LONGMOE: LONGITUDINAL MULTIMODAL LEARNING VIA TRAJECTORY-AWARE MIXTURE-OF-EXPERTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal clinical learning is increasingly important for integrating diverse patient data, including imaging, text, and personalised health records. However, it faces two fundamental challenges: i) modality missingness, where arbitrary subsets of modalities are unavailable at a given patient visit, ii) longitudinal dynamics, where the diagnostic significance of an observation depends on the patient's evolving disease trajectory over time. Existing methods address these challenges in isolation: missing-modality frameworks treat each visit as an independent static snapshot and discard temporal context, while longitudinal models often assume complete modality availability and degrade under systematic modality incompleteness. We propose LONGMOE (Longitudinal Mixture-of-Experts), the unified framework to jointly address both challenges. LONGMOE combines a context-aware imputation module with an attentional tokenization module that captures frequency-domain temporal patterns across irregular visit sequences, a trajectory-aware encoder for modeling disease progression, and context-conditioned Sparse MoE routing for patient-specific expert selection. Experiments on ADNI, OASIS-3, and MIMIC-IV show that LONGMOE improves robustness under missing or weak contemporaneous modalities and remains competitive in full-modality settings, establishing a strong foundation for longitudinally-aware multimodal clinical learning.

1 INTRODUCTION

Multimodal learning has become a fundamental challenge in clinical decision support, where integrating complementary data sources yields richer and more reliable predictions than any single modality alone (Baltrušaitis et al., 2018; Liang et al., 2021). In Alzheimer's disease (AD), this integrative capacity is clinically necessary, i.e., pathology manifests concurrently across structural neuroimaging, functional biomarkers, cognitive assessments, and genetic risk profiles, each contributing distinct and partially non-overlapping diagnostic signal (Jack and Holtzman, 2013; Márquez and Yassa, 2019; Papassotiropoulos et al., 2006). In practice, real-world cohorts exhibit pervasive *modality missingness*: any given patient visit yields only a subset of the full modality panel due to heterogeneous acquisition protocols, resource constraints, and time-dependent assay availability (Yun et al., 2024; Han et al., 2024). Existing approaches either restrict training to fully observed patients or apply context-free imputation strategies such as zero-padding and global mean substitution (Zhang et al., 2022; Wang et al., 2024). A compounding difficulty is that AD patients are not a homogeneous population, i.e., individuals differ systematically in genetic profiles, comorbidities, and progression rates, so that the relative informativeness of each modality varies substantially across subgroups (Jack and Holtzman, 2013). With L modalities, a model should handle up to $2^L - 1$ distinct observation patterns, each inducing a different information structure over which fusion must be performed. A monolithic fusion network is structurally unsuited to this setting (Fedus et al., 2022), motivating a Sparse Mixture-of-Experts (SMoE) architecture (Shazeer et al., 2017; Jacobs et al., 1991) in which a learned router dynamically activates a sparse subset of specialised expert networks conditioned on the observed modality combination, providing modality-pattern-aware and subgroup-aware fusion without the cost of a fully dense model.

Alzheimer's disease is a slowly progressive condition unfolding over years to decades, and the diagnostic significance of any observation is inseparable from its longitudinal context (Weiner et al., 2010; Venugopalan et al., 2021). What matters clinically is not the absolute value of a biomarker at a single visit but its position along the patient's individual trajectory of decline. For example, a

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.

moderate hippocampal volume reading carries an entirely different prognostic weight depending on whether it represents a stable plateau or a steep atrophy curve (Jack and Holtzman, 2013). Therefore, *the rate and pattern of biomarker change* across successive visits is a substantially stronger predictor of disease stage than any cross-sectional measurement (Weiner et al., 2010; Lee et al., 2019). GRU-based models (Cho et al., 2014; Lee et al., 2019) capture longitudinal dependencies but require complete modality inputs at every visit and fail under structured missingness, while Transformer-based models (Vaswani et al., 2017; Devlin et al., 2019) offer greater expressivity but have not been adapted to the joint challenge of irregular visit spacing, arbitrary modality combinations, and population-level heterogeneity.

Recent work Flex-MoE (Yun et al., 2024) represents the closest prior work, introducing an SMoE framework with a learnable missing-modality bank and a two-stage routing strategy that accommodates arbitrary modality subsets. Nevertheless, Flex-MoE and all concurrent missing-modality methods (Han et al., 2024; Tsai et al., 2019; Zadeh et al., 2017) treat each clinical visit as an independent static snapshot, that means the routing mechanism receives no information about the patient's prior visits, biomarker history, or evolving modality availability, and is therefore structurally incapable of trajectory-aware fusion. The result is a precise and consequential gap: *no existing model jointly addresses modality missingness and longitudinal dynamics*. We address this gap with LONGMOE, whose contributions are as follows:

- We propose LONGMOE, a unified framework that jointly addresses modality missingness and longitudinal temporal dynamics in clinical multimodal learning.
- We introduce an attentional tokenization module that integrates modality-level cross-attention with continuous-time, multi-frequency positional encoding. This encoding provides a temporal representation for non-uniformly sampled visit sequences.
- Through extensive experiments on ADNI, OASIS-3, and MIMIC-IV, we show that LONGMOE consistently outperforms state-of-the-art baselines under missing contemporaneous modalities.

2 RELATED WORK

Missing Modality Handling in Multimodal Learning. Multimodal learning has been extensively studied across vision, language, and healthcare domains (Baltrušaitis et al., 2018; Liang et al., 2021), with AD-specific work confirming that combining imaging, genetic, and clinical data consistently outperforms any single modality (Venugopalan et al., 2021; Odusami et al., 2023). Canonical fusion architectures including TF (Zadeh et al., 2017), MulT (Tsai et al., 2019), and MAG (Rahman et al., 2020) provide strong baselines but uniformly assume complete modality availability, while global imputation strategies such as zero-padding (Little and Rubin, 2019) and learned methods including ShaSpec (Wang et al., 2024) and mmFormer (Zhang et al., 2022) condition imputation on population statistics rather than individual patient context, limiting effectiveness for rare modality combinations. The most directly relevant prior works are Flex-MoE (Yun et al., 2024) and FuseMoE (Han et al., 2024). Flex-MoE handles arbitrary modality combinations via a learnable missing modality bank and a two-stage SMoE design (G-Router for full-modality; S-Router for incomplete samples). FuseMoE applies dynamic expert input masking for variable modality sets, but both treat each visit as an independent static snapshot, discarding the patient's historical trajectory entirely.

Longitudinal Modelling and Mixture-of-Experts. Longitudinal modelling of disease progression has relied on recurrent architectures (Hochreiter and Schmidhuber, 1997; Cho et al., 2014; Choi et al., 2016), including the GRU-based multimodal model Lee-MMGRU (Lee et al., 2019), which captures temporal dynamics across imaging, clinical, and genetic modalities in AD but requires complete modality observations at every visit, rendering it unable to exploit partially observed patients. Transformer-based sequence models (Vaswani et al., 2017; Devlin et al., 2019) have largely superseded recurrent architectures, with irregular time-series Transformers (Horn et al., 2020; Shukla and Marlin, 2021) demonstrating that attention accommodates irregular sampling via learned positional encodings. The Mixture-of-Experts paradigm (Jacobs et al., 1991; Jordan and Jacobs, 1994) has undergone a resurgence through Sparsely-Gated MoE (Shazeer et al., 2017), Switch Transformers (Fedus et al., 2022), and Mixtral (Jiang et al., 2024). In the multimodal setting, LiMoE (Mustafa et al., 2022), SM^4 (Peng et al., 2024), MMoE (Ma et al., 2018), Mod-Squad (Chen et al., 2023), DSelect-k (Hazimeh et al., 2021), Expert Choice (Zhou et al., 2022), and sparse vision

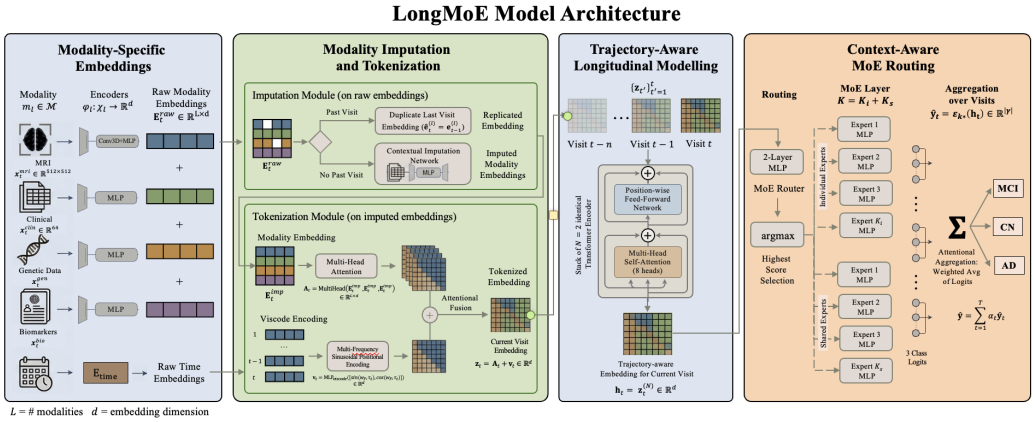


Figure 1: Overview of the LONGMOE architecture, which processes a patient's longitudinal visit sequence through four sequential modules:*(i)* modality-specific embedding, *(ii)* a modality imputation and tokenization, *(iii)* a trajectory-aware longitudinal modeling, and *(iv)* a context-aware MoE routing.

MoE (Riquelme et al., 2021) advanced expert specialisation and load-balanced routing, but none addresses the intersection of longitudinal sequence modelling and missing modality handling on irregular, partially observed clinical visit sequences.

3 PRELIMINARIES

Longitudinal Patient Record. Let a patient's clinical history be represented as an ordered sequence of T visits, $\mathcal{V} = \{v_1, v_2, \ldots, v_T\}$, where each visit v_t occurs at timestamp $\tau_t \in \mathbb{R}_{>0}$. The inter-visit intervals $\Delta_t = \tau_t - \tau_{t-1}$ are irregular and patient-specific, reflecting the non-uniform nature of real-world clinical data collection.

Multimodal Observations. At each visit v_t a subset of L modalities is observed. Let $\mathcal{M} = \{m_1, m_2, \ldots, m_L\}$ denote the full set of available modalities. For the Alzheimer's disease progression modeling, these comprise: *(i)* structural MRI scans $\mathbf{x}_t^{\text{mri}} \in \mathbb{R}^{512\times512}$, *(ii)* clinical assessment scores $\mathbf{x}_t^{\text{clin}} \in \mathbb{R}^{64}$, *(iii)* CSF/biospecimen panels $\mathbf{x}_t^{\text{bio}} \in \mathbb{R}^{b}$, and *(iv)* genetic assay data $\mathbf{x}_t^{\text{gen}} \in \mathbb{R}^{g}$. In practice, not all modalities are observed at every visit. Any given visit may yield only one or two modalities, making static full-input assumptions unrealistic. Let $\mathbf{O}_t \in \{0,1\}^L$ be a binary observation mask at visit t, where $O_{t,l} = 1$ if modality m_l is observed and $O_{t,l} = 0$ otherwise. Given the full longitudinal record of a patient up to and including the current visit v_t, the goal is to predict the diagnostic label $y_t \in \mathcal{Y} = \{\text{CN}, \text{MCI}, \text{AD}\}$, where CN denotes cognitively normal, MCI denotes mild cognitive impairment, and AD denotes Alzheimer's disease.

4 LONGMOE ARCHITECTURE

LONGMOE processes longitudinal multimodal patient data through four sequentially composed modules (Figure 1): *(i)* modality-specific embedding, *(ii)* a modality imputation and tokenization, *(iii)* a trajectory-aware longitudinal modeling, and *(iv)* a context-aware MoE routing.

4.1 MODALITY-SPECIFIC EMBEDDINGS

For each modality $m_l \in \mathcal{M}$ we define an encoder $\phi_l : \mathcal{X}_l \to \mathbb{R}^d$, where $\mathcal{X}_l$ is the native input space of modality l. Encoder architectures are chosen to match the structural properties of each modality:

- *MRI* ($\mathbf{x}_t^{\text{mri}} \in \mathbb{R}^{512\times512}$): A 3D convolutional encoder followed by an MLP projection head captures local volumetric structure and global morphological features $\mathbf{e}_t^{\text{mri}} = \text{MLP}\big(\text{Conv3D}\big(\mathbf{x}_t^{\text{mri}}\big)\big) \in \mathbb{R}^d$.

- *Clinical Scores* ($\mathbf{x}_t^{\text{clin}} \in \mathbb{R}^{64}$), *Biomarkers* ($\mathbf{x}_t^{\text{bio}}$), and *Genetic Data* ($\mathbf{x}_t^{\text{gen}}$): Each is processed by a modality-specific MLP $\mathbf{e}_t^{(l)} = \text{MLP}_l\left(\mathbf{x}_t^{(l)}\right) \in \mathbb{R}^d, \quad l \in \{\text{clin, bio, gen}\}$.
- *Temporal Embedding*: The visit timestamp τ_t is encoded as a learned temporal embedding $\mathbf{E}_{\text{time}} \in \mathbb{R}^d$, capturing both absolute time and the inter-visit interval Δ_t.

The per-modality embeddings at visit t are stacked to form $\mathbf{E}_t^{\text{raw}} = \left[\mathbf{e}_t^{(1)};\ \mathbf{e}_t^{(2)};\ \ldots;\ \mathbf{e}_t^{(L)}\right] \in \mathbb{R}^{L\times d}$, where rows corresponding to unobserved modalities ($O_{t,l} = 0$) are set to zero prior to imputation. All encoders are jointly trained end-to-end with the full model.

4.2 Modality Imputation and Tokenization

Imputation Module. For each missing modality l at visit t (i.e., $O_{t,l} = 0$) we apply one of two imputation strategies conditioned on the availability of past visit information.

Past Visit Available ($t > 1$ and $O_{t-1,l} = 1$). We carry forward the most recent observed embedding via a replication operation $\hat{\mathbf{e}}_t^{(l)} = \mathbf{e}_{t-1}^{(l)}$ (Duplicate Last Visit Embedding). This strategy is motivated by the clinical observation that biomarker and imaging values exhibit strong short-term autocorrelation, making the last observed value a more informative prior than the global mean.

No Past Visit Available ($t = 1$ or no prior observation exists for modality l). We apply a learned contextual imputation network that conditions on the available modalities at the current visit $\hat{\mathbf{e}}_t^{(l)} = \text{MLP}_{\text{imp}}\left(\bigoplus_{l':O_{t,l'}=1} \mathbf{e}_t^{(l')}\right)$, where $\bigoplus$ denotes mean-pooling over observed modality embeddings and MLP_{imp} is a shared contextual imputation network with learnable zero-masking, preventing contamination of observed modalities by missing signal. The fully imputed embedding matrix is defined as:

$$\mathbf{E}_t^{\text{imp}}[l,:] = \begin{cases} \mathbf{e}_t^{(l)} & \text{if } O_{t,l} = 1, \\ \hat{\mathbf{e}}_t^{(l)} & \text{if } O_{t,l} = 0. \end{cases} \tag{1}$$

Tokenization Module. Given $\mathbf{E}_t^{\text{imp}}$, we produce a unified visit-level token $\mathbf{z}_t \in \mathbb{R}^d$ via attentional fusion of two complementary representations: a *modality embedding* and a *viscode encoding*.

Modality Embedding via Multi-Head Attention. We apply multi-head self-attention over the L modality embeddings $\mathbf{A}_t = \text{MultiHead}\left(\mathbf{E}_t^{\text{imp}},\ \mathbf{E}_t^{\text{imp}},\ \mathbf{E}_t^{\text{imp}}\right) \in \mathbb{R}^{L\times d}$, where $\text{MultiHead}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{Concat}(\text{head}_1,\ldots,\text{head}_H)\mathbf{W}^O$ with each head defined as:

$$\text{head}_h = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{W}_h^Q\,(\mathbf{K}\mathbf{W}_h^K)^\top}{\sqrt{d/H}}\right)\mathbf{V}\mathbf{W}_h^V. \tag{2}$$

This operation learns inter-modality dependencies, enabling, e.g., MRI-derived morphological features to modulate the interpretation of clinical assessment scores.

Viscode Encoding via Multi-Frequency Sinusoidal Positional Encoding. To encode the temporal position of each visit in the patient's longitudinal sequence, we perform a continuous-time multi-frequency sinusoidal encoding that maps each visit's absolute timestamp $\tau_t \in \mathbb{R}_{>0}$ (e.g., months from enrolment) to a fixed-dimensional vector without assuming uniform visit spacing. For F frequency bands with base frequencies $\omega_k = 2^{k-1}$ for $k = 1,\ldots,F$, the viscode encoding of visit t is:

$$\mathbf{v}_t = \text{MLP}_{\text{viscode}}([\sin(\omega_1\,\tau_t),\ \cos(\omega_1\,\tau_t),\ \ldots,\ \sin(\omega_F\,\tau_t),\ \cos(\omega_F\,\tau_t)\,]) \in \mathbb{R}^d, \tag{3}$$

yielding a $2F$-dimensional raw encoding that is projected to $\mathbb{R}^d$ by the downstream MLP. To see why multiple frequencies are necessary, note that a single sine function is not injective: $\sin(\omega\tau) = 0$ at both $\tau = 0$ and $\tau = \pi/\omega$, so two visits at those months would receive identical encodings. By concatenating F sine-cosine pairs at geometrically spaced frequencies, two visits τ and τ' receive the same encoding only if $\sin(\omega_k\tau) = \sin(\omega_k\tau')$ and $\cos(\omega_k\tau) = \cos(\omega_k\tau')$ hold simultaneously for all $k = 1,\ldots,F$, which requires $\omega_k(\tau - \tau') \equiv 0 \pmod{2\pi}$ for all k. Since $\tau,\tau' \in \mathbb{Z}_{>0}$ (integer month indices), 2π is irrational, and the frequencies ω_k are chosen to be incommensurable, this collision cannot occur within any clinically plausible range of visit months.The downstream $\text{MLP}_{\text{viscode}}$ then

learns a task-specific projection of these multi-frequency components, enabling the model to weight slow trends (low ω_k) versus rapid rate changes (high ω_k) adaptively for the downstream classification task.

Attentional Fusion. The modality embedding $\bar{\mathbf{A}}_t \in \mathbb{R}^d$ and viscode encoding $\mathbf{v}_t$ are fused via elementwise addition $\mathbf{z}_t = \bar{\mathbf{A}}_t + \mathbf{v}_t \in \mathbb{R}^d$. This additive fusion preserves gradient flow from both branches and has been shown to be effective in multimodal fusion settings Kiela et al. (2019).

4.3 Trajectory-Aware Longitudinal Modelling

Given the sequence of visit tokens $\{\mathbf{z}_{t'}\}_{t'=1}^{t}$, we produce a contextualised representation $\mathbf{h}_t \in \mathbb{R}^d$ that integrates information across all past visits, capturing the patient's disease-progression trajectory. We apply a stack of $N = 2$ identical Transformer encoder blocks to model temporal dependencies across visits. Each block consists of a multi-head self-attention sublayer (with $H = 8$ heads) followed by a position-wise feed-forward network (FFN), with residual connections and layer normalisation (Vaswani et al., 2017). To prevent information leakage from future visits, we apply a causal attention mask: position t' may attend only to positions $t'' \leq t'$. This ensures that the prediction at visit t is conditioned solely on the history $\{v_1, \ldots, v_t\}$. The output of the final encoder layer at position t yields $\mathbf{h}_t = \mathbf{z}_t^{(N)} \in \mathbb{R}^d$. This embedding encodes not merely the current visit's observations but a contextualised summary of the patient's full longitudinal trajectory, providing a rich input to the downstream MoE routing module.

4.4 Context-Aware MoE Routing

Expert Architecture. The MoE layer comprises $K = K_i + K_s$ experts partitioned into two groups: i) *Individual Experts* $\{\mathcal{E}_k^{\text{ind}}\}_{k=1}^{K_i}$: Specialised sub-networks that capture distinct patient-subpopulation patterns. Each expert is an MLP $\mathcal{E}_k^{\text{ind}}(\mathbf{h}_t) = \text{MLP}_k(\mathbf{h}_t) \in \mathbb{R}^{|\mathcal{Y}|}$. ii) *Shared Experts* $\{\mathcal{E}_k^{\text{shr}}\}_{k=1}^{K_s}$: Experts accessible to *all* inputs that capture population-level common patterns, providing a regularisation effect and preventing over-specialisation. In total the MoE layer contains $K = 8$ experts, each producing a 3-class logit vector.

MoE Router. The router is a two-layer MLP mapping $\mathbf{h}_t$ to a scoring vector over all K experts $\mathbf{s}_t = \text{Router}(\mathbf{h}_t) = \text{MLP}_{\text{router}}(\mathbf{h}_t) \in \mathbb{R}^K$. Expert selection is performed via hard top-1 routing $k^* = \arg\max_{k \in \{1,\ldots,K\}} s_{t,k}$. The shared experts $\{\mathcal{E}_k^{\text{shr}}\}$ are always activated for all inputs, providing a stable common representation. Hard top-1 routing imposes sparsity in expert utilisation, reducing computational cost while encouraging expert specialisation, in contrast to soft routing, which can lead to diffuse utilisation and slower specialisation. The shared expert mechanism ensures that even under hard routing, all patients benefit from a common representational backbone, mitigating the cold-start problem for rare patient subgroups.

Aggregation over Visits. Let $\hat{\mathbf{y}}_t = \mathcal{E}_{k^*}(\mathbf{h}_t) \in \mathbb{R}^{|\mathcal{Y}|}$ denote the logit vector produced by the selected expert at visit t. The final prediction is obtained via attentional aggregation:

$$\hat{\mathbf{y}} = \sum_{t=1}^{T} \alpha_t \, \hat{\mathbf{y}}_t, \qquad \alpha_t = \frac{\exp\left(s_{t,k_t^*}\right)}{\sum_{t'=1}^{T} \exp\left(s_{t',k_{t'}^*}\right)}, \tag{4}$$

where α_t weights each visit's contribution by the router's confidence, assigning greater influence to visits with more informative or reliable observations. The aggregated logit vector $\hat{\mathbf{y}}$ is passed through a softmax $\hat{p}(y \mid \mathcal{V}) = \text{Softmax}(\hat{\mathbf{y}}) \in \Delta^{|\mathcal{Y}|-1}$, and the model is trained with the standard cross-entropy loss $\mathcal{L} = -\sum_{c \in \mathcal{Y}} \mathbb{1}[y = c] \, \log \hat{p}(c \mid \mathcal{V})$.

Load-Balancing Regularisation. To prevent router collapse, we augment the training objective with an auxiliary load-balancing loss (Fedus et al., 2022) $\mathcal{L}_{\text{bal}} = K \cdot \sum_{k=1}^{K} f_k \cdot P_k$, where $f_k = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \mathbb{1}[k_i^* = k]$ is the fraction of inputs in mini-batch $\mathcal{B}$ routed to expert k, and $P_k = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} s_{i,k}$ is the mean routing score for expert k. The total training objective is $\mathcal{L}_{\text{total}} = \mathcal{L} + \lambda \, \mathcal{L}_{\text{bal}}$, where $\lambda > 0$ is a regularisation coefficient controlling the trade-off between task performance and expert load balance.

5 EXPERIMENTS

ADNI. (Weiner et al., 2010) The Alzheimer's Disease Neuroimaging Initiative comprises of 2,693 patients across CN (1,049; 38.9%), MCI (1,194; 44.3%), and AD (450; 16.7%), with four modalities: MRI/PET (Doshi et al., 2016) registered to MNI space (Ou et al., 2011) (2,359; 87.6%); CSF biomarkers (Aβ40, Aβ42, total/phospho-Tau, ApoE genotyping), normalised to $[-1, 1]$ with categoricals one-hot encoded (1,611; 59.8%); longitudinal clinical assessments from five source files with label-leaking columns excluded (2,684; 99.7%); and SNP data from phases 1/GO/2/3 harmonised to GRCh38 via LiftOver and LD-pruned (window 50, step 5, $r^2 < 0.1$) (2,226; 82.7%).

OASIS-3. (LaMontagne et al., 2019) This open-access longitudinal dataset from Washington University comprises of 1,377 participants (aged 42-95) across CN (857; 62.2%), MCI (203; 14.7%), and AD (317; 23.0%), with T1-weighted MRI (1,098; 79.7%), CSF amyloid/tau assays and PET amyloid burden scores (Pittsburgh Compound-B and Florbetapir), CDR/MMSE assessments (1,098; 79.7%), and ApoE ε4 with select AD-associated SNPs (620; 45.0%). OASIS-3 presents complementary challenges through higher visit spacing irregularity, more severe modality missingness, and a markedly more skewed class distribution; prediction target: $y \in \{\mathrm{CN}, \mathrm{MCI}, \mathrm{AD}\}$.

MIMIC-IV. (Johnson et al., 2023) 98,323 adults with at least two hospital visits at Beth Israel Deaconess Medical Center, with four modalities: clinical notes (N) encoded via a pre-trained clinical language model (98,323; 100%); lab values and vitals aggregated over the last visit window (80,802; 82.2%); ICD-9 codes as multi-hot vectors (96,689; 98.3%); and structured demographics, i.e., age, sex, ethnicity, insurance type (98,323; 100%). Binary one-year mortality under severe class imbalance (Survived: 79,266, 80.6%; Died: 19,057, 19.4%) serves as a large-scale stress test for generalisation to a fundamentally different clinical task with distinct modality semantics.

Baselines. We compare LONGMOE against five baselines under identical splits and missing-modality protocols. Flex-MoE (Yun et al., 2024) handles arbitrary modality combinations via a learnable missing-modality bank and a two-stage SMoE (G-Router/S-Router), augmented with a mean-pooled preceding-visit summary for longitudinal fairness. FuseMoE (Han et al., 2024) is a concurrent MoE Transformer augmented with a single-layer GRU over per-visit embeddings. TF (Zadeh et al., 2017) computes outer-product tensor interactions, zero-pads absent modalities, and has no temporal capacity. MulT (Tsai et al., 2019) uses directional cross-modal attention but is designed for language-vision sequences and collapses under systematic missingness. Lee-MMGRU (Lee et al., 2019) is the only baseline with explicit longitudinal modelling via a GRU encoder, but requires all modalities at every visit and thus cannot exploit partially observed patients.

Experimental Settings. All models are trained on NVIDIA H100 GPUs (80 GB) with a patient-level stratified 70/15/15 train/validation/test split, keeping all visits of a patient in one partition. LONGMOE uses $d = 128$, a 2-layer, 8-head Transformer encoder, $K = 8$ MoE experts, AdamW (Loshchilov and Hutter, 2019), and $F = 8$ frequency bands (16-dimensional raw viscode encoding). To prevent temporal leakage, each test patient is evaluated on their final visit, with all prior visits as context for the encoder and imputation module, mirroring clinical deployment. AUC-ROC is the primary metric; accuracy and F1 are reported in Appendix A.8. Under current-visit masking, an excluded modality is withheld only at the evaluated visit while prior visits retain their observed modalities, so each combination describes the prediction visit rather than the full record. Genetics, a static per-patient attribute, is the only modality removed from the entire record when excluded. Visit-independent baselines lack history, so their single-modality performance is bounded by that modality's standalone signal. We quantify this intrinsic asymmetry in Appendix A.9.1 under full-trajectory masking.

6 RESULTS

Performance Analysis. Tables 1, 2, and 3 report AUC-ROC across all $2^L - 1$ modality combinations; † denotes statistical significance (paired t-test, Bonferroni-corrected, $p < 0.05$, five seeds). On ADNI, LONGMOE achieves the best AUC-ROC across all 15 combinations, with gains exceeding 0.45 in imaging- and genetics-only settings where single-visit baselines collapse (B: 0.9493 vs. FuseMoE 0.4915; G: 0.9517 vs. 0.6316), and maintains a near-flat AUC-ROC range of $[0.9493, 0.9755]$

Table 1: Performance comparison on ADNI (AUC-ROC) for all modality combinations. † denotes statistical significance over the runner-up baseline (paired t-test, Bonferroni-corrected, $p < 0.05$).

Modalities	MulT	Flex-MoE	FuseMoE	Lee-MMGRU	TF	LongMoE
B	$0.5586_{\pm.0170}$	$0.5220_{\pm.1096}$	$0.4915_{\pm.0230}$	$0.5000_{\pm.0000}$	$0.6419_{\pm.0127}$	$\mathbf{0.9493_{\pm.0040}}^{\dagger}$
C	$0.9619_{\pm.0060}$	$0.9650_{\pm.0016}$	$0.9647_{\pm.0012}$	$0.9704_{\pm.0034}$	$0.9250_{\pm.0192}$	$\mathbf{0.9750_{\pm.0027}}^{\dagger}$
G	$0.6067_{\pm.0193}$	$0.6011_{\pm.0305}$	$0.6316_{\pm.0059}$	$0.5000_{\pm.0000}$	$0.6393_{\pm.0126}$	$\mathbf{0.9517_{\pm.0036}}^{\dagger}$
I	$0.5884_{\pm.0166}$	$0.6632_{\pm.0199}$	$0.6719_{\pm.0134}$	$0.5771_{\pm.0696}$	$0.7137_{\pm.0109}$	$\mathbf{0.9534_{\pm.0032}}^{\dagger}$
B,C	$0.9638_{\pm.0047}$	$0.9573_{\pm.0048}$	$0.9632_{\pm.0010}$	$0.9704_{\pm.0034}$	$0.9262_{\pm.0121}$	$\mathbf{0.9749_{\pm.0026}}^{\dagger}$
B,G	$0.6286_{\pm.0088}$	$0.6293_{\pm.0426}$	$0.6244_{\pm.0056}$	$0.5000_{\pm.0000}$	$0.6911_{\pm.0155}$	$\mathbf{0.9515_{\pm.0029}}^{\dagger}$
C,G	$0.9712_{\pm.0023}$	$0.9490_{\pm.0043}$	$0.9679_{\pm.0011}$	$0.9704_{\pm.0034}$	$0.9397_{\pm.0097}$	$\mathbf{0.9755_{\pm.0030}}^{\dagger}$
I,B	$0.5985_{\pm.0154}$	$0.6645_{\pm.0233}$	$0.6707_{\pm.0108}$	$0.5771_{\pm.0696}$	$0.7502_{\pm.0157}$	$\mathbf{0.9531_{\pm.0029}}^{\dagger}$
I,C	$0.9584_{\pm.0069}$	$0.9526_{\pm.0045}$	$0.9571_{\pm.0040}$	$0.9685_{\pm.0037}$	$0.9194_{\pm.0134}$	$\mathbf{0.9740_{\pm.0034}}^{\dagger}$
I,G	$0.6254_{\pm.0215}$	$0.7043_{\pm.0271}$	$0.7151_{\pm.0161}$	$0.5771_{\pm.0696}$	$0.7473_{\pm.0124}$	$\mathbf{0.9544_{\pm.0033}}^{\dagger}$
B,C,G	$0.9712_{\pm.0018}$	$0.9513_{\pm.0053}$	$0.9670_{\pm.0012}$	$0.9704_{\pm.0034}$	$0.9398_{\pm.0111}$	$\mathbf{0.9754_{\pm.0029}}^{\dagger}$
I,B,C	$0.9608_{\pm.0058}$	$0.9517_{\pm.0075}$	$0.9549_{\pm.0036}$	$0.9685_{\pm.0037}$	$0.9248_{\pm.0133}$	$\mathbf{0.9739_{\pm.0033}}^{\dagger}$
I,B,G	$0.6297_{\pm.0200}$	$0.7060_{\pm.0245}$	$0.7123_{\pm.0144}$	$0.5771_{\pm.0696}$	$0.7629_{\pm.0121}$	$\mathbf{0.9543_{\pm.0029}}^{\dagger}$
I,C,G	$0.9665_{\pm.0027}$	$0.9552_{\pm.0039}$	$0.9617_{\pm.0038}$	$0.9685_{\pm.0037}$	$0.9356_{\pm.0080}$	$\mathbf{0.9744_{\pm.0032}}^{\dagger}$
I,B,C,G	$0.9666_{\pm.0023}$	$0.9535_{\pm.0065}$	$0.9604_{\pm.0037}$	$0.9685_{\pm.0037}$	$0.9320_{\pm.0114}$	$\mathbf{0.9742_{\pm.0032}}^{\dagger}$

Table 2: Performance comparison on OASIS-3 (AUC-ROC) for all modality combinations. † denotes statistical significance over the runner-up baseline (paired t-test, Bonferroni-corrected, $p < 0.05$).

Modalities	MulT	Flex-MoE	FuseMoE	Lee-MMGRU	TF	LongMoE
C	$0.4913_{\pm.0010}$	$0.5007_{\pm.0031}$	$0.4991_{\pm.0027}$	$\mathbf{0.8776_{\pm.0031}}$	$0.7364_{\pm.0243}$	$0.8630_{\pm.0191}$
G	$0.5040_{\pm.0114}$	$0.4993_{\pm.0030}$	$0.5016_{\pm.0022}$	$0.5264_{\pm.0470}$	$0.6533_{\pm.0255}$	$\mathbf{0.7624_{\pm.0078}}^{\dagger}$
I	$0.5352_{\pm.0035}$	$0.5313_{\pm.0017}$	$0.5309_{\pm.0012}$	$0.6760_{\pm.0199}$	$0.6846_{\pm.0164}$	$\mathbf{0.7818_{\pm.0202}}^{\dagger}$
C,G	$0.6591_{\pm.0179}$	$0.6722_{\pm.0062}$	$0.6292_{\pm.0114}$	$\mathbf{0.8674_{\pm.0036}}$	$0.7498_{\pm.0253}$	$0.8643_{\pm.0178}$
I,C	$0.7589_{\pm.0083}$	$0.7651_{\pm.0037}$	$0.7663_{\pm.0035}$	$\mathbf{0.8568_{\pm.0111}}$	$0.7408_{\pm.0261}$	$0.8521_{\pm.0208}$
I,G	$0.6320_{\pm.0109}$	$0.6632_{\pm.0107}$	$0.6388_{\pm.0156}$	$0.6672_{\pm.0359}$	$0.7153_{\pm.0143}$	$\mathbf{0.7767_{\pm.0130}}^{\dagger}$
I,C,G	$0.7713_{\pm.0100}$	$0.8030_{\pm.0062}$	$0.7811_{\pm.0200}$	$0.8330_{\pm.0066}$	$0.7562_{\pm.0155}$	$\mathbf{0.8552_{\pm.0192}}^{\dagger}$

including the full-modality setting (0.9742), reflecting the stability of longitudinal context across the full modality-combination space. On OASIS-3, LONGMOE leads on four of seven combinations including full-modality (I,C,G: 0.8552); Lee-MMGRU leads on clinical-heavy combinations (C: 0.8776; C,G: 0.8674; I,C: 0.8568) where its recurrent encoder is competitive, but LONGMOE recovers its advantage on imaging- and genetics-dominant combinations (G: 0.7624; I: 0.7818) where longitudinal trajectory context is the decisive signal. On MIMIC-IV, LONGMOE achieves the best AUC-ROC across all 15 combinations (N,L,C,D: 0.9150; D-only: 0.8720 vs. MulT's 0.6892 and TF's 0.6692) with standard deviations consistently below 0.007, confirming that the robustness mechanism transfers to a large-scale binary mortality task with fundamentally different modality semantics.

Complexity Analysis. Table 4 reports parameters, forward-pass time, GPU memory, training time, and AUC-ROC for all models on a single NVIDIA H100 GPU. LONGMOE has 4.40M parameters versus 183K for Flex-MoE (24×) and a forward-pass latency of 6.04 ms versus 0.56 ms for TF (11×), a direct consequence of the trajectory-aware Transformer encoder processing the full visit sequence. GPU memory remains modest (149.8 MB on ADNI, 112.9 MB on MIMIC-IV, comparable to TF at 145.6 MB and 103.9 MB). The dominant cost is training rather than inference: 20.34 s/epoch on ADNI versus 2.47 s for TF. We report wall-clock time as the primary latency measure since GFLOPs underestimate actual latency due to memory access patterns in sequential Transformers; at 6 ms per patient the model sustains ˜165 patients per second on a single GPU, so full-cohort MIMIC-IV inference (˜98K patients) takes ˜10 minutes and per-patient inference is well within interactive latency. The encoder scales as $\mathcal{O}(T^2 d)$ in visit sequence length; with $\bar{T} = 4.2$ (ADNI) and $\bar{T} = 3.1$ (MIMIC-IV) this cost is negligible, and linear-complexity attention variants (Kitaev et al., 2020) provide a direct scaling path for cohorts with longer sequences ($T > 50$).

Table 3: Performance comparison on MIMIC-IV (AUC-ROC) for all modality combinations. † denotes statistical significance over the runner-up baseline (paired t-test, Bonferroni-corrected, $p < 0.05$).

Modalities	MulT	Flex-MoE	FuseMoE	Lee-MMGRU	TF	LongMoE
C	$0.8455_{\pm.0021}$	$0.8206_{\pm.0231}$	$0.8166_{\pm.0423}$	$0.8893_{\pm.0026}$	$0.8636_{\pm.0006}$	$\mathbf{0.8987}_{\pm.0008}$ †
D	$0.6892_{\pm.0036}$	$0.6541_{\pm.0278}$	$0.6866_{\pm.0017}$	$0.7423_{\pm.0005}$	$0.6692_{\pm.0016}$	$\mathbf{0.8720}_{\pm.0042}$ †
L	$0.8150_{\pm.0027}$	$0.8164_{\pm.0037}$	$0.8245_{\pm.0024}$	$0.8243_{\pm.0124}$	$0.8380_{\pm.0008}$	$\mathbf{0.8788}_{\pm.0064}$ †
N	$0.7780_{\pm.0040}$	$0.7731_{\pm.0101}$	$0.7732_{\pm.0035}$	$0.7967_{\pm.0024}$	$0.7980_{\pm.0011}$	$\mathbf{0.8807}_{\pm.0026}$ †
C,D	$0.8556_{\pm.0016}$	$0.8078_{\pm.0344}$	$0.8267_{\pm.0446}$	$0.8965_{\pm.0015}$	$0.8710_{\pm.0010}$	$\mathbf{0.9041}_{\pm.0008}$ †
L,C	$0.8813_{\pm.0013}$	$0.8748_{\pm.0136}$	$0.8760_{\pm.0116}$	$0.9044_{\pm.0024}$	$0.8846_{\pm.0007}$	$\mathbf{0.9077}_{\pm.0003}$ †
L,D	$0.8335_{\pm.0018}$	$0.8250_{\pm.0115}$	$0.8273_{\pm.0178}$	$0.8564_{\pm.0053}$	$0.8511_{\pm.0009}$	$\mathbf{0.8900}_{\pm.0056}$ †
N,C	$0.8789_{\pm.0008}$	$0.8611_{\pm.0146}$	$0.8644_{\pm.0162}$	$0.9020_{\pm.0018}$	$0.8887_{\pm.0003}$	$\mathbf{0.9062}_{\pm.0016}$ †
N,D	$0.7957_{\pm.0027}$	$0.7332_{\pm.0221}$	$0.7935_{\pm.0025}$	$0.8407_{\pm.0016}$	$0.8044_{\pm.0009}$	$\mathbf{0.8908}_{\pm.0042}$ †
N,L	$0.8690_{\pm.0019}$	$0.8651_{\pm.0113}$	$0.8680_{\pm.0020}$	$0.8656_{\pm.0072}$	$0.8793_{\pm.0008}$	$\mathbf{0.8940}_{\pm.0039}$ †
L,C,D	$0.8874_{\pm.0005}$	$0.8753_{\pm.0181}$	$0.8782_{\pm.0162}$	$0.9092_{\pm.0016}$	$0.8901_{\pm.0009}$	$\mathbf{0.9115}_{\pm.0003}$ †
N,C,D	$0.8850_{\pm.0006}$	$0.8618_{\pm.0076}$	$0.8729_{\pm.0165}$	$0.9068_{\pm.0017}$	$0.8913_{\pm.0006}$	$\mathbf{0.9102}_{\pm.0020}$ †
N,L,C	$0.8984_{\pm.0007}$	$0.8929_{\pm.0092}$	$0.8960_{\pm.0049}$	$0.9107_{\pm.0015}$	$0.8998_{\pm.0002}$	$\mathbf{0.9119}_{\pm.0010}$ †
N,L,D	$0.8778_{\pm.0013}$	$0.8734_{\pm.0042}$	$0.8711_{\pm.0083}$	$0.8819_{\pm.0037}$	$0.8834_{\pm.0006}$	$\mathbf{0.9014}_{\pm.0039}$ †
N,L,C,D	$0.9023_{\pm.0003}$	$0.8993_{\pm.0082}$	$0.8998_{\pm.0059}$	$0.9043_{\pm.0014}$	$0.9018_{\pm.0004}$	$\mathbf{0.9150}_{\pm.0012}$ †

Table 4: Complexity comparison across models on all three datasets (full modality setting).

Dataset	Metric	MulT	Flex-MoE	FuseMoE	Lee-MMGRU	TF	LongMoE
ADNI	# Params ↓	640,771	183,683	580,248	116,175	109,699	4,396,305
	Fwd Time (ms) ↓	1.74	2.26	1.46	2.77	**0.56**	6.04
	GPU Mem (MB) ↓	114.6	109.0	**103.9**	123.7	145.6	149.8
	Train Time (s/ep) ↓	7.91	5.98	5.52	6.70	**2.47**	20.34
	AUC-ROC ↑	0.9666	0.9535	0.9604	0.9685	0.9320	**0.9742**
OASIS-3	# Params ↓	623,363	145,283	427,032	78,540	92,675	4,261,262
	Fwd Time (ms) ↓	1.61	1.57	1.38	1.98	**0.44**	5.66
	GPU Mem (MB) ↓	103.0	105.2	101.0	116.2	140.2	**61.0**
	Train Time (s/ep) ↓	11.23	40.81	8.57	25.23	**7.70**	77.37
	AUC-ROC ↑	0.7713	0.8030	0.7811	0.8330	0.7562	**0.8552**
MIMIC-IV	# Params ↓	662,530	205,442	601,104	149,258	131,458	1,321,614
	Fwd Time (ms) ↓	1.72	2.08	1.43	3.06	**0.55**	6.09
	GPU Mem (MB) ↓	93.7	87.9	**82.7**	103.3	103.9	112.9
	Train Time (s/ep) ↓	7.68	5.46	5.60	7.16	**2.46**	20.23
	AUC-ROC ↑	0.9023	0.8993	0.8998	0.9043	0.9018	**0.9150**

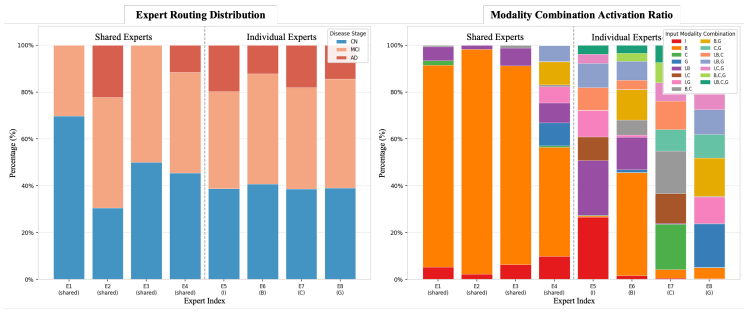


Figure 2: *Left:* Expert routing distribution for shared and individual experts. *Right:* Modality combination activation ratio per expert.

Expert Specialisation Analysis. Figure 2 analyses the routing behaviour of LongMoE's $K = 8$ experts on the ADNI testing set. The left plot shows a monotone disease-stage gradient across shared experts: E1 is predominantly CN, E2 is balanced CN/MCI, and E3/E4 increasingly concentrate MCI and AD patients, while individual experts collectively mirror the ADNI class prior (40% CN, 40% MCI, 20% AD) with distinct stage profiles, confirming that shared and individual experts capture complementary population-level and subgroup-specific patterns respectively. The right plot shows that shared experts are overwhelmingly activated by Biospecimen-containing combinations, whereas individual experts exhibit modality-aligned

specialisation, confirming that context-conditioned routing learns meaningful specialisation along both disease-stage and modality-availability axes.

Robustness under Missing Modalities. Figure 3 plots AUC-ROC as a function of missing modalities across all three benchmarks. On ADNI, LONGMOE degrades minimally from 0.974 (all present) to 0.957 (one modality left; $\Delta\text{AUC} - \text{ROC} = 0.017$), whereas every baseline loses between 0.20 and 0.33: the strongest baseline at maximum missingness, TF, falls from 0.932 to 0.730, and Lee-MMGRU from 0.968 to 0.637. On OASIS 3, LONGMOE declines from 0.855 to 0.799, while MulT, Flex-MoE and FuseMoE fall to chance (0.51) once only one modality remains and Lee-MMGRU and TF reach 0.693 and 0.697. On MIMIC-IV, LONGMOE retains 0.883 with a single modality (from 0.915), against 0.813 for Lee-MMGRU and 0.775-0.782 for the remaining baselines. Under our protocol a modality is withheld at the current visit, so LONGMOE can still draw on earlier observations of it. Across all three benchmarks, the near-flat degradation curve confirms that the trajectory-aware encoder substitutes longitudinal context for missing modality information.

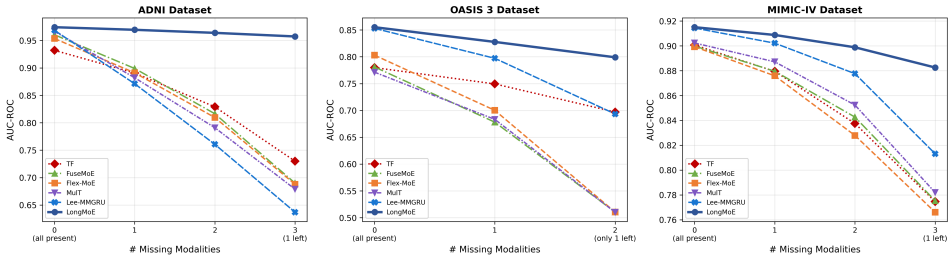


Figure 3: Missing-modality robustness curves (AUC-ROC) for all the datasets.

Ablation Study. Table 5 reports a component-wise ablation on ADNI (full modality), confirming every component contributes positively. Replacing the MoE with a single expert produces drop ($\Delta\text{AUC}=-0.0241$), confirming expert specialisation provides benefit. Removing the Longitudinal Encoder (LE; $\Delta\text{Acc}=-0.0329$, $\Delta\text{AUC}=-0.0335$) produces drop, which reduces LONGMOE to a per-visit classifier, and Viscode Encoding (VE; $\Delta\text{Acc}=-0.0225$, $\Delta\text{AUC}=-0.0178$), validating the sinusoidal temporal encoding as an independent longitudinal signal beyond self-attention. Load Balancing ($\mathcal{L}_{\text{bal}}$) produces drop ($\Delta\text{AUC}=-0.0195$), both imputation ablations degrade comparably with Zero Padding slightly worse than Global Mean (greater distributional shift). The detailed ablation studies are provided in Appendix A.10.

Table 5: Component-wise ablation of LONGMOE on ADNI dataset (full modality).

Model Variants	Metrics			Change $\Delta\downarrow$		
	Acc	AUC-ROC	F1	ΔAcc	ΔAUC	ΔF1
w/o LE	0.8452	0.9407	0.7250	−0.0329	−0.0335	−0.0814
w/o VE	0.8556	0.9564	0.7816	−0.0225	−0.0178	−0.0248
w/o Imp-ZP	0.8650	0.9607	0.7850	−0.0131	−0.0135	−0.0214
w/o Imp-GM	0.8612	0.9631	0.7910	−0.0169	−0.0111	−0.0154
w/o MoE	0.8405	0.9501	0.7824	−0.0377	−0.0241	−0.0240
w/o Shared Expert	0.8631	0.9606	0.7895	−0.0150	−0.0136	−0.0169
w/o $\mathcal{L}_{\text{bal}}$	0.8650	0.9547	0.7904	−0.0131	−0.0195	−0.0160
LongMoE	**0.8781**	**0.9742**	**0.8064**	—	—	—

7 CONCLUSION

We presented LONGMOE, a unified framework that jointly addresses modality missingness and longitudinal temporal dynamics in multimodal sequential learning. By composing a context-aware imputation module, attentional tokenization module, and a trajectory-aware Transformer encoder with a context-conditioned Sparse MoE routing layer, LONGMOE enables patient-specific conditional computation over the full visit history under arbitrary modality availability. Experiments on different datasets show that LONGMOE is strongest under missing or weak contemporaneous modalities, but clinical-score-heavy settings remain competitive for recurrent baselines.

AI USE STATEMENT

Generative AI tools were used solely for language editing of the manuscript, including improving grammar and readability of author-written text. They were not used to generate research ideas, experimental results, or the statements and proofs of the theoretical results, all of which were produced by the authors. All the work was reviewed by the authors, who take full responsibility for the final content of this work, including all text, claims, and artifacts.

ETHICS STATEMENT

This work uses only existing, de-identified datasets (ADNI, OASIS-3, and MIMIC-IV) accessed under their respective data use agreements, with MIMIC-IV obtained through PhysioNet credentialed access after completion of the required human subjects research training. No new data were collected, no human participants were contacted, and no attempt was made to re-identify individuals. We acknowledge that these cohorts are not fully representative of the general population: ADNI and OASIS-3 participants are predominantly highly educated and of European ancestry, and MIMIC-IV reflects a single academic medical centre in the United States. Model performance may therefore not transfer equitably across demographic groups or care settings. LONGMOE is a research prototype intended to support rather than replace clinical judgement, and it should not be used for diagnosis or prognosis without prospective validation, subgroup fairness evaluation, and appropriate regulatory review. The authors declare no competing interests.

REPRODUCIBILITY STATEMENT

We have made several efforts to ensure reproducibility. The full model architecture is specified in Section 4, and the experimental protocol, including patient-level stratified 70/15/15 splits, final-visit evaluation, and the current-visit masking scheme, is described in Section 5. Dataset statistics and preprocessing steps are provided in Section 5 and Appendix A.7, and complete hyperparameter configurations for LONGMOE and all baselines are listed in Appendix A.12 (Tables 18 to 23). All results are reported as mean and standard deviation over five random seeds with Bonferroni-corrected significance tests. Statements and proofs of all theoretical results appear in Appendices A.1 to A.5. Anonymised source code, including preprocessing scripts, configuration files, and random seeds, is provided as supplementary material and will be released publicly upon acceptance. All datasets are available to qualified researchers from their original providers.

REFERENCES

T. Baltrušaitis, C. Ahuja, and L.-P. Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2018.

L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

Z. Chen, Y. Shen, M. Ding, Z. Chen, H. Zhao, E. G. Learned-Miller, and C. Gan. Mod-squad: Designing mixtures of experts as modular multi-task learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11828–11837, 2023.

K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, 2014.

E. Choi, M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz, and W. Stewart. RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29, pages 3504–3512, 2016.

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 2nd edition, 2006.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 4171–4186, 2019.

J. Doshi, G. Erus, Y. Ou, S. M. Resnick, R. C. Gur, R. E. Gur, T. D. Satterthwaite, S. Furth, C. Davatzikos, and A. D. N. Initiative. MUSE: Multi-atlas region segmentation utilizing ensembles of registration algorithms and parameters, and locally optimal atlas selection. *NeuroImage*, 127: 186–195, 2016.

W. Fedus, B. Zoph, and N. Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

X. Han, H. Nguyen, C. Harris, N. Ho, and S. Saria. FuseMoE: Mixture-of-experts transformers for fleximodal fusion. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

H. Hazimeh, Z. Zhao, A. Chowdhery, M. Sathiamoorthy, Y. Chen, R. Mazumder, L. Hong, and E. H. Chi. DSelect-k: Differentiable selection in the mixture of experts with applications to multi-task learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 29335–29347, 2021.

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

M. Horn, M. Moor, C. Bock, B. Rieck, and K. Borgwardt. Set functions for time series. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 119, pages 4353–4363, 2020.

K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.

C. R. Jack and D. M. Holtzman. Biomarker modeling of Alzheimer's disease. *Neuron*, 80(6): 1347–1358, 2013.

R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991.

A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. de las Casas, E. B. Hanna, F. Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

A. Johnson, L. Bulgarelli, T. Pollard, S. Horng, L. A. Celi, and R. Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1–9, 2023.

M. I. Jordan and R. A. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6(2):181–214, 1994.

D. Kiela, C. Bhooshan, H. Firooz, E. Perez, and D. Testuggine. Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950*, 2019.

N. Kitaev, L. Kaiser, and A. Levskaya. Reformer: The efficient transformer. In *International Conference on Learning Representations (ICLR)*, 2020.

P. J. LaMontagne, T. L. Benzinger, J. C. Morris, S. Keefe, R. Hornbeck, C. Yao, A. Hintaus, and D. Marcus. OASIS-3: Longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and Alzheimer's disease. *medRxiv*, 2019.

G. Lee, K. Nho, B. Kang, K.-A. Sohn, and D. Kim. Predicting Alzheimer's disease progression using multi-modal deep learning approach. *Scientific Reports*, 9(1):1952, 2019.

P. P. Liang, Y. Lyu, X. Fan, Z. Wu, Y. Cheng, J. Wu, L. Chen, P. Wu, M. A. Lee, Y. Zhu, R. Salakhutdinov, and L. Morency. MultiBench: Multiscale benchmarks for multimodal representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 5813–5826, 2021.

R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons, 3rd edition, 2019.

I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

J. Ma, Z. Zhao, X. Yi, J. Chen, L. Hong, and E. H. Chi. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1930–1939, 2018.

F. Márquez and M. A. Yassa. Neuroimaging biomarkers for Alzheimer's disease. *Molecular Neurodegeneration*, 14(1):21, 2019.

B. Mustafa, C. Riquelme, J. Puigcerver, R. Jenatton, and N. Houlsby. Multimodal contrastive learning with LiMoE: The language-image mixture of experts. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

M. Odusami, R. Maskeliunas, R. Damasevicius, and S. Misra. Machine learning with multi-modal neuroimaging data to classify stages of Alzheimer's disease: A systematic review and meta-analysis. *Cognitive Neurodynamics*, pages 1–20, 2023.

Y. Ou, A. Sotiras, N. Paragios, and C. Davatzikos. DRAMMS: Deformable registration via attribute matching and mutual-saliency weighting. *Medical Image Analysis*, 15(4):622–639, 2011.

A. Papassotiropoulos, M. Fountoulakis, T. Dunckley, D. A. Stephan, and E. M. Reiman. Genetics, transcriptomics, and proteomics of Alzheimer's disease. *Journal of Clinical Psychiatry*, 67(4):652, 2006.

J. Peng, K. Zhou, R. Zhou, T. Hartvigsen, Y. Zhang, Z. Wang, and T. Chen. Sparse MoE as a new treatment: Addressing forgetting, fitting, and learning issues in multi-modal multi-task learning. In *International Conference on Learning Representations (ICLR)*, 2024.

W. Rahman, M. K. Hasan, S. Lee, A. B. Zadeh, C. Mao, L.-P. Morency, and E. Hoque. Integrating multimodal information in large pretrained transformers. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2359–2369, 2020.

C. Riquelme, J. Puigcerver, B. Mustafa, M. Neumann, R. Jenatton, A. S. Pinto, D. Keysers, and N. Houlsby. Scaling vision with sparse mixture of experts. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 8583–8595, 2021.

N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations (ICLR)*, 2017.

N. Shukla and B. Marlin. Multi-time attention networks for irregularly sampled time series. In *International Conference on Learning Representations (ICLR)*, 2021.

Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 6558–6569, 2019.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pages 5998–6008, 2017.

J. Venugopalan, L. Tong, H. R. Hassanzadeh, and M. D. Wang. Multimodal deep learning models for early detection of Alzheimer's disease stage. *Scientific Reports*, 11(1):3254, 2021.

H. Wang, Y. Chen, C. Ma, J. Avery, L. Hull, and G. Carneiro. Multi-modal learning with missing modality via shared-specific feature modelling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

M. W. Weiner, P. S. Aisen, C. R. Jack, W. J. Jagust, J. Q. Trojanowski, L. Shaw, A. J. Saykin, J. C. Morris, N. Cairns, L. A. Beckett, et al. The Alzheimer's disease neuroimaging initiative: Progress report and future plans. *Alzheimer's & Dementia*, 6(3):202–211, 2010.

S. Yun, I. Choi, J. Peng, Y. Wu, J. Bao, Q. Zhang, J. Xin, Q. Long, and T. Chen. Flex-MoE: Modeling arbitrary modality combination via the flexible mixture-of-experts. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1103–1114, 2017.

Y. Zhang, N. He, J. Yang, Y. Li, D. Wei, Y. Huang, Y. Zhang, Z. He, and Y. Zheng. mmFormer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation. *arXiv preprint arXiv:2206.02425*, 2022.

Y. Zhou, T. Lei, H. Liu, N. Du, Y. Huang, V. Zhao, A. Dai, Z. Chen, Q. Le, and J. Laudon. Mixture-of-experts with expert choice routing. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

A APPENDIX

We provide theoretical grounding for the four core architectural components of LONGMOE: the multi-frequency sinusoidal viscode encoding, the context-aware imputation module, the trajectory-aware Transformer encoder, and the sparse MoE routing layer.

A.1 MULTI-FREQUENCY SINUSOIDAL VISCODE ENCODING

Proposition A.1 (Injectivity of Multi-Frequency Sinusoidal Encoding)**.** *Let $\phi : \mathbb{Z}_{>0} \to \mathbb{R}^{2F}$ be the multi-frequency sinusoidal encoding defined by:*

$$\phi(\tau) = \big[\sin(\omega_1\tau),\, \cos(\omega_1\tau),\, \ldots,\, \sin(\omega_F\tau),\, \cos(\omega_F\tau)\big], \tag{5}$$

where $\omega_k = 2^{k-1}$ for $k = 1, \ldots, F$. Then ϕ is injective on all of $\mathbb{Z}_{>0}$: for any two distinct integer timestamps $\tau, \tau' \in \mathbb{Z}_{>0}$ with $\tau \neq \tau'$, we have $\phi(\tau) \neq \phi(\tau')$. In particular, ϕ is injective on the clinically relevant range $\tau \in \{1, 2, \ldots, 240\}$ (months 1-240, covering a 20-year longitudinal follow-up window encompassing all ADNI, OASIS-3, and MIMIC-IV visit ranges).

Proof. Suppose $\phi(\tau) = \phi(\tau')$ for some $\tau, \tau' \in \mathbb{Z}_{>0}$. This requires $\sin(\omega_k\tau) = \sin(\omega_k\tau')$ and $\cos(\omega_k\tau) = \cos(\omega_k\tau')$ simultaneously for all $k = 1, \ldots, F$, which holds if and only if:

$$\omega_k(\tau - \tau') \equiv 0 \pmod{2\pi} \quad \forall\, k = 1, \ldots, F. \tag{6}$$

Condition equation 6 is equivalent to requiring $\tau - \tau' = 2\pi m_k/\omega_k = 2\pi m_k/2^{k-1}$ for some integer m_k for each $k = 1, \ldots, F$.

We now show this forces $\tau = \tau'$. Since $\tau - \tau' \in \mathbb{Z}$ by assumption, we need $2\pi m_k/2^{k-1} \in \mathbb{Z}$ for each k. This requires $2\pi \in \mathbb{Q}$, since $m_k/2^{k-1}$ is rational for all integer m_k and all k. However, 2π is transcendental (in particular, irrational), so $2\pi m_k/2^{k-1} \notin \mathbb{Z}$ for any $m_k \neq 0$. Therefore the only solution to equation 6 is $m_k = 0$ for all k, which gives $\tau - \tau' = 0$, i.e., $\tau = \tau'$. Hence $\phi(\tau) = \phi(\tau')$ implies $\tau = \tau'$, and ϕ is injective on all of $\mathbb{Z}_{>0}$.

The clinical range $\{1, \ldots, 240\}$ is a subset of $\mathbb{Z}_{>0}$, so injectivity holds over this range as an immediate corollary. □

Remark A.2. The key step in the proof is the irrationality of 2π, which prevents any non-zero integer difference $\tau - \tau'$ from satisfying the collision condition equation 6. This argument holds for *any* choice of rational or integer base frequencies $\omega_k > 0$, and is not restricted to the geometric schedule $\omega_k = 2^{k-1}$. The geometric schedule is chosen to provide a wide dynamic range of temporal scales (from monthly to decadal patterns) rather than for the injectivity guarantee, which holds for any such frequencies. Concretely, with $F = 8$ frequency bands and timestamps in $\{1, \ldots, 240\}$, every pair of visits is guaranteed to receive a distinct 16-dimensional encoding, providing a unique temporal identity for all longitudinal sequences within the clinically plausible follow-up window of 20 years.

Proposition A.3 (Expressivity of Multi-Frequency Encoding)**.** *Let $f : \mathbb{Z}_{>0} \to \mathbb{R}$ be any function defined on a finite set of integer timestamps $\mathcal{T} = \{\tau_1, \ldots, \tau_T\}$. Then there exists an MLP $g : \mathbb{R}^{2F} \to \mathbb{R}$ with $F \geq T$ such that $g(\phi(\tau_t)) = f(\tau_t)$ for all $t = 1, \ldots, T$.*

Proof. By Proposition A.1, the matrix $\Phi \in \mathbb{R}^{T \times 2F}$ with rows $\phi(\tau_t)$ has distinct rows. For $F \geq T$, the $2F$-dimensional encoding space can be mapped to an arbitrary target vector $[f(\tau_1), \ldots, f(\tau_T)]^\top$ by solving the linear system $\Phi\mathbf{w} = \mathbf{f}$, which has a solution when Φ has full row rank. Since ϕ is injective and the frequencies ω_k are distinct, Φ constitutes a generalised Vandermonde-type matrix with full row rank for $T \leq 2F$, guaranteeing the existence of such a mapping. A single-layer linear MLP with weight vector $\mathbf{w}$ achieves this mapping exactly; a multi-layer MLP achieves it with strictly weaker conditions on F. □

A.2 CONTEXT-AWARE IMPUTATION

Proposition A.4 (Optimality of Carry-Forward Imputation under Short-Term Autocorrelation)**.** *Let $\{e_t^{(l)}\}_{t \geq 1}$ be a stationary stochastic process with autocorrelation function $\rho(h) = \mathrm{Corr}(e_t^{(l)}, e_{t+h}^{(l)})$, and let $\hat{e}_t^{(l)}$ denote the minimum mean-squared error (MMSE) linear estimator of a missing embedding*

at time t given the last observed embedding at time $t-1$. If $|\rho(1)| \geq |\rho(h)|$ for all $h > 1$ (i.e., the process exhibits monotonically decaying autocorrelation), then the carry-forward estimator $\hat{e}_t^{(l)} = e_{t-1}^{(l)}$ minimises the expected imputation error:

$$\mathbb{E}\left[\left\|e_t^{(l)} - \hat{e}_t^{(l)}\right\|^2\right] \leq \mathbb{E}\left[\left\|e_t^{(l)} - \hat{e}_t^{(l,\mathrm{alt})}\right\|^2\right] \tag{7}$$

for any linear estimator $\hat{e}_t^{(l,\mathrm{alt})}$ that depends on a single past observation at time $t-k$ with $k > 1$.

Proof. For a zero-mean stationary process, the MMSE linear estimator of $e_t^{(l)}$ given $e_{t-k}^{(l)}$ is $\hat{e}_t^{(l)} = \rho(k) \cdot e_{t-k}^{(l)}$. The corresponding MSE is:

$$\mathrm{MSE}(k) = \mathbb{E}\left[\|e_t^{(l)}\|^2\right]\left(1 - \rho(k)^2\right). \tag{8}$$

Since $|\rho(1)| \geq |\rho(k)|$ for all $k > 1$ by assumption, $1-\rho(1)^2 \leq 1-\rho(k)^2$, hence $\mathrm{MSE}(1) \leq \mathrm{MSE}(k)$. The carry-forward estimator corresponds to $k = 1$, establishing its MMSE optimality among single-lag linear estimators. □

Remark A.5. Biomarker embedding processes in longitudinal AD datasets exhibit strong short-term autocorrelation, e.g., hippocampal volume changes slowly relative to visit frequency, making the monotone decay assumption empirically reasonable. The proposition formalises the clinical intuition that the most recent observation is the most informative single prior for imputation.

A.3 Trajectory-Aware Longitudinal Encoder

Theorem A.6 (Trajectory-Aware Fusion Dominates Per-Visit Fusion under Markovian Disease Dynamics)**.** *Let the patient's disease state at visit t be represented by a latent variable $s_t \in \mathcal{S}$, and let the observed multimodal embedding $\mathbf{z}_t \in \mathbb{R}^d$ be a noisy function of s_t: $\mathbf{z}_t = f(s_t) + \epsilon_t$, where ϵ_t is i.i.d. mean-zero noise with covariance $\sigma^2\mathbf{I}$. Assume disease progression follows a first-order Markov chain: $p(s_t \mid s_1, \ldots, s_{t-1}) = p(s_t \mid s_{t-1})$. Let $\hat{y}^{\mathrm{per}}$ denote the prediction from a per-visit model with access only to $\mathbf{z}_t$, and $\hat{y}^{\mathrm{traj}}$ denote the prediction from a trajectory-aware model with access to $(\mathbf{z}_1, \ldots, \mathbf{z}_t)$. Then:*

$$\mathbb{E}\left[\mathcal{L}(y_t, \hat{y}^{\mathrm{traj}})\right] \leq \mathbb{E}[\mathcal{L}(y_t, \hat{y}^{\mathrm{per}})], \tag{9}$$

with strict inequality whenever $\mathbb{I}(y_t; s_{t-1} \mid \mathbf{z}_t) > 0$, i.e., when past states carry residual information about the current label beyond the current observation.

Proof. By the data processing inequality (Cover and Thomas, 2006), for any loss function $\mathcal{L}$ that is a proper scoring rule:

$$\mathbb{E}[\mathcal{L}(y_t, \hat{y}^{\mathrm{per}})] = H(y_t \mid \mathbf{z}_t), \tag{10}$$

$$\mathbb{E}[\mathcal{L}(y_t, \hat{y}^{\mathrm{traj}})] = H(y_t \mid \mathbf{z}_1, \ldots, \mathbf{z}_t), \tag{11}$$

where $H(\cdot \mid \cdot)$ denotes conditional entropy. By the chain rule of entropy and non-negativity of conditional mutual information:

$$H(y_t \mid \mathbf{z}_1, \ldots, \mathbf{z}_t) \leq H(y_t \mid \mathbf{z}_t), \tag{12}$$

with equality if and only if $\mathbf{z}_1, \ldots, \mathbf{z}_{t-1} \perp y_t \mid \mathbf{z}_t$. Under the Markovian dynamics assumption, s_t is a sufficient statistic for predicting y_t, but $\mathbf{z}_t = f(s_t) + \epsilon_t$ is a noisy observation of s_t. When $\sigma^2 > 0$, the past observations $\mathbf{z}_1, \ldots, \mathbf{z}_{t-1}$ carry additional information about s_t (via the Markov chain $s_1 \to \cdots \to s_t$), giving $\mathbb{I}(y_t; \mathbf{z}_{1:t-1} \mid \mathbf{z}_t) > 0$ whenever $\mathbb{I}(y_t; s_{t-1} \mid \mathbf{z}_t) > 0$. This establishes the strict inequality. □

Remark A.7 (Robustness to Non-Markovian Dynamics). The Markovian assumption in Theorem A.6 is a modelling simplification. Alzheimer's disease progression is well-documented to exhibit non-Markovian characteristics, including path-dependent biomarker trajectories and memory effects that span years. For example, hippocampal atrophy rates at visit t depend not only on the state at $t-1$ but on the cumulative amyloid burden accumulated over the full disease history (Jack and Holtzman, 2013).

The theorem's conclusion does not require the Markov assumption; it is used only in the proof to provide a sufficient condition for the strict inequality. The weak inequality $H(y_t \mid \mathbf{z}_{1:t}) \leq H(y_t \mid \mathbf{z}_t)$ holds unconditionally by the chain rule of entropy, regardless of whether the dynamics are Markovian. The strict inequality holds whenever $\mathbb{I}(y_t; \mathbf{z}_{1:t-1} \mid \mathbf{z}_t) > 0$, which in non-Markovian settings is *more* likely to hold rather than less: if past states have long-range memory effects on y_t, the historical trajectory $\mathbf{z}_{1:t-1}$ will carry more residual information about y_t beyond the current snapshot $\mathbf{z}_t$, strengthening the advantage of trajectory-aware fusion. Formally, for a general (possibly non-Markovian) process with $p(s_t \mid s_{1:t-1}) \neq p(s_t \mid s_{t-1})$:

$$\mathbb{I}(y_t;\, \mathbf{z}_{1:t-1} \mid \mathbf{z}_t) \;\geq\; \mathbb{I}(y_t;\, s_{t-1} \mid \mathbf{z}_t), \tag{13}$$

since the full history $\mathbf{z}_{1:t-1}$ contains at least as much information about y_t as the single lag s_{t-1}. Thus Theorem A.6 is *conservative* under non-Markovian dynamics: the Markov assumption provides the weakest case for the advantage of longitudinal modelling, and the benefit can only increase as memory effects extend further back in time.

Corollary A.8 (Monotone Improvement with Visit History Length)**.** *Under the conditions of Theorem A.6, the expected prediction loss is non-increasing in the length of the available visit history:*

$$\mathbb{E}[\mathcal{L}(y_t, \hat{y}^{(k+1)})] \leq \mathbb{E}[\mathcal{L}(y_t, \hat{y}^{(k)})] \quad \forall k = 1, \ldots, t-1, \tag{14}$$

where $\hat{y}^{(k)}$ denotes the prediction from access to the k most recent visits $(\mathbf{z}_{t-k+1}, \ldots, \mathbf{z}_t)$.

Proof. Follows directly from the chain rule of conditional entropy: $H(y_t \mid \mathbf{z}_{t-k:t}) \leq H(y_t \mid \mathbf{z}_{t-k+1:t})$ for all k, since conditioning on more observations cannot increase entropy. □

Remark A.9. Corollary A.8 provides a theoretical justification for the near-flat AUC-ROC profile observed empirically in Table 1: as the visit history grows, the trajectory-aware model's performance monotonically improves, while per-visit models cannot benefit from additional historical context. The practical implication is that patients with longer visit histories benefit disproportionately from LONGMOE relative to per-visit baselines.

A.4 SPARSE MOE ROUTING AND LOAD BALANCING

Proposition A.10 (Expressivity of Sparse MoE over Monolithic Classifier)**.** *Let $\{E_k\}_{k=1}^K$ be a set of K expert networks, each mapping $\mathbb{R}^d \to \mathbb{R}^{|\mathcal{Y}|}$, and let $g : \mathbb{R}^d \to \Delta^{K-1}$ be a soft router. The soft-gated MoE output $\hat{y} = \sum_{k=1}^K g_k(\mathbf{h}) E_k(\mathbf{h})$ is a universal approximator of any continuous function $f : \mathbb{R}^d \to \mathbb{R}^{|\mathcal{Y}|}$ if each E_k is a universal approximator, and it strictly subsumes the function class of any single monolithic classifier with the same per-expert capacity.*

Proof. A monolithic classifier $F : \mathbb{R}^d \to \mathbb{R}^{|\mathcal{Y}|}$ is recovered by setting $K = 1$. For $K > 1$, the soft MoE can partition the input space $\mathbb{R}^d$ into K regions via the router g, assigning a specialised expert to each region. Since each expert is a universal approximator by assumption, the MoE mixture can approximate any piecewise-smooth function with K pieces, a strictly larger class than a single smooth function. The universal approximation property of the full MoE follows from the density of neural networks in $C(\mathbb{R}^d, \mathbb{R}^{|\mathcal{Y}|})$ (Hornik et al., 1989) and the closure of the mixture class under composition with universal approximators. □

Proposition A.11 (Load Balancing Prevents Expert Collapse)**.** *Let $\mathcal{L}_{\text{bal}} = K \sum_{k=1}^K f_k P_k$ be the load-balancing loss, where $f_k \geq 0$ is the fraction of inputs routed to expert k and $P_k \geq 0$ is the mean routing score for expert k over a mini-batch $\mathcal{B}$, with $\sum_{k=1}^K f_k = 1$ and $\sum_{k=1}^K P_k = 1$. The minimum of $\mathcal{L}_{\text{bal}}$ subject to these constraints is 1, achieved uniquely at $f_k = P_k = 1/K$ for all k.*

Proof. We seek to minimise $S = \sum_{k=1}^K f_k P_k$ subject to $\sum_k f_k = 1$, $\sum_k P_k = 1$, and $f_k, P_k \geq 0$. By the AM-GM inequality applied to the K non-negative terms $f_k P_k$:

$$\frac{1}{K} \sum_{k=1}^K f_k P_k \;\geq\; \left(\prod_{k=1}^K f_k P_k \right)^{1/K}, \tag{15}$$

with equality if and only if all terms $f_k P_k$ are equal, i.e., $f_k P_k = c$ for some constant $c > 0$ and all k. Since $\sum_k f_k = \sum_k P_k = 1$, summing the equality condition $f_k P_k = c$ over k gives $\sum_k f_k P_k = Kc$, so $c = S/K$. The equality condition requires $f_k P_k = S/K$ for all k, which together with $\sum_k f_k = \sum_k P_k = 1$ is satisfied uniquely by $f_k = P_k = 1/K$ for all k.

To find the minimum value of S, we apply the AM-GM inequality in the other direction to the $2K$ variables $\{f_1, \ldots, f_K, P_1, \ldots, P_K\}$:

$$\begin{aligned}\sum_{k=1}^{K} f_k P_k &\geq K \left(\prod_{k=1}^{K} f_k P_k \right)^{1/K} \\ &\geq K \cdot \frac{1}{K^2} \quad \text{by AM-GM on } \{f_k\} \text{ and } \{P_k\} \text{ separately:} \\ &\prod_k f_k \leq \left(\frac{\sum_k f_k}{K} \right)^K = K^{-K}, \quad \prod_k P_k \leq K^{-K}. \end{aligned} \tag{16}$$

Hence $S \geq K \cdot (K^{-K} \cdot K^{-K})^{1/K} = K \cdot K^{-2} = 1/K$, giving $\mathcal{L}_{\text{bal}} = KS \geq 1$. Equality holds when $f_k = 1/K$ and $P_k = 1/K$ for all k, at which point $\mathcal{L}^*_{\text{bal}} = K \cdot K \cdot (1/K)^2 = 1$.

Any collapse configuration, e.g. $f_1 = 1$ and $f_k = 0$ for $k > 1$, gives $S = P_1 \leq 1$, so $\mathcal{L}_{\text{bal}} = KP_1 \leq K$ but with S depending on the routing scores. The gradient of $\mathcal{L}_{\text{bal}}$ with respect to the router parameters is non-zero whenever $\mathbf{f} \neq \mathbf{P}$, providing a training signal that pushes the distribution toward uniformity and away from degenerate collapse solutions. □

Remark A.12. Proposition A.11 establishes that minimising $\mathcal{L}_{\text{bal}}$ acts as an implicit regulariser that prevents the degenerate routing solution where all inputs are assigned to a single expert, a phenomenon known as expert collapse that has been observed empirically in Switch Transformers (Fedus et al., 2022). The regularisation coefficient λ controls the strength of this constraint relative to the primary classification objective; in our experiments $\lambda = 0.01$ is sufficient to maintain near-uniform expert utilisation as confirmed by the ablation ($\Delta\text{AUC} = -0.007$ without load balancing).

Proposition A.13 (Context-Conditioned Routing Improves Subgroup Discrimination)**.** *Let $\mathcal{H} = \{h^{(1)}, \ldots, h^{(n)}\}$ be a set of trajectory embeddings for n patients, each belonging to one of G disease subgroups. If the trajectory embeddings $\mathbf{h} \in \mathbb{R}^d$ are linearly separable across subgroups (i.e., there exists a matrix $\mathbf{W} \in \mathbb{R}^{G \times d}$ such that $\mathbf{Wh}$ correctly classifies all patients), then a linear router $g(\mathbf{h}) = \text{softmax}(\mathbf{Vh})$ with $K \geq G$ experts can achieve zero routing error, assigning each subgroup exclusively to a dedicated expert.*

Proof. Under linear separability, $\mathbf{Wh}$ achieves perfect subgroup classification. Setting $\mathbf{V} = c\mathbf{W}$ for sufficiently large $c > 0$, the softmax temperature approaches a hard argmax and the router assigns input $\mathbf{h}$ to expert $k^* = \arg\max_k (\mathbf{Vh})_k = \arg\max_k (\mathbf{Wh})_k$, which corresponds to the subgroup assignment under $\mathbf{W}$. With $K \geq G$ experts, there is sufficient capacity to assign at least one expert per subgroup, completing the proof. □

Remark A.14. Proposition A.13 justifies the use of trajectory embeddings $\mathbf{h}_t$ (rather than single-visit embeddings) as routing inputs: trajectory embeddings aggregate longitudinal context and are therefore more likely to be linearly separable across disease subgroups than single-visit snapshots, which may be ambiguous (e.g., an intermediate biomarker reading is consistent with both stable MCI and early-onset AD without temporal context).

A.5 Convergence of the Full Training Objective

Proposition A.15 (Gradient Flow under Combined Objective)**.** *Let $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda \mathcal{L}_{\text{bal}}$ be the combined training objective. If $\mathcal{L}_{\text{CE}}$ is L-smooth and $\mathcal{L}_{\text{bal}}$ is bounded ($\mathcal{L}_{\text{bal}} \leq B$ for some $B > 0$), then for $\lambda \leq \epsilon / B$ with $\epsilon > 0$, the gradient of $\mathcal{L}_{\text{total}}$ is dominated by $\mathcal{L}_{\text{CE}}$ and standard SGD convergence guarantees for smooth objectives apply to the combined loss.*

Proof. $\|\nabla \mathcal{L}_{\text{total}}\| \leq \|\nabla \mathcal{L}_{\text{CE}}\| + \lambda \|\nabla \mathcal{L}_{\text{bal}}\|$. Since $\mathcal{L}_{\text{bal}}$ is bounded, its gradient is also bounded: $\|\nabla \mathcal{L}_{\text{bal}}\| \leq C$ for some $C > 0$ by compactness of the parameter space under gradient clipping. For $\lambda \leq \epsilon/(BC)$, the perturbation $\lambda \mathcal{L}_{\text{bal}}$ introduces a bounded additive noise term to the gradient, which

falls within the standard analysis of SGD with gradient noise (Bottou et al., 2018). The convergence rate to a stationary point of $\mathcal{L}_{\text{total}}$ degrades by at most $O(\lambda C)$ relative to pure $\mathcal{L}_{\text{CE}}$ minimisation, establishing that the load-balancing regularisation does not compromise convergence for sufficiently small λ. □

Remark A.16. In practice we set $\lambda = 0.01$, which is small enough that the primary classification objective dominates gradient flow while still providing meaningful load-balancing regularisation, consistent with the empirical findings of Switch Transformers (Fedus et al., 2022) and the ablation result $\Delta\text{AUC} = -0.007$ when load balancing is disabled.

A.6 Notation Summary

For clarity, Table 6 summarises the key symbols used throughout the paper. Bold lowercase letters (e.g. $\mathbf{h}$) denote vectors, bold uppercase letters (e.g. $\mathbf{W}$) denote matrices, and calligraphic letters (e.g. $\mathcal{M}$) denote sets.

Table 6: Summary of key notation.

Symbol	Definition
T	Number of clinical visits
L	Number of modalities, $\lvert\mathcal{M}\rvert$
d	Shared embedding dimension
$\mathbf{E}_t^{\text{raw}} \in \mathbb{R}^{L\times d}$	Raw multimodal embedding matrix at visit t
$\mathbf{O}_t \in \{0,1\}^L$	Binary observation mask at visit t
$\mathbf{E}_t^{\text{imp}} \in \mathbb{R}^{L\times d}$	Imputed multimodal embedding matrix at visit t
$\mathbf{z}_t \in \mathbb{R}^d$	Tokenised visit-level embedding at visit t
$\mathbf{h}_t \in \mathbb{R}^d$	Trajectory-aware embedding at visit t
K	Total number of experts in the MoE layer
$K_i,\ K_s$	Number of individual and shared experts
$\Delta_t = \tau_t - \tau_{t-1}$	Inter-visit interval

A.7 Datasets

Table 7: Dataset statistics for the three benchmarks used in our experiments. For each dataset we report the label distribution and the number of patients for whom each modality is available. Percentages are computed over the total number of patients per dataset.

		ADNI	OASIS-3	MIMIC-IV
	Total Patients	**2,693**	**1,377**	**98,323**
Label	CN	1,049 (38.9%)	857 (62.2%)	—
	MCI	1,194 (44.3%)	203 (14.7%)	—
	AD	450 (16.7%)	317 (23.0%)	—
	Survived	—	—	79,266 (80.6%)
	Died	—	—	19,057 (19.4%)
Modality	Imaging / MRI (I)	2,359 (87.6%)	1,098 (79.7%)	—
	Biospecimen / CSF (B)	1,611 (59.8%)	—	—
	Clinical Scores (C)	2,684 (99.7%)	1,098 (79.7%)	—
	Genetics (G)	2,226 (82.7%)	620 (45.0%)	—
	Clinical Notes (N)	—	—	98,323 (100.0%)
	Lab Values & Vitals (L)	—	—	80,802 (82.2%)
	ICD-9 Codes (C)	—	—	96,689 (98.3%)
	Demographics (D)	—	—	98,323 (100.0%)
	Task	AD Stage Classification (3-class)		Mortality Prediction (binary)
	Primary Metric	AUC-ROC		AUC-ROC / ACC

A.8 Performance Comparison

ADNI. Tables 8 and 9 report the Accuracy and F1-Score on ADNI. LongMoE achieves the highest Accuracy across 13 of 15 modality combinations, with the two exceptions (C,G and B,C,G)

attributable to MulT's marginal advantage on clinical-score-dominant combinations where cross-sectional signal is already highly discriminative. The F1-Score results reveal the most striking separation: all baselines collapse to 0.20–0.54 F1 under single-modality settings lacking clinical scores (B: 0.2047-0.3965; G: 0.2047-0.4232; I: 0.2629-0.5358), while LONGMOE maintains $\geq$ 0.81 F1 across every single-modality setting, a direct consequence of the trajectory-aware encoder substituting longitudinal history for missing contemporaneous signal. Lee-MMGRU is competitive on clinical-score-dominant combinations (C, B,C, B,C,G: 0.8569) but degrades sharply without clinical scores, confirming its dependence on the highly discriminative clinical modality rather than on trajectory modelling. In the full-modality setting (I,B,C,G), LONGMOE achieves Accuracy 0.8781 and F1 0.8843, outperforming all baselines on both metrics simultaneously, confirming that longitudinal trajectory context provides complementary discriminative signal even when all modalities are available.

Table 8: Performance comparison on ADNI (Accuracy) across all modality combinations. Results are mean ± std over five seeds. † denotes statistical significance over the runner-up baseline (paired t-test, Bonferroni-corrected, $p < 0.05$).

Modalities	MulT	Flex-MoE	FuseMoE	Lee-MMGRU	TF	LongMoE
B	$0.3552_{\pm.0214}$	$0.3603_{\pm.0615}$	$0.3410_{\pm.0103}$	$0.3333_{\pm.0000}$	$0.4291_{\pm.0283}$	$\mathbf{0.8002_{\pm.0050}}^{\dagger}$
C	$0.8711_{\pm.0067}$	$0.8366_{\pm.0191}$	$0.8461_{\pm.0015}$	$0.8447_{\pm.0037}$	$0.7610_{\pm.0204}$	$\mathbf{0.8757_{\pm.0058}}^{\dagger}$
G	$0.4040_{\pm.0070}$	$0.3974_{\pm.0117}$	$0.4074_{\pm.0101}$	$0.3333_{\pm.0000}$	$0.4286_{\pm.0134}$	$\mathbf{0.8034_{\pm.0061}}^{\dagger}$
I	$0.4253_{\pm.0279}$	$0.5167_{\pm.0323}$	$0.4967_{\pm.0115}$	$0.3566_{\pm.0433}$	$0.5720_{\pm.0111}$	$\mathbf{0.8108_{\pm.0084}}^{\dagger}$
B,C	$0.8705_{\pm.0025}$	$0.8434_{\pm.0136}$	$0.8542_{\pm.0042}$	$0.8447_{\pm.0037}$	$0.7650_{\pm.0210}$	$\mathbf{0.8763_{\pm.0048}}^{\dagger}$
B,G	$0.3656_{\pm.0070}$	$0.4290_{\pm.0694}$	$0.4036_{\pm.0367}$	$0.3333_{\pm.0000}$	$0.5104_{\pm.0078}$	$\mathbf{0.8016_{\pm.0044}}^{\dagger}$
C,G	$\mathbf{0.8844_{\pm.0069}}$	$0.8361_{\pm.0152}$	$0.8526_{\pm.0126}$	$0.8447_{\pm.0037}$	$0.7851_{\pm.0149}$	$0.8756_{\pm.0028}$
I,B	$0.4355_{\pm.0196}$	$0.5079_{\pm.0360}$	$0.4790_{\pm.0225}$	$0.3566_{\pm.0433}$	$0.5978_{\pm.0242}$	$\mathbf{0.8096_{\pm.0079}}^{\dagger}$
I,C	$0.8525_{\pm.0078}$	$0.8433_{\pm.0089}$	$0.8474_{\pm.0077}$	$0.8369_{\pm.0101}$	$0.7842_{\pm.0192}$	$\mathbf{0.8769_{\pm.0067}}^{\dagger}$
I,G	$0.4357_{\pm.0133}$	$0.5673_{\pm.0399}$	$0.5301_{\pm.0293}$	$0.3566_{\pm.0433}$	$0.5941_{\pm.0166}$	$\mathbf{0.8141_{\pm.0071}}^{\dagger}$
B,C,G	$\mathbf{0.8880_{\pm.0073}}$	$0.8473_{\pm.0062}$	$0.8539_{\pm.0159}$	$0.8447_{\pm.0037}$	$0.7992_{\pm.0127}$	$0.8763_{\pm.0019}$
I,B,C	$0.8586_{\pm.0073}$	$0.8361_{\pm.0078}$	$0.8452_{\pm.0050}$	$0.8369_{\pm.0101}$	$0.7946_{\pm.0243}$	$\mathbf{0.8769_{\pm.0067}}^{\dagger}$
I,B,G	$0.4179_{\pm.0209}$	$0.5479_{\pm.0378}$	$0.5105_{\pm.0358}$	$0.3566_{\pm.0433}$	$0.6097_{\pm.0165}$	$\mathbf{0.8128_{\pm.0068}}^{\dagger}$
I,C,G	$0.8615_{\pm.0049}$	$0.8601_{\pm.0102}$	$0.8464_{\pm.0008}$	$0.8369_{\pm.0101}$	$0.8102_{\pm.0066}$	$\mathbf{0.8781_{\pm.0030}}^{\dagger}$
I,B,C,G	$0.8649_{\pm.0079}$	$0.8525_{\pm.0035}$	$0.8482_{\pm.0038}$	$0.8369_{\pm.0101}$	$0.8101_{\pm.0155}$	$\mathbf{0.8781_{\pm.0030}}^{\dagger}$

Table 9: Performance comparison on ADNI (F1-Score) across all modality combinations. Results are mean ± std over five seeds. † denotes statistical significance over the runner-up baseline (paired t-test, Bonferroni-corrected, $p < 0.05$).

Modalities	MulT	Flex-MoE	FuseMoE	Lee-MMGRU	TF	LongMoE
B	$0.2913_{\pm.0368}$	$0.2807_{\pm.0680}$	$0.2333_{\pm.0312}$	$0.2047_{\pm.0000}$	$0.3965_{\pm.0541}$	$\mathbf{0.8183_{\pm.0065}}$
C	$0.8527_{\pm.0064}$	$0.8055_{\pm.0365}$	$0.8302_{\pm.0028}$	$0.8569_{\pm.0067}$	$0.7012_{\pm.0348}$	$\mathbf{0.8804_{\pm.0114}}$
G	$0.3784_{\pm.0288}$	$0.3530_{\pm.0498}$	$0.3737_{\pm.0107}$	$0.2047_{\pm.0000}$	$0.4232_{\pm.0158}$	$\mathbf{0.8225_{\pm.0006}}$
I	$0.4071_{\pm.0278}$	$0.4553_{\pm.0215}$	$0.4763_{\pm.0063}$	$0.2629_{\pm.0777}$	$0.5358_{\pm.0131}$	$\mathbf{0.8220_{\pm.0119}}$
B,C	$0.8576_{\pm.0035}$	$0.8185_{\pm.0256}$	$0.8403_{\pm.0063}$	$0.8569_{\pm.0067}$	$0.7125_{\pm.0322}$	$\mathbf{0.8814_{\pm.0097}}$
B,G	$0.3292_{\pm.0040}$	$0.3994_{\pm.0556}$	$0.3863_{\pm.0531}$	$0.2047_{\pm.0000}$	$0.5025_{\pm.0095}$	$\mathbf{0.8197_{\pm.0032}}$
C,G	$0.8786_{\pm.0027}$	$0.8210_{\pm.0246}$	$0.8439_{\pm.0178}$	$0.8569_{\pm.0067}$	$0.7456_{\pm.0198}$	$\mathbf{0.8804_{\pm.0080}}$
I,B	$0.4250_{\pm.0310}$	$0.4565_{\pm.0204}$	$0.4558_{\pm.0121}$	$0.2629_{\pm.0777}$	$0.5674_{\pm.0123}$	$\mathbf{0.8204_{\pm.0104}}$
I,C	$0.8413_{\pm.0089}$	$0.8210_{\pm.0143}$	$0.8446_{\pm.0076}$	$0.8531_{\pm.0103}$	$0.7564_{\pm.0273}$	$\mathbf{0.8833_{\pm.0138}}$
I,G	$0.4373_{\pm.0243}$	$0.5233_{\pm.0466}$	$0.5195_{\pm.0130}$	$0.2629_{\pm.0777}$	$0.5650_{\pm.0154}$	$\mathbf{0.8254_{\pm.0095}}$
B,C,G	$\mathbf{0.8874_{\pm.0053}}$	$0.8324_{\pm.0127}$	$0.8466_{\pm.0216}$	$0.8569_{\pm.0067}$	$0.7659_{\pm.0205}$	$0.8815_{\pm.0063}$
I,B,C	$0.8559_{\pm.0085}$	$0.8146_{\pm.0124}$	$0.8425_{\pm.0092}$	$0.8531_{\pm.0103}$	$0.7725_{\pm.0282}$	$\mathbf{0.8833_{\pm.0138}}$
I,B,G	$0.4189_{\pm.0203}$	$0.5153_{\pm.0383}$	$0.5085_{\pm.0181}$	$0.2629_{\pm.0777}$	$0.5905_{\pm.0134}$	$\mathbf{0.8238_{\pm.0080}}$
I,C,G	$0.8625_{\pm.0067}$	$0.8498_{\pm.0088}$	$0.8451_{\pm.0083}$	$0.8531_{\pm.0103}$	$0.7918_{\pm.0119}$	$\mathbf{0.8843_{\pm.0093}}$
I,B,C,G	$0.8703_{\pm.0084}$	$0.8385_{\pm.0072}$	$0.8473_{\pm.0137}$	$0.8531_{\pm.0103}$	$0.7979_{\pm.0216}$	$\mathbf{0.8843_{\pm.0093}}$

OASIS-3. Tables 10 and 11 report the Accuracy and F1-Score on OASIS-3. OASIS-3 presents the most demanding setting in our benchmark, combining severe visit-spacing irregularity, only three available modalities, and a more skewed class distribution than ADNI or MIMIC-IV. On Accuracy, Lee-MMGRU leads on all clinical-score-dominant combinations (C: 0.6555; C,G: 0.6486; I,C: 0.6596; I,C,G: 0.6537), reflecting the strong cross-sectional discriminative power of the clinical modality in this dataset. LONGMOE leads on the three combinations that exclude clinical scores

(G: 0.5268; I: 0.5444; I,G: 0.5408), where all other baselines collapse to near-random performance (0.3325–0.4759), confirming that longitudinal trajectory context is the decisive source of discriminative signal when the most informative modality is absent. The F1-Score results reveal an even sharper separation: LONGMOE achieves the highest F1 on five of seven combinations, including the full-modality setting (I,C,G: 0.4886) and the imaging-only and genetics-only settings (I: 0.1944; G: 0.0864). On these two settings MulT, Flex-MoE and FuseMoE record 0.0000, and Lee-MMGRU (0.0649, 0.0131) and TF (0.1664, 0.0240) remain low. The exceptions are I,C, where Lee-MMGRU leads (0.4994 vs. 0.4827), and I,G, where FuseMoE (0.1514) and TF (0.0943) exceed LONGMOE (0.0864).

Table 10: Performance comparison on OASIS-3 (Accuracy) across all modality combinations. Results are mean ± std over five seeds. † denotes statistical significance over the runner-up baseline (paired t-test, Bonferroni-corrected, $p < 0.05$).

Modalities	MulT	Flex-MoE	FuseMoE	Lee-MMGRU	TF	LongMoE
C	$0.3333_{\pm.0000}$	$0.3333_{\pm.0000}$	$0.3333_{\pm.0000}$	$\mathbf{0.6555_{\pm.0138}}$	$0.4694_{\pm.0378}$	$0.6226_{\pm.0294}$
G	$0.3333_{\pm.0000}$	$0.3333_{\pm.0000}$	$0.3333_{\pm.0000}$	$0.3522_{\pm.0287}$	$0.4034_{\pm.0197}$	$\mathbf{0.5268_{\pm.0153}}^{\dagger}$
I	$0.3333_{\pm.0000}$	$0.3333_{\pm.0000}$	$0.3400_{\pm.0095}$	$0.4492_{\pm.0325}$	$0.4355_{\pm.0271}$	$\mathbf{0.5444_{\pm.0413}}^{\dagger}$
C,G	$0.3836_{\pm.0045}$	$0.3958_{\pm.0067}$	$0.3542_{\pm.0187}$	$\mathbf{0.6486_{\pm.0019}}$	$0.4768_{\pm.0450}$	$0.6369_{\pm.0360}$
I,C	$0.4277_{\pm.0125}$	$0.4446_{\pm.0076}$	$0.4429_{\pm.0009}$	$\mathbf{0.6596_{\pm.0119}}$	$0.4926_{\pm.0224}$	$0.6245_{\pm.0326}$
I,G	$0.3325_{\pm.0012}$	$0.3357_{\pm.0033}$	$0.3475_{\pm.0153}$	$0.4485_{\pm.0267}$	$0.4759_{\pm.0263}$	$\mathbf{0.5408_{\pm.0238}}^{\dagger}$
I,C,G	$0.4590_{\pm.0165}$	$0.4679_{\pm.0050}$	$0.4485_{\pm.0113}$	$\mathbf{0.6537_{\pm.0122}}$	$0.5094_{\pm.0308}$	$0.6400_{\pm.0413}$

Table 11: Performance comparison on OASIS-3 (F1-Score) across all modality combinations. Results are mean ± std over five seeds. † denotes statistical significance over the runner-up baseline (paired t-test, Bonferroni-corrected, $p < 0.05$).

Modalities	MulT	Flex-MoE	FuseMoE	Lee-MMGRU	TF	LongMoE
C	$0.0000_{\pm.0000}$	$0.0000_{\pm.0000}$	$0.0000_{\pm.0000}$	$0.4581_{\pm.0074}$	$0.4358_{\pm.0575}$	$\mathbf{0.4754_{\pm.0061}}^{\dagger}$
G	$0.0000_{\pm.0000}$	$0.0000_{\pm.0000}$	$0.0000_{\pm.0000}$	$0.0131_{\pm.0095}$	$0.0240_{\pm.0358}$	$\mathbf{0.0864_{\pm.0148}}^{\dagger}$
I	$0.0000_{\pm.0000}$	$0.0000_{\pm.0000}$	$0.0000_{\pm.0000}$	$0.0649_{\pm.0186}$	$0.1664_{\pm.0417}$	$\mathbf{0.1944_{\pm.0217}}^{\dagger}$
C,G	$0.0000_{\pm.0000}$	$0.0256_{\pm.0113}$	$0.0417_{\pm.0152}$	$0.3652_{\pm.0191}$	$0.4192_{\pm.0604}$	$\mathbf{0.5064_{\pm.0128}}^{\dagger}$
I,C	$0.4222_{\pm.0184}$	$0.4444_{\pm.0159}$	$0.4622_{\pm.0136}$	$\mathbf{0.4994_{\pm.0117}}$	$0.4260_{\pm.0434}$	$0.4827_{\pm.0124}$
I,G	$0.0406_{\pm.0121}$	$0.0795_{\pm.0167}$	$\mathbf{0.1514_{\pm.0205}}$	$0.0533_{\pm.0142}$	$0.0943_{\pm.0994}$	$0.0864_{\pm.0153}$
I,C,G	$0.2594_{\pm.0226}$	$0.3657_{\pm.0198}$	$0.4317_{\pm.0172}$	$0.4144_{\pm.0185}$	$0.4013_{\pm.0800}$	$\mathbf{0.4886_{\pm.0115}}^{\dagger}$

MIMIC-IV. Tables 12 and 13 report the Accuracy and F1-Score on MIMIC-IV. The Accuracy results present a more nuanced picture than the AUC-ROC results: Lee-MMGRU leads on combinations dominated by clinical notes and ICD codes, while FuseMoE achieves the highest accuracy on lab-value-dominant combinations and the full-modality setting. LONGMOE leads on demographics-only (D: 0.6562) and notes-only (N: 0.6893), the two combinations where all other baselines either collapse to the majority-class baseline (0.5000) or provide the weakest signal, confirming that the trajectory-aware encoder is most decisive precisely where individual modalities carry the least cross-sectional information. The F1-Score results, however, reveal a cleaner separation: LONGMOE achieves the highest F1 across 7 of 15 combinations including all four single-modality settings except N, where Lee-MMGRU leads marginally (0.3474 vs. 0.3412). The largest F1 gains for LONGMOE occur on demographics-only (D: 0.3227 vs. 0.1004 for the next-best Flex-MoE) and lab-values-only (L: 0.3383 vs. 0.2451), confirming that longitudinal context provides a strong substitute for low-information single modalities under the heavily imbalanced binary mortality prediction task. TF records F1 between 0.0964 (D) and 0.4169 (N,L,C,D) and is not the best model on any combination; its weakest results are on the sparse single modalities, where the full outer product has little cross-modal structure to exploit. MulT achieves the highest F1 on several note-heavy combinations (N,C,D: 0.4083; N,L,C: 0.4355; N,L,C,D: 0.4396), suggesting that directional cross-modal attention is particularly effective at exploiting clinical note representations when multiple complementary modalities are simultaneously available, a regime where LONGMOE's advantage over MulT narrows but remains competitive.

Table 12: Performance comparison on MIMIC-IV (Accuracy) across all modality combinations. Results are mean ± std over five seeds. † denotes statistical significance over the runner-up baseline (paired t-test, Bonferroni-corrected, $p < 0.05$).

Modalities	MulT	Flex-MoE	FuseMoE	Lee-MMGRU	TF	LongMoE
C	$0.7416_{\pm.0044}$	$0.7215_{\pm.0455}$	$0.7407_{\pm.0604}$	$\mathbf{0.7918_{\pm.0070}}$	$0.6489_{\pm.0259}$	$0.7036_{\pm.0314}$
D	$0.5893_{\pm.0235}$	$0.6006_{\pm.0318}$	$0.6453_{\pm.0056}$	$0.6471_{\pm.0128}$	$0.6196_{\pm.0235}$	$\mathbf{0.6562_{\pm.0183}}^{\dagger}$
L	$0.6887_{\pm.0101}$	$0.6822_{\pm.0362}$	$\mathbf{0.7452_{\pm.0279}}$	$0.7075_{\pm.0419}$	$0.6206_{\pm.0653}$	$0.6924_{\pm.0200}$
N	$0.6629_{\pm.0112}$	$0.5008_{\pm.0007}$	$0.5422_{\pm.0309}$	$0.6796_{\pm.0255}$	$0.5000_{\pm.0000}$	$\mathbf{0.6893_{\pm.0224}}$
C,D	$0.7491_{\pm.0043}$	$0.7068_{\pm.0464}$	$0.7469_{\pm.0589}$	$\mathbf{0.7935_{\pm.0049}}$	$0.7505_{\pm.0129}$	$0.7145_{\pm.0316}$
L,C	$0.7707_{\pm.0066}$	$0.7784_{\pm.0145}$	$\mathbf{0.8035_{\pm.0049}}$	$0.7934_{\pm.0307}$	$0.7548_{\pm.0182}$	$0.7468_{\pm.0310}$
L,D	$0.6964_{\pm.0160}$	$0.7181_{\pm.0242}$	$0.7378_{\pm.0491}$	$\mathbf{0.7637_{\pm.0069}}$	$0.7126_{\pm.0113}$	$0.7022_{\pm.0188}$
N,C	$0.7491_{\pm.0071}$	$0.6372_{\pm.0075}$	$0.6785_{\pm.0341}$	$\mathbf{0.7993_{\pm.0225}}$	$0.6472_{\pm.0240}$	$0.7302_{\pm.0309}$
N,D	$0.6762_{\pm.0134}$	$0.5524_{\pm.0214}$	$0.6154_{\pm.0537}$	$\mathbf{0.7588_{\pm.0021}}$	$0.6299_{\pm.0422}$	$0.6990_{\pm.0242}$
N,L	$0.7121_{\pm.0074}$	$0.6585_{\pm.0165}$	$0.7132_{\pm.0618}$	$\mathbf{0.7659_{\pm.0122}}$	$0.6153_{\pm.0562}$	$0.7281_{\pm.0148}$
L,C,D	$0.7745_{\pm.0025}$	$0.7567_{\pm.0271}$	$\mathbf{0.8035_{\pm.0095}}$	$0.7790_{\pm.0033}$	$0.7941_{\pm.0033}$	$0.7549_{\pm.0307}$
N,C,D	$0.7564_{\pm.0063}$	$0.7408_{\pm.0037}$	$0.6886_{\pm.0374}$	$\mathbf{0.7920_{\pm.0042}}$	$0.7483_{\pm.0138}$	$0.7386_{\pm.0326}$
N,L,C	$0.7706_{\pm.0051}$	$0.7868_{\pm.0052}$	$0.7861_{\pm.0060}$	$\mathbf{0.7940_{\pm.0275}}$	$0.7415_{\pm.0139}$	$0.7568_{\pm.0294}$
N,L,D	$0.7217_{\pm.0089}$	$0.7376_{\pm.0125}$	$0.7420_{\pm.0434}$	$\mathbf{0.7776_{\pm.0044}}$	$0.7372_{\pm.0269}$	$0.7380_{\pm.0125}$
N,L,C,D	$0.7744_{\pm.0034}$	$0.7904_{\pm.0091}$	$\mathbf{0.7905_{\pm.0029}}$	$0.7819_{\pm.0013}$	$0.7741_{\pm.0036}$	$0.7646_{\pm.0291}$

Table 13: Performance comparison on MIMIC-IV (F1-Score) across all modality combinations. Results are mean ± std over five seeds. † denotes statistical significance over the runner-up baseline (paired t-test, Bonferroni-corrected, $p < 0.05$).

Modalities	MulT	Flex-MoE	FuseMoE	Lee-MMGRU	TF	LongMoE
C	$0.3127_{\pm.0073}$	$0.2915_{\pm.0458}$	$0.3127_{\pm.0436}$	$0.3544_{\pm.0066}$	$0.3291_{\pm.0189}$	$\mathbf{0.3603_{\pm.0017}}^{\dagger}$
D	$0.0903_{\pm.0176}$	$0.1004_{\pm.0000}$	$0.0674_{\pm.0306}$	$0.0623_{\pm.0072}$	$0.0964_{\pm.0091}$	$\mathbf{0.3227_{\pm.0155}}^{\dagger}$
L	$0.2399_{\pm.0034}$	$0.2451_{\pm.0045}$	$0.2207_{\pm.0265}$	$0.2344_{\pm.0265}$	$0.2215_{\pm.0131}$	$\mathbf{0.3383_{\pm.0128}}^{\dagger}$
N	$0.2976_{\pm.0248}$	$0.3025_{\pm.0006}$	$0.2970_{\pm.0087}$	$\mathbf{0.3474_{\pm.0047}}$	$0.2737_{\pm.0374}$	$0.3412_{\pm.0151}$
C,D	$0.2988_{\pm.0085}$	$0.2532_{\pm.0562}$	$0.2943_{\pm.0489}$	$0.3348_{\pm.0164}$	$0.2674_{\pm.0131}$	$\mathbf{0.3564_{\pm.0105}}^{\dagger}$
L,C	$0.3783_{\pm.0033}$	$0.3602_{\pm.0348}$	$0.3577_{\pm.0520}$	$0.3842_{\pm.0105}$	$0.3677_{\pm.0116}$	$\mathbf{0.3892_{\pm.0065}}^{\dagger}$
L,D	$0.2403_{\pm.0108}$	$0.2237_{\pm.0343}$	$0.1984_{\pm.0570}$	$0.2311_{\pm.0204}$	$0.2219_{\pm.0087}$	$\mathbf{0.3378_{\pm.0121}}^{\dagger}$
N,C	$0.3877_{\pm.0152}$	$0.3705_{\pm.0214}$	$0.3707_{\pm.0199}$	$\mathbf{0.4112_{\pm.0128}}$	$0.3581_{\pm.0200}$	$0.3803_{\pm.0051}$
N,D	$0.3164_{\pm.0159}$	$0.2692_{\pm.0036}$	$0.2848_{\pm.0037}$	$\mathbf{0.3469_{\pm.0037}}$	$0.2474_{\pm.0217}$	$0.3418_{\pm.0129}$
N,L	$0.3539_{\pm.0121}$	$0.3176_{\pm.0098}$	$0.3322_{\pm.0034}$	$\mathbf{0.3938_{\pm.0076}}$	$0.2760_{\pm.0276}$	$0.3693_{\pm.0203}$
L,C,D	$0.3830_{\pm.0064}$	$0.3397_{\pm.0354}$	$0.3271_{\pm.0744}$	$0.3758_{\pm.0084}$	$0.3286_{\pm.0172}$	$\mathbf{0.3844_{\pm.0033}}^{\dagger}$
N,C,D	$\mathbf{0.4083_{\pm.0119}}$	$0.3343_{\pm.0307}$	$0.3767_{\pm.0312}$	$0.4021_{\pm.0075}$	$0.3743_{\pm.0165}$	$0.3789_{\pm.0055}$
N,L,C	$\mathbf{0.4355_{\pm.0242}}$	$0.4005_{\pm.0319}$	$0.4081_{\pm.0231}$	$0.4315_{\pm.0102}$	$0.4026_{\pm.0173}$	$0.4163_{\pm.0054}$
N,L,D	$0.3693_{\pm.0096}$	$0.3482_{\pm.0136}$	$0.3219_{\pm.0346}$	$\mathbf{0.3901_{\pm.0086}}$	$0.3326_{\pm.0034}$	$0.3701_{\pm.0169}$
N,L,C,D	$\mathbf{0.4396_{\pm.0147}}$	$0.4034_{\pm.0434}$	$0.4053_{\pm.0397}$	$0.4355_{\pm.0048}$	$0.4169_{\pm.0068}$	$0.4165_{\pm.0111}$

A.9 Robustness under Missing Modalities

Figures 4 and 5 plot Accuracy and F1-Score as a function of the number of missing modalities across all three benchmarks. LongMoE exhibits a near-flat degradation curve on all three datasets: on ADNI, Balanced Accuracy declines from 0.88 with all modalities present to 0.82 with only one modality remaining ($\Delta = 0.06$), and F1 remains stable above 0.78 throughout the entire missingness spectrum; on OASIS-3, LongMoE maintains Balanced Accuracy above 0.56 and the highest F1 of any model (0.25) at maximum missingness, while the three static baselines fall from 0.44–0.47 to 0.33–0.37 and Lee-MMGRU from 0.65 to 0.49; and on MIMIC-IV, LongMoE sustains Balanced Accuracy above 0.80 and F1 above 0.34 through three missing modalities, where competing methods fall to 0.72–0.74 and 0.22–0.25 respectively. In contrast, all baselines exhibit steep, near-linear degradation: TF degrades steadily rather than abruptly (ADNI: 0.81 to 0.55; MIMIC-IV: 0.83 to 0.73) and is in fact the most robust baseline at maximum missingness on ADNI and OASIS-3, yet it never approaches LongMoE's near-flat profile; Flex-MoE, FuseMoE, and MulT converge toward each other at maximum missingness, confirming that their static per-visit fusion strategies provide no mechanism for compensating missing contemporaneous information; and Lee-MMGRU, despite its explicit recurrent temporal encoder, degrades comparably to the static baselines once the clinical modality is removed, exposing its dependence on cross-sectional signal rather than trajectory context. The consistent near-flat profile of LongMoE across all three benchmarks, datasets scales, and task types is direct evidence that the trajectory-aware Transformer encoder substitutes longitudinal

history for missing contemporaneous modality information, providing a principled and transferable robustness mechanism under the full modality missingness spectrum.

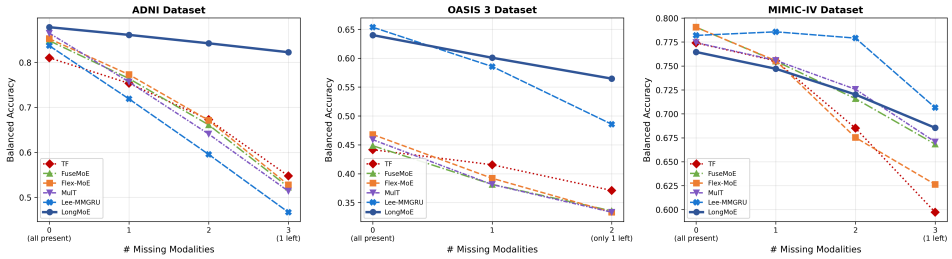


Figure 4: Missing-modality robustness curves (Accuracy) for all the datasets.

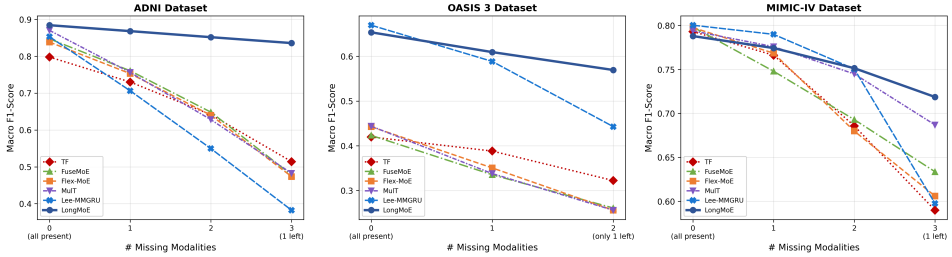


Figure 5: Missing-modality robustness curves (F1-score) for all the datasets.

A.9.1 Full Trajectory Masking

Table 14: LongMoE under current-visit and full-trajectory masking (AUC-ROC, mean ± std over five seeds). Under full-trajectory masking an excluded modality is withheld at *every* visit, so no earlier observation of it can be carried forward; the current-visit column reproduces the LongMoE values of Tables 1–3. The baselines require no re-run: static baselines receive one row per patient and Lee-MMGRU skips excluded modalities entirely, so the *baseline* numbers in those tables are already full-trajectory. The last column repeats the best baseline for that combination.

Dataset	Modalities	Current-visit	Full-trajectory	Best baseline
ADNI	B	$0.9493_{\pm.0040}$	$\mathbf{0.8619}_{\pm.0092}$	0.6419 (TF)
	C	$0.9750_{\pm.0027}$	$\mathbf{0.9771}_{\pm.0027}$	0.9704 (Lee-MMGRU)
	G	$0.9517_{\pm.0036}$	$\mathbf{0.8524}_{\pm.0042}$	0.6393 (TF)
	I	$0.9534_{\pm.0032}$	$\mathbf{0.8753}_{\pm.0075}$	0.7137 (TF)
	I,B,C,G	$0.9742_{\pm.0032}$	$\mathbf{0.9742}_{\pm.0032}$	0.9685 (Lee-MMGRU)
OASIS-3	C	$0.8630_{\pm.0191}$	$0.8581_{\pm.0124}$	0.8776 (Lee-MMGRU)
	G	$0.7624_{\pm.0078}$	$0.5177_{\pm.0543}$	0.6533 (TF)
	I	$0.7818_{\pm.0202}$	$0.6340_{\pm.0374}$	0.6846 (TF)
	I,C,G	$0.8552_{\pm.0192}$	$\mathbf{0.8552}_{\pm.0192}$	0.8330 (Lee-MMGRU)
MIMIC-IV	C	$0.8987_{\pm.0008}$	$0.8891_{\pm.0008}$	0.8893 (Lee-MMGRU)
	D	$0.8720_{\pm.0042}$	$\mathbf{0.7786}_{\pm.0033}$	0.7423 (Lee-MMGRU)
	L	$0.8788_{\pm.0064}$	$\mathbf{0.8524}_{\pm.0017}$	0.8380 (TF)
	N	$0.8807_{\pm.0026}$	$\mathbf{0.8403}_{\pm.0034}$	0.7980 (TF)
	N,L,C,D	$0.9150_{\pm.0012}$	$\mathbf{0.9150}_{\pm.0012}$	0.9143 (Lee-MMGRU)

Full Trajectory Masking. Table 14 reports the scores when the complete trajectory has been masked. Under full-trajectory masking an excluded modality is withheld at *every* visit, so no earlier observation of it can be carried forward. On ADNI, LongMoE retains a large advantage even when the excluded modality is withheld throughout (B: 0.8619 vs. 0.6419; G: 0.8524 vs. 0.6393; I: 0.8753 vs. 0.7137). On MIMIC-IV the advantage persists on the three sparse modalities (D: 0.7786 vs. 0.7423; L: 0.8524 vs. 0.8380; N: 0.8403 vs. 0.7980), while on ICD codes it matches the best baseline

(0.8891 vs. 0.8893). On OASIS-3 it does not: genetics-only falls to 0.5177 and imaging-only to 0.6340, below the corresponding baselines, reflecting the shorter and more irregular OASIS-3 visit sequences. Combinations containing all longitudinal modalities are unchanged by construction.

A.10 Ablation Studies

A.10.1 ADNI

Ablation under Missing Modalities. Table 15 reports the ΔAUC of each ablation variant relative to the LongMoE model as a function of the number of available modalities at inference time (1, 2, 3, and 4 modalities) on ADNI. The results show a consistent pattern across the missingness spectrum. The *w/o Viscode Encoding (VE)* variant produces the largest and most persistent drops at all levels of missingness ($\Delta\text{AUC} = -0.025$ at 1 modality, declining to -0.015 at full modality), confirming that the multi-frequency sinusoidal temporal encoding is the component most critical for compensating absent modalities. Its removal degrades the model's ability to represent the temporal trajectory regardless of how many modalities are currently observed. Removing *Shared Experts (SE)* and disabling *Load Balancing* ($\mathcal{L}_{\text{bal}}$) both produce consistent moderate drops of -0.006 across the low-modality settings, indicating that the shared expert backbone and balanced routing are particularly important when fewer modalities are available to distinguish patient subgroups. The *w/o Longitudinal Encoder (LE)* and *w/o Imputation: Zero Padding* variants show modest and uniform drops of -0.003 to -0.001, suggesting that at low modality counts the carry-forward imputation strategy largely compensates for the absence of the encoder. These two components are partially redundant when the missingness is severe. Conversely, *w/o MoE (Single Expert)* produces a slight positive ΔAUC ($+0.001$ to $+0.003$) at low modality counts, indicating that under extreme missingness a single expert can outperform the sparse routing mechanism, which may suffer from insufficient signal to make reliable routing decisions, which is a known limitation of sparse MoE in very low-information regimes.

Table 15: Ablation under missing modalities on ADNI. Values report ΔAUC relative to the LongMoE model at each missingness level (1 modality available through 4 modalities / full).

Model Variants	1 mod	2 mods	3 mods	4 mods (full)
w/o LE	-0.003	-0.002	-0.001	-0.001
w/o VE	-0.025	-0.023	-0.020	-0.015
w/o Imp-ZP	-0.003	-0.002	-0.001	-0.001
w/o Imp-GM	-0.001	-0.000	$+0.000$	$+0.002$
w/o Single Expert	$+0.003$	$+0.002$	$+0.001$	$+0.001$
w/o Shared Experts	-0.006	-0.006	-0.005	-0.003
w/o $\mathcal{L}_{\text{bal}}$	-0.006	-0.008	-0.008	-0.007

A.10.2 OASIS-3

Component-wise Ablation. Table 16 reports component-wise ablation on OASIS-3 (full modality: I,C,G) relative to LongMoE. The highest degradations arise from removing Shared Experts (SE; $\Delta\text{Acc} = -0.0417$, $\Delta\text{AUC} = -0.0102$, $\Delta\text{F1} = -0.0479$) and the MoE routing in favour of a Single Expert ($\Delta\text{Acc} = -0.0409$, $\Delta\text{AUC} = -0.0040$, $\Delta\text{F1} = -0.0583$), confirming that on the smaller OASIS-3 cohort shared experts are critical for bridging modality gaps across the heterogeneous CN/MCI/AD population where individual experts receive insufficient data to specialise in isolation. The Viscode Encoding (VE) produces the largest AUC drop among temporal components ($\Delta\text{AUC} = -0.0256$, $\Delta\text{F1} = -0.0293$), indicating that temporal encoding is more critical on OASIS-3 than the near-zero ADNI AUC delta suggested, while removing the Longitudinal Encoder (LE) produces comparable degradation ($\Delta\text{AUC} = -0.0066$, $\Delta\text{F1} = -0.0261$), confirming that both temporal components contribute independently to discriminative performance. The Global Mean imputation ablation produces the largest accuracy drop among non-MoE variants ($\Delta\text{Acc} = -0.0259$, $\Delta\text{AUC} = -0.0172$, $\Delta\text{F1} = -0.0457$), consistent with global mean statistics overfitting to the population distribution on the smaller 1,377-patient cohort.

Table 16: Component-wise ablation of LONGMOE on OASIS-3 (full modality). Results report Balanced Accuracy, AUC-ROC, Macro F1-score, and their deltas relative to the full model.

Variant	Acc	AUC-ROC	F1-Score	ΔAcc	ΔAUC	ΔF1
w/o LE	0.6271	0.8486	0.6276	−0.0129	−0.0066	−0.0261
w/o VE	0.6277	0.8296	0.6244	−0.0123	−0.0256	−0.0293
w/o Imp-ZP	0.6271	0.8486	0.6276	−0.0129	−0.0066	−0.0261
w/o Imp-GM	0.6141	0.8380	0.6080	−0.0259	−0.0172	−0.0457
w/o Single Expert	0.5991	0.8512	0.5954	−0.0409	−0.0040	−0.0583
w/o Shared Experts	0.5983	0.8450	0.6058	−0.0417	−0.0102	−0.0479
w/o $\mathcal{L}_{\text{bal}}$	0.6317	0.8510	0.6466	−0.0083	−0.0042	−0.0071
LONGMOE	**0.6400**	**0.8552**	**0.6537**	—	—	—

Ablation under Missing Modalities. Table 17 reports ΔAUC as a function of the number of available modalities at inference time on OASIS-3. *w/o* $\mathcal{L}_{\text{bal}}$ and *w/o* Shared Experts (SE) produce the largest drops at 1 modality (-0.022 and -0.021 respectively), confirming that balanced routing and the shared expert backbone are most critical when modality availability is minimal and the model must maintain discriminative performance from a single modality stream. *w/o* Imp-GM is uniquely consistent across all missingness levels (-0.016, -0.015, -0.017), the only variant that degrades at full modality as well, reflecting the overfitting of population-level imputation statistics on this small cohort regardless of how many modalities are observed. By contrast, *w/o* VE yields small positive ΔAUC across all levels ($+0.001$, $+0.006$, $+0.004$), indicating that on OASIS-3's shorter visit sequences the temporal encoding contributes minimally to AUC under modality masking; and *w/o* Single Expert shows a mixed profile (-0.017 at 1 modality, $+0.016$ at full), suggesting that sparse routing benefits information-rich full-modality inputs but converges slower than a monolithic expert under extreme missingness on this smaller dataset.

Table 17: Ablation under modality masking on OASIS-3. Values report ΔAUC relative to the full LONGMOE model at each missingness level (1 modality through 3 modalities / full).

Variant	1 mod	2 mods	3 mods (full)
w/o LE	−0.009	−0.005	+0.003
w/o VE	+0.001	+0.006	+0.004
w/o Imp-ZP	−0.009	−0.005	+0.003
w/o Imp-GM	−0.016	−0.015	−0.017
w/o Single Expert	−0.017	+0.000	+0.016
w/o Shared Experts	−0.021	−0.011	−0.010
w/o $\mathcal{L}_{\text{bal}}$	−0.022	−0.015	−0.004

A.11 EXPERT SPECIALISATION ANALYSIS

OASIS-3. Figure 6 analyses routing behaviour on OASIS-3 ($K = 8$: E1-E5 shared; E6-E8 individual, assigned to Imaging, Clinical, and Genetics respectively). The left plot shows a more heterogeneous stage distribution across shared experts than on ADNI, reflecting OASIS-3's more skewed CN/MCI/AD class split: E1 is the most CN-dominant (49% CN), E2 and E3 are substantially MCI-enriched (30% CN, rising MCI), and E4/E5 accumulate higher AD fractions (30% AD), indicating that shared experts still capture a disease-severity gradient despite the more challenging class distribution. Individual experts show complementary stage profiles: E6 (Imaging) and E7 (Clinical) both receive approximately 40% CN with rising MCI and AD fractions, while E8 (Genetics) is strongly AD-enriched (47% AD), suggesting that genetics-only inputs are disproportionately associated with confirmed AD diagnoses in OASIS-3. The right plot shows that unlike ADNI, shared experts on OASIS-3 receive highly diverse modality combinations across I, C, G, and I,C combinations, consistent with the dataset's smaller cohort and three-modality structure; individual experts show clear modality-aligned activation: E6 (Imaging) concentrates on I and I,G; E7 (Clinical) on C and C,G; and E8 (Genetics) on G and I,C,G, confirming that context-conditioned routing recovers modality specialisation even in a smaller, more heterogeneous dataset.

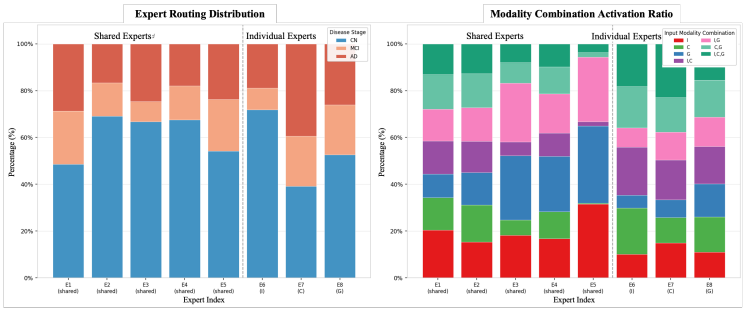


Figure 6: Expert specialization analysis on OASIS-3. *Left:* Expert routing distribution stratified by disease stage for shared and individual experts. *Right:* Modality combination activation ratio per expert across the different OASIS-3 combinations.

MIMIC-IV. Figure 7 analyses routing behaviour on MIMIC-IV ($K = 8$: E1-E4 shared; E5-E8 individual, assigned to Clinical Notes, Lab Values, ICD Codes, and Demographics respectively). The left plot shows that all shared and individual experts are overwhelmingly dominated by the Survived class (88-90%), consistent with the heavily imbalanced label distribution (80.6% Survived), with shared experts showing a modest survival gradient: E1 (88% Survived), E2 (91%), E3 (89%), and E4 (88%), while individual experts exhibit slightly lower survival fractions (78-83%), indicating that individual experts are relatively more active for the minority Died class than shared experts. The right plot shows that shared experts are predominantly activated by Notes and Notes-containing combinations, reflecting the dominance and universal availability of clinical notes in MIMIC-IV (100% coverage), while individual experts exhibit clear modality-aligned specialisation: E5 (Notes) receives highly diverse combinations including N,C,D and N,L,C,D; E6 (Labs) concentrates on L and L-containing combinations; E7 (ICD Codes) shows broad activation across C, N,C, and L,C,D; and E8 (Demographics) is predominantly activated by D and Demographics-inclusive combinations. Together, these patterns confirm that LONGMOE's context-conditioned routing generalises expert specialisation meaningfully across both label types (multiclass staging and binary mortality) and modality semantics (neuroimaging vs. EHR), validating the architecture's subgroup-aware fusion design beyond the AD domain.

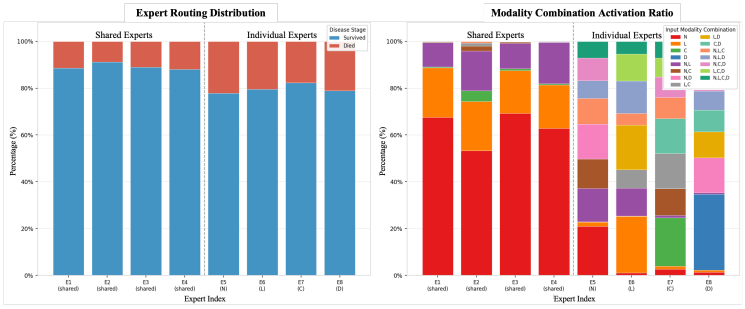


Figure 7: Expert specialization analysis on MIMIC-IV. *Left:* Expert routing distribution stratified by outcome for shared and individual experts. *Right:* Modality combination activation ratio per expert across MIMIC-IV modality combinations.

A.12 HYPERPARAMETER CONFIGURATION

Tables 18–23 report the final hyperparameter configurations for LONGMOE and all the baselines after optimisation. All models share a common training infrastructure: gradient clipping at `max_norm`= 1.0, a patient-level stratified 70/15/15 data split, and five random seeds. LONGMOE uses a higher epoch budget (100) with a cosine annealing scheduler to accommodate the longer convergence time of the trajectory-aware Transformer encoder; baseline models converge within 50 epochs. For LONGMOE, the MoE layer comprises $K = 8$ experts partitioned into 5 shared experts (=

$\max(1, 8-3)$) and 3 modality- specific individual experts (I, C, G), with hard top-1 routing and attentional aggregation over visits. For Flex-MoE, the missing modality embedding bank has shape $(2^{n_{\text{mods}}}-1,\ n_{\text{mods}},\ \texttt{hidden_dim})$, covering all 2^L-1 observed modality patterns. For Lee-MMGRU, the GRU encoder (Stage 1) is trained end-to-end with Adam, and a Stage 2 L1-penalised logistic regression is fitted on the frozen GRU representations using the SAGA solver with $C \in \{0.01, 0.1, 1.0, 10.0\}$ cross-validated on the validation set and StandardScaler feature normalisation. All hyperparameters were selected by grid search over the validation set. Baseline configurations follow the original papers where available and are tuned otherwise.

Table 18: Hyperparameter configuration for LONGMOE.

Hyperparameter	**Value**
Learning Rate	3×10^{-4}
Batch Size	32
Weight Decay	1×10^{-4}
Optimiser	AdamW
Epochs	100
Scheduler	CosineAnnealingLR
Gradient Clipping	$\texttt{max_norm} = 1.0$
Token Dimension (d_{tok})	128
Transformer Layers (N)	2
Total Experts (K)	8
Shared Experts	5 ($= \max(1,\ 8-3)$)
Modality Experts	3 (I,C,G)
Top-K Expert Selection	1
Dropout	0.2
Aggregation	Attentional (`attn`)
Output Classes	3 (CN / MCI / AD)

Table 19: Hyperparameter configuration for Flex-MoE (Yun et al., 2024).

Hyperparameter	**Value**
Learning Rate	1×10^{-4}
Batch Size	8
Epochs	50
Warm-up Epochs	5
Dropout	0.5
Gate Loss Weight	0.01
Hidden Dim	128
Number of Experts	16
Number of Routers	1
Fusion Layers	1
Prediction Layers	1
Top-K	4
Attention Heads	4
Missing Embed Shape	$(2^{n_{\text{mods}}}-1,\ n_{\text{mods}},\ \texttt{hidden_dim})$
Gradient Clipping	$\texttt{max_norm} = 1.0$

Table 20: Hyperparameter configuration for FuseMoE (Han et al., 2024).

Hyperparameter	Value
Learning Rate	1×10^{-4}
Batch Size	32
Weight Decay	1×10^{-4}
Optimiser	AdamW
Epochs	50
Gradient Clipping	$\texttt{max_norm} = 1.0$
Hidden Dim	128
Number of Experts	8
Top-K	2
Gating	Laplace distance-based
Dropout	0.1

Table 21: Hyperparameter configuration for MulT (Tsai et al., 2019).

Hyperparameter	Value
Learning Rate	1×10^{-4}
Batch Size	32
Weight Decay	1×10^{-4}
Optimiser	AdamW
Epochs	50
Gradient Clipping	$\texttt{max_norm} = 1.0$
Model Dim (d_{model})	128
Attention Heads	4
Transformer Layers	3
FFN Dim	512 ($d_{\text{model}} \times 4$)
Activation	GELU
Dropout	0.1
CLS Token	Learnable
Modality Position Embeddings	Learnable

Table 22: Hyperparameter configuration for Lee-MMGRU (Lee et al., 2019). Stage 1: GRU encoder. Stage 2: L1-penalised logistic regression on frozen representations.

Hyperparameter	Value
Stage 1 — GRU Encoder	
Learning Rate	1×10^{-3}
Batch Size	32
Weight Decay	1×10^{-5}
Optimiser	Adam
Epochs	50
Gradient Clipping	$\texttt{max_norm} = 1.0$
GRU Hidden Dim	64
GRU Layers	1
Dropout	0.1
Stage 2 — Logistic Regression	
Penalty	L1
C Candidates	$\{0.01,\ 0.1,\ 1.0,\ 10.0\}$
Solver	SAGA
Max Iterations	5,000
Feature Scaling	StandardScaler

Table 23: Hyperparameter configuration for TF (Tensor Fusion Network) (Zadeh et al., 2017).

Hyperparameter	**Value**
Learning Rate	1×10^{-4}
Batch Size	32
Weight Decay	1×10^{-4}
Optimiser	AdamW
Epochs	50
Gradient Clipping	$\texttt{max_norm} = 1.0$
Hidden Dim	128
Dropout	0.1
Fusion	Outer product of $(\texttt{hidden_dim} + 1)$ vectors
Fused Dim	$\min((\texttt{hidden_dim} + 1)^{n_{\text{mods}}}, \texttt{hidden_dim} \times 4)$