.59}$	$84.85_{\pm 3.16}$	$75.41_{\pm 4.65}$	$90.53_{\pm 4.39}$
MIRIAD [AD vs CN]	SBT_{PL}	MLP	$96.97_{\pm 4.41}$	$96.54_{\pm 3.58}$	$97.57_{\pm 4.15}$	$98.33_{\pm 3.00}$
	$SBT_{PL\text{-}B}$	MLP	$92.93_{\pm 4.31}$	$93.16_{\pm 3.81}$	$94.14_{\pm 3.08}$	$96.96_{\pm 3.56}$
	SBT_{SL}	MLP	$97.41_{\pm 3.45}$	$96.90_{\pm 3.55}$	$98.14_{\pm 3.89}$	$98.05_{\pm 3.22}$
	SBT_{PL}	3D Densenet	$95.73_{\pm 4.69}$	$95.47_{\pm 3.98}$	$96.57_{\pm 3.83}$	$98.00_{\pm 2.58}$
	$SBT_{PL\text{-}B}$	3D Densenet	$97.20_{\pm 3.93}$	$96.43_{\pm 4.05}$	$97.86_{\pm 3.00}$	$\mathbf{98.40_{\pm 1.11}}$
	SBT_{SL}	3D Densenet	$\mathbf{98.05_{\pm 2.48}}$	$\mathbf{98.44_{\pm 2.00}}$	$\mathbf{98.39_{\pm 2.17}}$	$98.33_{\pm 1.11}$

Table A12: Classification performance on the held-out controlled synthetic test set.

Model	Acc. (%)	BAcc. (%)	F1 (%)	AUC (%)
SBT_{SS}	97.57	98.18	95.79	99.82
SBT_{SL}	**98.77**	**99.00**	**97.82**	**99.97**
SBT_{PL}	97.96	98.21	96.42	99.88

MIRIAD cohorts without downstream adaptation. These cohorts differ from ADNI in acquisition characteristics, follow-up structure, and age distribution, providing a complementary test of whether performance depends strongly on absolute-age encoding. Using the trained checkpoints directly for inference on the external AIBL and MIRIAD cohorts, we obtain the results in Table A13. Performance remains comparable across the external cohorts under both temporal encoding strategies, despite differences in their age distributions. Importantly, relative-time encoding transfers comparably to absolute-age encoding, and is superior for several metrics.

Table A13: Direct external-cohort inference under relative-time and absolute-age encoding.

Model	Age encoding	AIBL [AD/CN]		MIRIAD [AD/CN]	
		BAcc (%)	AUC (%)	BAcc (%)	AUC (%)
SBT_{SL}	Relative time	$83.13_{(3.12)}$	$91.02_{(2.32)}$	$89.14_{(3.48)}$	$94.31_{(3.17)}$
SBT_{SL}	Absolute age	$82.75_{(3.07)}$	$90.58_{(2.18)}$	$90.83_{(4.19)}$	$93.42_{(3.62)}$
SBT_{PL}	Relative time	$82.78_{(3.62)}$	$90.67_{(2.86)}$	$89.01_{(4.01)}$	$94.46_{(3.09)}$
SBT_{PL}	Absolute age	$82.02_{(3.13)}$	$90.04_{(3.01)}$	$90.21_{(3.15)}$	$94.02_{(3.52)}$

Together with the age-only and no-time ablations reported in the main paper, these results indicate that the longitudinal gains cannot be attributed primarily to absolute age as a predictive shortcut. Imaging history provides the dominant signal, while temporal encoding supplies complementary information about when the observations were acquired.

D.3 Survival analysis

Binary sMCI/pMCI classification captures whether conversion occurs but not *when*. We therefore perform censoring-aware landmark survival analysis with SBT_{SL} to assess whether the learned representations predict time to MCI-to-AD conversion. The main paper reports comparisons at 0-, 12-,

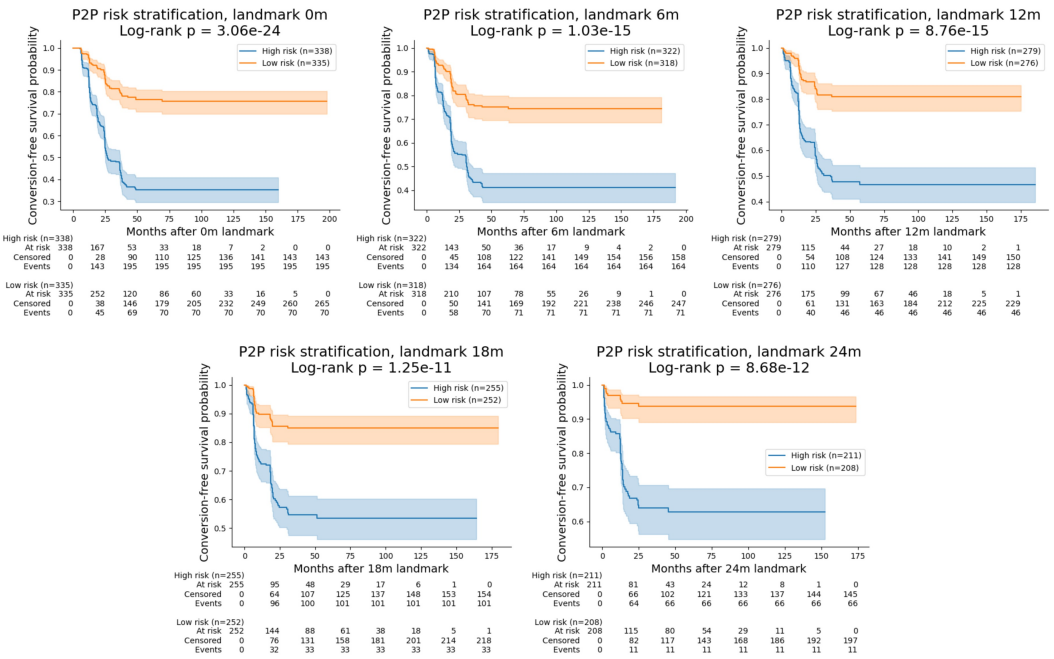

Figure A6: **Kaplan-Meier curves** at landmarks 0, 6, 12, 18, 24 months showing the separation of the high and low risk MCI subjects from ADNI.

and 24-month landmarks; here we provide additional intermediate-landmark, time-dependent AUC, and Kaplan–Meier analyses.

Setting: For each subject, we constructed a subject-level survival record from the scan-level ADNI metadata. The model uses longitudinal MRI information available up to a predefined landmark month and predicts future conversion risk after that landmark. At a landmark L, all visits with visit month less than or equal to L are used as model input. Subjects who converted before or at the landmark are excluded, because their event has already occurred before the prediction origin. For subjects with observed conversion, the survival duration is computed as conversion time minus landmark time. For censored subjects, the duration is computed as last known follow-up or censoring time minus landmark time.

Methodology: At each landmark $t_L \in \{0, 6, 12, 18, 24\}$ months, subjects not yet converted form the at-risk cohort, and all MRI examinations observed up to t_L are used; non-converters are right-censored. Frozen parcel representations are aggregated at each visit and integrated across the available history using $\mathrm{SBT}_{\mathrm{SL}}$, yielding one landmark-specific subject representation. The parcel scorer, longitudinal Transformer, and Cox proportional-hazards (Cox (1972)) risk head are jointly optimized for MCI-to-AD conversion per landmark within 5-fold CV:

$$\lambda(t \mid \mathbf{h}_i^{(L)}) = \lambda_0(t) \exp\left(\boldsymbol{\beta}^\top \mathbf{h}_i^{(L)}\right). \tag{7}$$

All metrics are computed from out-of-fold predictions. For KM stratification, the high/low-risk threshold is determined by the median risk in each training fold and applied to its held-out fold before pooling predictions. We evaluate using: (i) Harrell's C-index (Harrell et al. (1982)), (ii) cumulative/dynamic time-dependent AUC at horizons $\Delta t \in \{12, 18, 24, 30, 36\}$ months, and (iii) Kaplan-Meier (Kaplan & Meier (1958)) stratification with log-rank testing. Note that AUC@Δt at landmark t_L predicts conversion by $t_L + \Delta t$ from study baseline.

Concordance index: Table A14 reports C-index across landmarks. The C-index measures whether the model correctly ranks subject pairs by survival time, where higher predicted risk should correspond to earlier conversion. Performance generally increases with landmark time from 69.1% at baseline to 76.7% at 24m, indicating that additional longitudinal history improves prognostic ranking.

Table A14: C-index across landmarks (mean$_{\pm\text{std}}$, 5-fold CV).

Landmark	N at risk	Events	Censored	C-index (%)
0m	640	235	405	$69.1_{\pm 4.7}$
12m	555	174	381	$70.6_{\pm 6.1}$
24m	419	77	342	$76.7_{\pm 8.3}$

Table A15: Time-dependent AUC (%) across landmarks and prediction horizons (mean$_{\pm\text{std}}$, 5-fold CV). All results use frozen SBT embeddings without clinical covariates.

Landmark	AUC@12m	AUC@18m	AUC@24m	AUC@30m	AUC@36m
0m	$66.2_{\pm 5.4}$	$70.3_{\pm 4.2}$	$72.6_{\pm 4.8}$	$72.9_{\pm 7.1}$	$72.9_{\pm 7.0}$
18m	$69.5_{\pm 8.6}$	$70.1_{\pm 8.6}$	$75.5_{\pm 9.0}$	$75.2_{\pm 8.8}$	$75.3_{\pm 11.0}$
24m	$77.4_{\pm 15.6}$	$79.3_{\pm 8.1}$	$79.4_{\pm 7.5}$	$79.8_{\pm 8.5}$	$80.6_{\pm 8.8}$

Time-dependent AUC: While the C-index evaluates overall ranking, time-dependent AUC assesses discrimination at specific future horizons: can the predicted risk separate subjects who convert within Δt months from those who remain event-free beyond that horizon? Table A15 reports AUC across all landmark–horizon combinations. Baseline from short-term prediction (AUC@12m $\sim$66%), improves to $\sim$73% at longer horizons. Later landmarks achieve consistently higher AUCs: 70–75% at 18m and 77–81% at 24m. Comparing the same absolute endpoint across landmarks (e.g., conversion by month 42 from baseline: landmark 18m AUC@24m $= 75.5\%$ vs. landmark 24m AUC@18m $= 79.3\%$) confirms that additional MRI history improves prediction of the same clinical outcome.

Kaplan-Meier stratification: Finally, we assessed whether the continuous risk scores translate to clinically meaningful patient-group separation. Figure A6 shows KM curves at landmarks 0, 6, 12, 18, and 24 months. The high-risk group exhibits rapid decline in conversion-free survival probability soon after each landmark, while the low-risk group maintains substantially higher event-free probability. Table A16 quantifies the separation: the event ratio between high- and low-risk groups increases monotonically from 2.3:1 at 6m to 6.0:1 at 24m, with all log-rank p-values below 10^{-11}. At landmark 24m, only 5% of low-risk subjects converted versus $\sim$31% in the high-risk group, the low-risk survival curve remains above 0.90 throughout follow-up.

Overall, these results suggest that SBT representations support a time-to-event formulation of MCI prognosis beyond static sMCI/pMCI classification. The model provides meaningful early risk stratification, while additional longitudinal history appears to refine prognostic ranking and horizon-specific discrimination under censoring-aware evaluation.

D.4 Confounder Analysis

We examined confounding at two complementary levels on the ADNI dataset (AD vs CN task), which covers a wide variability of confounders like sex, manufacturer, age, and field-strength. First, we analyzed the *frozen parcel-level embeddings* produced by the parcel encoder to determine whether demographic and acquisition variables remain decodable from the learned representation and whether diagnostic separability persists after explicitly attenuating these factors. Second, we analyze the final model predictions to determine whether representation-level dependence translates into systematic differences in classification performance across subgroups. We consider age, sex, scanner manufacturer, field strength, and acquisition site. All results are reported as mean $\pm$ standard deviation across the 5 folds. Note that the number of sites in the ADNI dataset is varied and high. There are certain sites that contributed very few scans in ADNI cohort. Hence, only BACC results could be reported in Table A17.

Embedding-level analysis: For the representation analysis, parcel embeddings from each scan were stacked and applied PCA to get a single frozen scan-level embedding to avoid loss of embedding information during aggregations. Linear probes were trained on the training-fold embeddings only. Using these embeddings, we evaluated: (i) diagnosis prediction, (ii) prediction of individ-

Table A16: KM stratification summary across landmarks. Event ratio = high-risk events / low-risk events.

Landmark	N	High n	Low n	High events	Low events	Ratio	Log-rank p
6m	640	322	318	164	71	2.3:1	1.0×10^{-15}
12m	555	279	276	128	46	2.8:1	8.8×10^{-15}
18m	507	255	252	101	33	3.1:1	1.3×10^{-11}
24m	419	211	208	66	11	6.0:1	8.7×10^{-12}

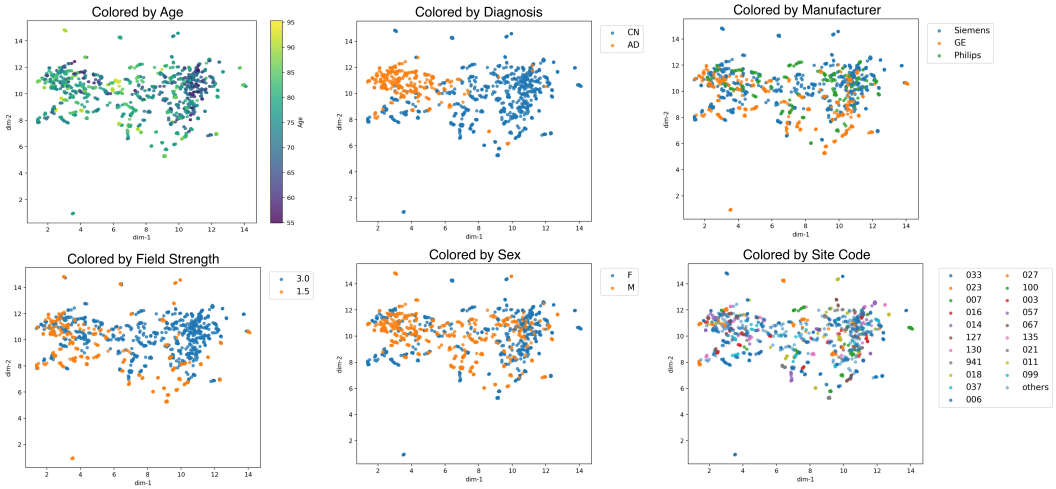

Figure A7: UMAPs of frozen scan-level embeddings from the parcel encoder on the test set, coloured by diagnosis, demographic or acquisition variables, including age, sex, field strength, manufacturer, and site.

ual confounders, and (iii) a diagnosis baseline based only on confound metadata. To ensure that our model's performance was driven by true pathological signatures rather than demographic or acquisition biases, we performed linear residualization. By regressing out the influence of available confounds from the embeddings, we created a 'clean' latent space and then re-evaluated the predictability of both the diagnosis and the confounds. This suggests that the diagnostic separability of the representation is not primarily explained by the linearly removable component associated with these measured demographic and acquisition variables.

The embedding-based diagnosis probe achieved strong performance (Table A17), confirming that the frozen representation preserves substantial disease-relevant information. In contrast, a diagnosis model trained only on metadata confounders performed worse (Table A19), indicating that the diagnostic signal captured by the embeddings cannot be explained by demographic and acquisition variables alone. To examine whether the learned representation encodes diagnostic signal independently of demographic and acquisition-related confounders, we performed a series of post-hoc analyses.

First, we applied linear residualization by fitting a ridge regression from metadata confounders (age, sex, manufacturer, field strength, and site) to the embedding vectors on the training split, and subtracting the predicted confound component from both splits. Table A17 evaluates the prediction of the respective confounder/diagnosis before and after residualization. Nuisance decodability dropped sharply following this adjustment: sex, manufacturer, and field-strength probes showed degraded performance, age-prediction error increased substantially, and the age correlation reversed sign. This reflects strong attenuation of confound structure. In contrast, diagnostic performance remained unchanged, suggesting that diagnostic separability is not primarily explained by the linearly removable component associated with these measured variables. Second, we assessed statistical dependence (non-linearities) between the embedding vectors and the confounder label space using the Hilbert-Schmidt Independence Criterion (HSIC) (Table A17). HSIC, a kernel-based measure of statistical dependence, makes no assumptions about the functional form of the relationship between variables and equals zero if and only if the two distributions are statistically independent, making it a prin-

Table A17: Diagnosis & confound predictability from embeddings, and embedding confound dependence before and after confound residualization. Results are reported as mean$_{\pm\mathrm{std}}$ over 5 folds. Classification tasks use BACC and AUC, age uses MAE, and Pearson r, and dependence is measured using HSIC. Each row in the table corresponds to a separate prediction task, where the target is the diagnosis/confounder itself (for example, sex, field strength, or manufacturer), evaluated before and after residualization.

Task	Metric	Before	After	Δ
Diagnosis	BACC	$0.84_{\pm 0.04}$	$0.84_{\pm 0.04}$	$-0.00_{\pm 0.02}$
	AUC	$0.91_{\pm 0.03}$	$0.91_{\pm 0.03}$	$-0.00_{\pm 0.01}$
Age	MAE ↓	$4.70_{\pm 0.18}$	$5.94_{\pm 0.04}$	$+1.24_{\pm 0.19}$
	Pearson r	$0.63_{\pm 0.02}$	$-0.52_{\pm 0.06}$	$-1.15_{\pm 0.06}$
Sex	BACC	$0.85_{\pm 0.01}$	$0.22_{\pm 0.03}$	$-0.63_{\pm 0.03}$
	AUC	$0.93_{\pm 0.01}$	$0.14_{\pm 0.03}$	$-0.79_{\pm 0.03}$
Manufacturer	BACC	$0.83_{\pm 0.04}$	$0.08_{\pm 0.02}$	$-0.76_{\pm 0.03}$
	AUC	$0.95_{\pm 0.02}$	$0.20_{\pm 0.04}$	$-0.75_{\pm 0.02}$
Field Strength	BACC	$0.96_{\pm 0.01}$	$0.35_{\pm 0.02}$	$-0.61_{\pm 0.02}$
	AUC	$0.99_{\pm 0.00}$	$0.28_{\pm 0.03}$	$-0.71_{\pm 0.03}$
Site	BACC	$0.05_{\pm 0.02}$	$0.00_{\pm 0.00}$	$-0.05_{\pm 0.02}$
Sex	HSIC	$0.003_{\pm 0.001}$	$0.001_{\pm 0.000}$	$-0.001_{\pm 0.000}$
Manufacturer	HSIC	$0.002_{\pm 0.001}$	$0.001_{\pm 0.000}$	$-0.001_{\pm 0.001}$
Field Strength	HSIC	$0.008_{\pm 0.003}$	$0.001_{\pm 0.000}$	$-0.007_{\pm 0.003}$

Table A18: Diagnosis performance from embeddings across demographic and acquisition subgroups.

Group	Subgroup	BACC	AUC
Sex	Female	$0.84_{\pm 0.03}$	$0.91_{\pm 0.03}$
	Male	$0.84_{\pm 0.05}$	$0.90_{\pm 0.04}$
Manufacturer	GE	$0.83_{\pm 0.02}$	$0.90_{\pm 0.02}$
	Philips	$0.81_{\pm 0.09}$	$0.89_{\pm 0.08}$
	Siemens	$0.85_{\pm 0.04}$	$0.92_{\pm 0.04}$
Field Strength	1.5T	$0.81_{\pm 0.05}$	$0.89_{\pm 0.04}$
	3.0T	$0.85_{\pm 0.04}$	$0.91_{\pm 0.04}$
Age bin	≤ 65	$0.84_{\pm 0.11}$	$0.90_{\pm 0.05}$
	66–75	$0.86_{\pm 0.03}$	$0.93_{\pm 0.02}$
	76–85	$0.82_{\pm 0.04}$	$0.90_{\pm 0.04}$
	> 85	$0.79_{\pm 0.04}$	$0.90_{\pm 0.04}$

Table A19: Diagnosis prediction using only metadata confounders.

Model	BACC	AUC
Metadata confounders only	$0.56_{\pm 0.02}$	$0.58_{\pm 0.03}$

Table A20: Equalized odds gaps from embeddings across confounder groups.

Confounder	TPR gap	FPR gap
Sex	$0.07_{\pm 0.04}$	$0.04_{\pm 0.02}$
Manufacturer	$0.14_{\pm 0.10}$	$0.09_{\pm 0.03}$
Field Strength	$0.05_{\pm 0.03}$	$0.11_{\pm 0.06}$
Age bin	$0.21_{\pm 0.16}$	$0.20_{\pm 0.04}$

Table A21: Subgroup performance of the final diagnosis predictions across demographic and acquisition factors. Values are reported as mean$_{\pm\mathrm{std}}$ over 5 folds.

Group	Subgroup	BACC	AUC
Sex	Female	$0.82_{\pm 0.02}$	$0.91_{\pm 0.02}$
	Male	$0.84_{\pm 0.06}$	$0.91_{\pm 0.05}$
Manufacturer	GE	$0.81_{\pm 0.03}$	$0.89_{\pm 0.02}$
	Philips	$0.82_{\pm 0.08}$	$0.90_{\pm 0.09}$
	Siemens	$0.85_{\pm 0.03}$	$0.92_{\pm 0.04}$
Field Strength	1.5T	$0.82_{\pm 0.04}$	$0.89_{\pm 0.04}$
	3.0T	$0.83_{\pm 0.04}$	$0.92_{\pm 0.03}$
Age Bin	≤ 65	$0.73_{\pm 0.12}$	$0.87_{\pm 0.09}$
	66–75	$0.85_{\pm 0.02}$	$0.95_{\pm 0.02}$
	76–85	$0.84_{\pm 0.05}$	$0.90_{\pm 0.04}$
	> 85	$0.82_{\pm 0.03}$	$0.86_{\pm 0.04}$

Table A22: Prediction-level fairness summaries measured using Equalized Odds gaps across confounder groups. Lower values indicate smaller disparity.

Confounder	FPR Gap	TPR Gap
Sex	$0.03_{\pm 0.02}$	$0.11_{\pm 0.05}$
Manufacturer	$0.06_{\pm 0.03}$	$0.16_{\pm 0.09}$
Field Strength	$0.06_{\pm 0.04}$	$0.07_{\pm 0.04}$
Age Bin	$0.16_{\pm 0.03}$	$0.19_{\pm 0.22}$

cipled and complete test. We applied an RBF kernel over the embedding space and a delta kernel over the categorical confounder labels. Across all tested confounders, the Hilbert–Schmidt Independence Criterion (HSIC) values were uniformly small and decreased further after residualization

(Table A17). Overall, the embeddings exhibit negligible non-linear confound-related dependence for diagnosis, both before and after linear residualization. Third, we evaluated diagnostic performance across demographic and acquisition subgroups, (Table A18). Performance remained broadly consistent across sex, manufacturer, field strength, and age strata, with only modest variation. This suggests diagnostic behavior does not disproportionately depend on specific demographic or scanner subpopulations. Finally, we evaluated equalized odds gaps. This metric directly quantifies whether the model's error structure is consistent across demographic and acquisition conditions, providing a more clinically grounded assessment of fairness. Equalized-odds gaps remained small across sex and field strength , confirming stable prediction behaviour across these demographic and acquisition axes in Table A20. Thus indicating that prediction behavior is broadly stable across demographic and acquisition conditions.

Taken together, these analyses indicate that although the representation contains measurable demographic and acquisition-related structure, diagnostic performance cannot be explained primarily by these confounders. Diagnostic separability remains stable after confound attenuation, confounders alone provide only weak diagnostic signal, and performance remains broadly consistent across major demographic and acquisition strata. This qualitative trend is further supported visually by the embedding projections in Fig. A7, where diagnosis remains well structured despite variation across demographic and acquisition-related factors.

Prediction-level analysis: To complement the representation analysis, we next examined the final model outputs at the level of predicted probabilities. We evaluated subgroup behaviour across sex, manufacturer, field strength, and age bins using subgroup performance and Equalized Odds gaps.

The subgroup analysis in Tables A21 and A22 indicates that the final predictions are broadly stable across demographic and acquisition-related factors. Performance remains comparable across sex, scanner manufacturer, field strength, and age bins, suggesting good generalization across clinically and technically relevant subgroups. The final predictions are not dominated by any single confounding axis.

D.5 Robustness to preprocessing perturbations.

We evaluate preprocessing-related robustness by perturbing the bias-field log-amplitude, contrast scale, Gibbs k-space truncation fraction, image rotation, and translation for the SBT_{SS} model on the AD/CN task. Results are averaged over five folds and summarized in Table A23.

Table A23: Robustness of SBT_{SS} to preprocessing perturbations on ADNI AD/CN classification.

Perturbation	Level	AUC (%)	ΔAUC
Original (SBT_{SS})	–	91.18	–
Bias field	0.1	90.60	−0.58
Bias field	0.2	90.54	−0.64
Bias field	0.3	90.32	−0.86
Contrast	0.5	87.80	−3.38
Contrast	1.5	92.29	+1.11
Contrast	2.0	92.72	+1.54
Gibbs	0.1	90.84	−0.34
Gibbs	0.2	90.91	−0.27
Gibbs	0.3	90.85	−0.33
Rotation	1°	90.01	−1.17
Rotation	2°	88.79	−2.39
Translation	1 voxel	88.43	−2.75
Translation	2 voxels	87.19	−3.99

The model is largely stable under moderate intensity inhomogeneity, Gibbs artifacts, and small geometric perturbations, with degradation becoming more noticeable only for stronger contrast changes or larger spatial misalignment. Importantly, the probabilistic atlas is registered to each preprocessed scan, and parcel membership is computed at the 8^3-voxel patch level rather than requiring exact voxel-wise correspondence with parcel boundaries. Consequently, small residual registration errors need not induce abrupt changes in parcel assignment, since boundary patches can retain graded membership across adjacent parcels.

E Explainability Details and Additional Analysis

E.1 Cohort-level analysis

Cohort level Age-binned parcel saliencies: For each test scan, we compute a normalized, non-negative parcel saliency vector $\hat{\mathbf{s}} \in \mathbb{R}^P$ and average these across the cohort $\mathcal{I}$ within diagnosis d and age bin b, $\bar{\mathbf{s}}_{d,b} = \frac{1}{|\mathcal{I}_{d,b}|} \sum_{i \in \mathcal{I}_{d,b}} \hat{\mathbf{s}}^{(i)}$. A sagittal and coronal view of this volume is visualized across all the bins in Fig A8A for AD, pMCI, and sMCI test subjects. Fig A8B shows top-parcel rank trajectories over age. Cohort-level explainability analysis reveals that our model's classification decisions

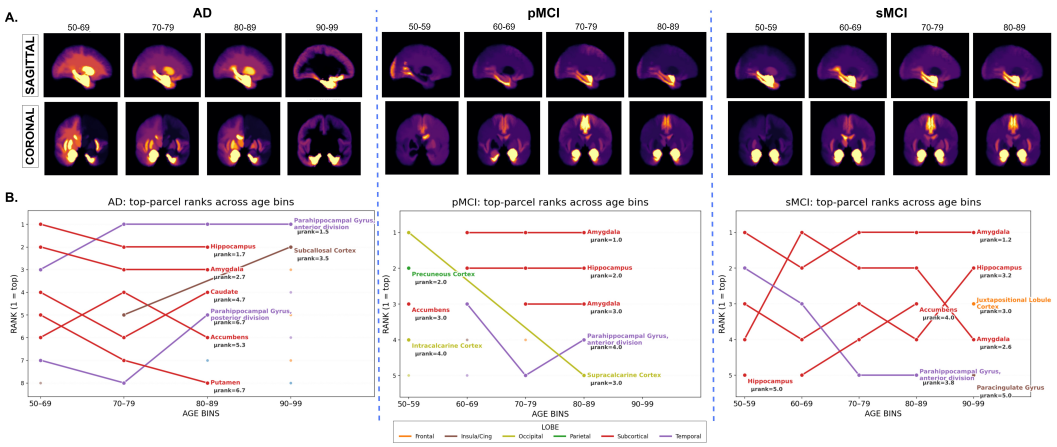

Figure A8: **A. Cohort-level parcel saliency in AD/pMCI/sMCI subjects (test set) across age strata.** Each panel is a sagittal (slice:64) / coronal (slice:122) view of a *voxel-space heatmap* constructed using *cohort-mean parcel saliencies*. Bright regions indicate parcels that contributed most to the model's prediction within the corresponding age bin. **B. Top-parcel ranks across age bins.** The chart shows *relative* importance of parcels across the age bins. Parcels, colored by lobe, are ranked within each age bin by absolute mean saliency score $|\bar{s}|$ (1=top). μrank indicates the mean rank over all bins.

are anchored in anatomically plausible and disease-consistent patterns. These cohort-aggregated maps provide an auditable summary of the anatomical regions most responsible for predictions across diagnoses and age.

Per-scan parcel saliency: Given a trained model and a target class logit f_y, let scores_p be the pre-softmax gate score for parcel $p \in \{1, \ldots, P\}$. Define parcel weights $w_p = \text{softmax}(\text{scores}/\tau)_p$. Then the gradient-based parcel saliency becomes:

$$s_p = \text{ReLU}\Big(w_p \cdot \frac{\partial f_y}{\partial \text{scores}_p}\Big), \qquad \hat{s}_p = \frac{s_p}{\sum_{q=1}^{P} s_q + \varepsilon},$$

yielding a per-scan vector $\hat{\mathbf{s}} \in \mathbb{R}^P$ that is nonnegative and L_1-normalized.

Age binning and cohort mean: Assign each scan i to an age bin b (in years). For diagnosis d and bin b, compute the cohort mean

$$\bar{s}_{d,b} = \frac{1}{|\mathcal{I}_{d,b}|} \sum_{i \in \mathcal{I}_{d,b}} \hat{s}^{(i)}.$$

Atlas-weighted 3D maps: A *voxel-space heatmap* constructed by linearly mixing Harvard–Oxford (HO) probabilistic parcel maps with *cohort-mean parcel saliencies*. Using the Harvard–Oxford probabilistic atlas $\{A_k(x)\}_{k=1}^{P}$ (voxel x), form the explanation volume

$$H_{d,b}(x) = \sum_{k=1}^{P} \bar{s}_{d,b}[k]\, A_k(x).$$

For Fig A8, we apply light smoothing and robust normalization (shared across panels in a row) and render sagittal/coronal views. Within each bin b, rank parcels by descending $|\bar{s}_{d,b}[k]|$ and keep the top N with integer ranks (1=highest).

Interpreting Fig A8. Cohort-level explainability analysis reveals that our model's classification decisions are anchored in anatomically plausible and disease- consistent patterns (Fig A8). Across age-stratified bins, gradient-based attribution concentrates persistently on medial temporal structures such as hippocampus (Pennanen et al. (2004)), amygdala (Poulin et al. (2011)), and parahippocampal gyrus in AD and pMCI (Pennanen et al. (2004)), mirroring the canonical Braak staging trajectory (Braak & Braak (1991); Braak et al. (2006); Planche et al. (2022); Lam et al. (2013);

Blinkouskaya & Weickenmeier (2021); Poulin et al. (2011)) and established MRI volumetry findings (Braak & Braak (1991); Braak et al. (2006); Jack Jr et al. (1999); Dickerson et al. (2009)). In pMCI, the relative ranking of salient regions changes across age strata: strong limbic emphasis (Ferreira et al. (2017)) in younger bins (50-69 years) transitions toward posterior association cortices (precuneus, supracalcarine) (Doody et al. (2010); Jellinger (2022)) in older bins (70-89 years). In contrast, sMCI shows unstable Medial Temporal Lobe attribution (hippocampus), reflecting its heterogeneous, non-progressive etiology (Petersen (2004); Chételat et al. (2005)).

Note for Fig A8: Parcel labels besides denote parcels with high persistence (appear in many bins with large summed $|\bar{s}|$). Parcels appearing twice in the chart correspond to the left and right hemispheres of the Brain.

E.2 Attention Rollout Construction

Context: For SBT_{SL}, each visit i produces parcel embeddings $\{z_{p,i}\}_{p=1}^{P}$ and post-softmax parcel weights $\alpha_{p,i}$, which are aggregated into a scan representation $s_i = \sum_{p=1}^{P} \alpha_{p,i} z_{p,i}$. The Longitudinal Transformer subsequently models the sequence $(s_1, \ldots, s_T)$.

Attention rollout heatmaps: Attention rollout is constructed hierarchically following the SBT factorization. We first compute attention rollout independently within each anatomical parcel. Let $A_{\ell,i}^{(p)}$ denote Parcel Encoder attention matrix at layer ℓ for parcel p at visit i. The accumulated rollout is

$$R_i^{(p)} = \prod_{\ell=1}^{L} \text{row_norm}\left(A_{\ell,i}^{(p)} + I\right),$$

from which the parcel-local token importance is

$$r_i^{(p)}[u] = \frac{1}{K_p} \sum_{v=1}^{K_p} R_i^{(p)}[v, u].$$

The parcel-local rollouts are then projected back to the global token lattice using the soft token–parcel memberships $M_i[k, p]$ and the corresponding parcel weights $\alpha_{p,i}$:

$$r_{\text{scan},i}[k] = \sum_{p=1}^{P} M_i[k, p]\, \alpha_{p,i}\, r_i^{(p)}[u_p(k)],$$

where $u_p(k)$ denotes the local index of global token k within parcel p.

To incorporate longitudinal information, we compute temporal rollout from the Longitudinal Transformer,

$$r_{\text{time}}[i] = \frac{1}{t} \sum_{u=1}^{t} \left(\prod_{b=1}^{B} \text{row_norm}(T_b + I) \right)[u, i],$$

where $\{T_b\}_{b=1}^{B}$ are the temporal attention matrices over the t available visits. The final token-level visualization for visit i is

$$h_i[k] = r_{\text{time}}[i]\, r_{\text{scan},i}[k].$$

Thus, the visualization first captures information routing within each anatomical parcel, then combines these parcel-local maps according to the learned parcel weights and memberships, and finally modulates each visit by its longitudinal attention contribution.

E.3 Additional Subject-Level Examples and Failure Analysis

Figure A9 provides additional subject-level examples across ADNI, AIBL, and MIRIAD, complementing the cohort-level analysis above.

Interpreting Fig 7C. Subject-level longitudinal parcel attribution trajectories demonstrate coherence with cohort-level analysis. AD exhibits sustained, convergent Medial Temporal Lobe (MTL)

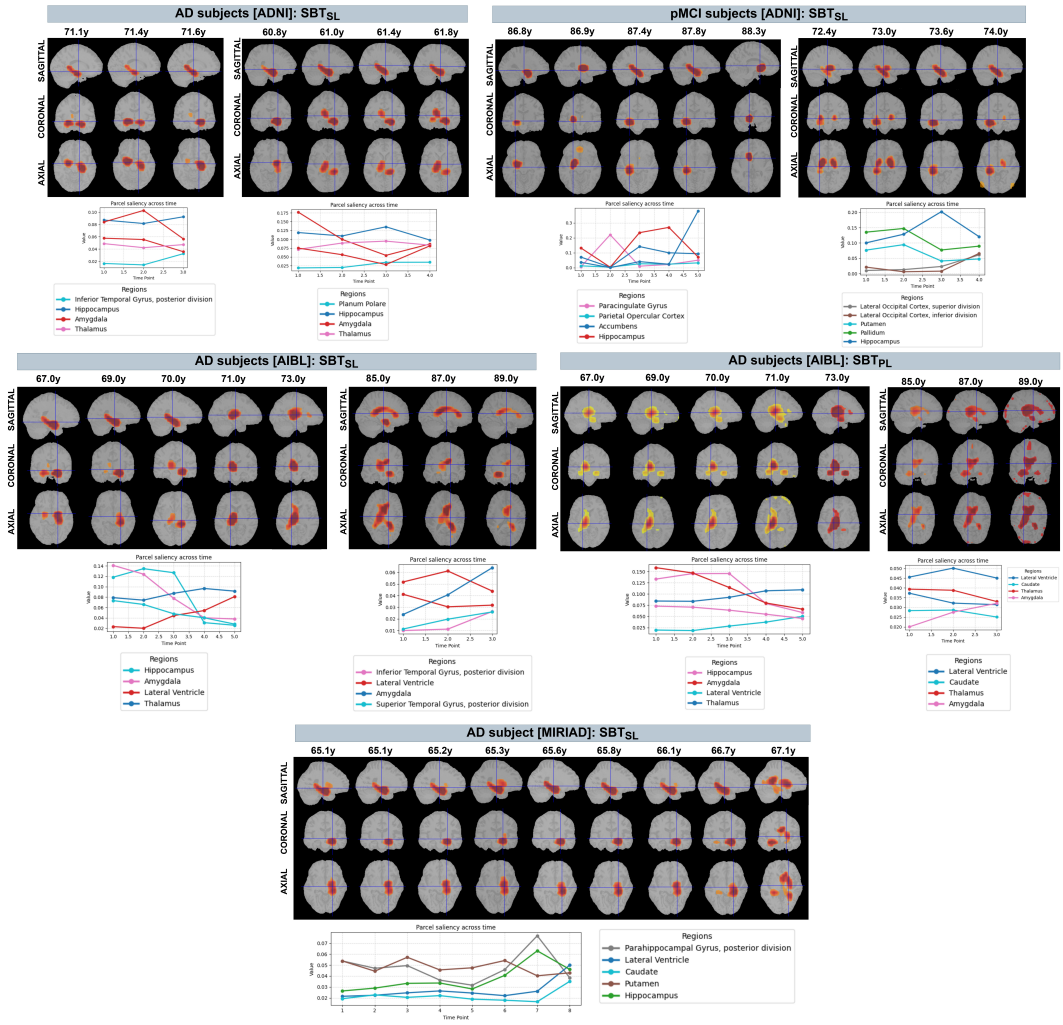

Figure A9: Attention Rollout visualizations across all three views (coronal, sagittal, and axial) along with their parcel saliency plots of correctly predicted longitudinal subjects of ADNI, AIBL and MIRIAD dataset.

weighting, such as amygdala and Hippocampus, rising to dominate and thalamus peaking (at time-point 4), mirroring progressive limbic neurodegeneration and secondary degeneration (Braak et al. (2006); Poulin et al. (2011); De Jong et al. (2008)). pMCI displays the canonical hippocampal and Medial Temporal Gyrus, recapitulating maximal atrophy rates during MCI-to-AD conversion (Jack Jr et al. (2000); Henneman et al. (2009)). It is followed by posterior/lateral cortical parcels (intracalcarine) progressively reflecting disease progression (Whitwell et al. (2008); Buckner et al. (2008)). In contrast, sMCI shows high-amplitude oscillations without stable rankings, i.e, hippocampus fluctuates, striatal parcels intermittently dominate, thus reflecting etiological heterogeneity (Petersen (2004); Stern (2012)). These subject-level patterns are qualitatively consistent with the cohort-level analysis, while illustrating that the regional evidence used by SBT can evolve across longitudinal observations.

ADNI analysis (Fig A9): In AD cases, SBT consistently allocates saliency to canonical medial temporal hubs such as hippocampus, amygdala, and posterior inferior temporal cortex, mirroring established staging and reinforcing that the framework recovers core loci of Alzheimer's pathology in a visit-consistent manner. In pMCI, the trajectories emphasize limbic and posterior association cortices whose evolving engagement is characteristic of elevated conversion risk. Notably, older converters often exhibit progressive involvement of paracingulate and parietal opercular regions,

whereas younger converters show a complementary posterior–subcortical pattern involving lateral occipital parcels and limbic nuclei.

AIBL/MIRIAD analysis (Fig A9): In AIBL, hippocampal saliency remains persistently high with increasing emphasis on hippocampus, amygdala, and lateral ventricles, the atrophy-consistent signals, whereas MIRIAD subject frequently accentuates the posterior division of the parahippocampal gyrus. In the SBT_{PL} variant's attention rollouts, clinically aligned parcel activations (i.e., hippocampus, Amygdala, Lateral ventricle) are visible across the scans of all time points, but with increasing focus towards the latest scans. Together, these visualizations provide qualitative evidence that the regional information emphasized by SBT overlaps with anatomically plausible and AD-relevant structures across independent cohorts.

Failure Case: Fig A10 shows a failure case of a misclassified CN subject predicted as AD. The false positive is accompanied by a spatially coherent attribution pattern involving medial-temporal and subcortical regions rather than diffuse attribution across the scan.

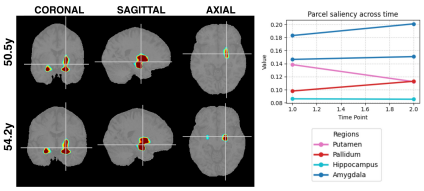

Figure A10: Attention rollouts with parcel-saliency plots of misclassified ADNI subject.

Critically, across all cohorts, variants, and diagnostic categories, attention rollouts and gradient saliency maps are largely concentrated in anatomically plausible regions. Their spatial agreement provides complementary qualitative evidence that information routing and prediction sensitivity are localized to similar structures. Together, the controlled and clinical analyses indicate that SBT preserves localized anatomical evidence through longitudinal modeling, while the clinical visualizations show qualitative concordance with established AD-related structural patterns without requiring regional disease labels.

E.4 SBT_{SL} vs SBT_{PL} on Controlled Progression

Figure A11 provides a controlled comparison of the longitudinal parcel weights produced by SBT_{SL} and SBT_{PL}, for which the location and severity of regional abnormalities are known exactly.

Spatial distribution. Both variants preserve anatomically localized progression information, but exhibit different weighting patterns. SBT_{SL} generally distributes relevance across multiple concurrently affected parcels, more closely reflecting the broader spatial distribution of evolving abnormalities. SBT_{PL} is comparatively more selective, often concentrating relevance on a smaller subset of strongly informative regional trajectories. This is consistent with their different longitudinal factorizations: SBT_{SL} aggregates regional information into a scan representation before temporal modeling, whereas SBT_{PL} retains individual parcel trajectories throughout the temporal stage.

Sensitivity before conversion. Importantly, both variants respond during the CN* interval, when localized abnormalities are already present but the cumulative subject-level burden has not yet crossed the ATR conversion threshold. Changes in parcel weighting are therefore visible before the observed class transition, indicating that both representations remain sensitive to emerging regional abnormalities during the pre-conversion period.

Taken together, these results motivate SBT_{SL} as the default longitudinal formulation when both distributed anatomical evidence and temporal efficiency are desired, while SBT_{PL} provides a complementary alternative for explicitly examining region-specific longitudinal trajectories.

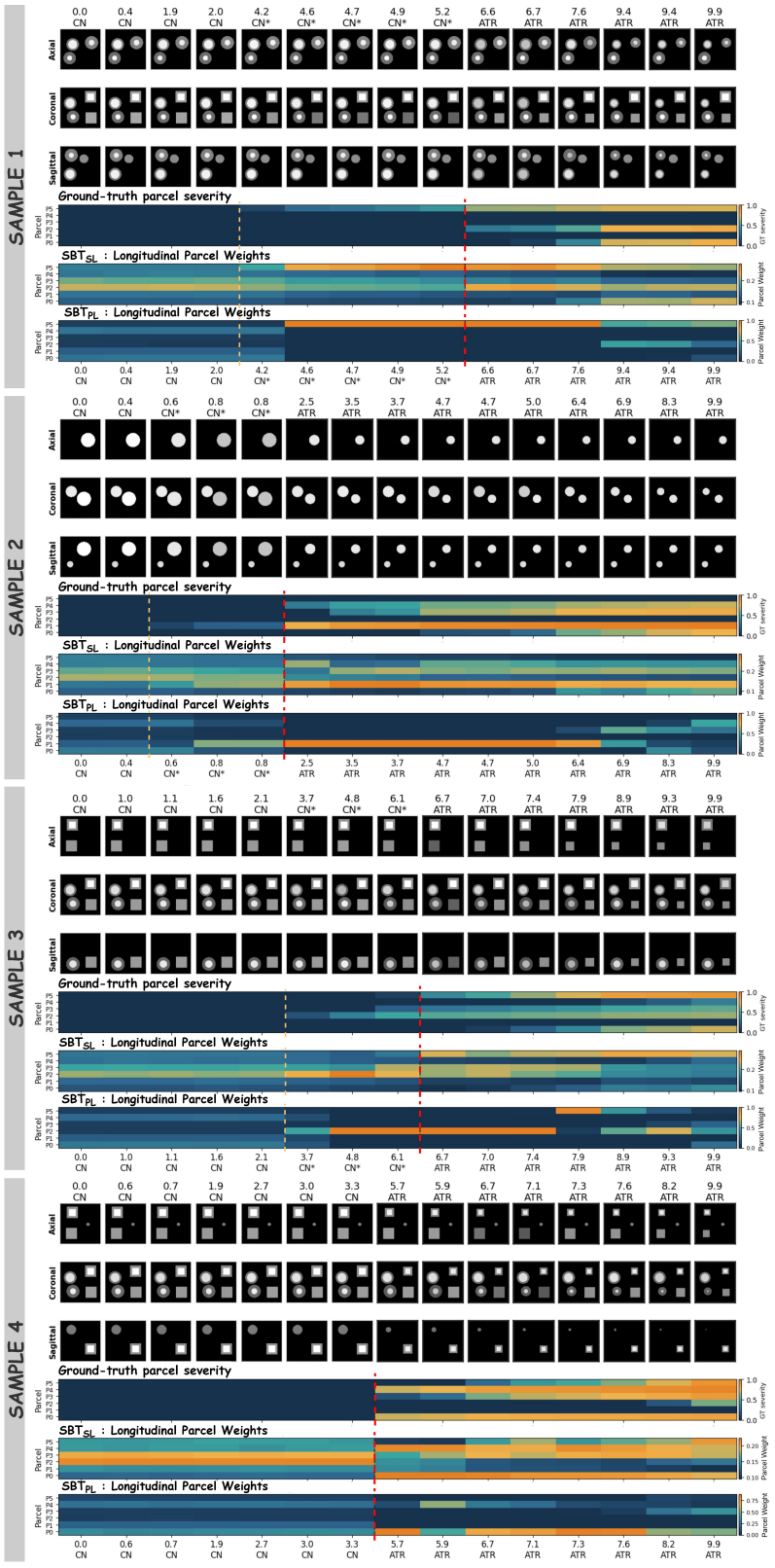

Figure A11: **Scan-level versus parcel-level longitudinal reasoning under controlled progression.** For each synthetic subject, the upper heatmap shows the known parcel-wise abnormality severity over time, followed by longitudinal parcel weights from $\mathrm{SBT}_{\mathrm{SL}}$ and $\mathrm{SBT}_{\mathrm{PL}}$.