Interpretable Longitudinal Disease Forecasting with Graph-based Multimodal Encoding and Population-aware Language Generation

Yuheng Fan[1†], Hongyi Qin[2†], Jianyang Xie[1], Zhengju Chen[3], Yimin Luo[4], Uazman Alam[2], Gregory Y.H. Lip[2], Li Cheng[5], Yalin Zheng[1], and He Zhao[1]

[1] Department of Eye and Vision Sciences, University of Liverpool, Liverpool, UK
[2] Department of Cardiovascular and Metabolic Medicine, University of Liverpool, Liverpool, UK
[3] Department of Radiology, Huaxi MR Research Center (HMRRC), West China Hospital of Sichuan University, Chengdu, China.
[4] Department of Radiology, Weill Cornell Medicine, New York City, US
[5] Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Canada
He.Zhao@liverpool.ac.uk

Abstract. Multimodal longitudinal records provide critical insights into disease evolution and support disease understanding and early intervention. However, existing methods struggle with irregular and incomplete multimodal data, overlook the population-level progressions, and offer limited interpretability for clinical forecasting. To address these challenges, we propose a unified multimodal longitudinal forecasting framework that jointly predicts future disease status while generating explanations. Our approach employs a multimodal language model (MLM) with a novel longitudinal language tuning scheme that learns from generated hierarchical interpretations, enabling forecasting alongside structured explanatory outputs. Furthermore, a multimodal bipartite graph module is proposed to model missing and heterogeneous records, while a dual-memory module retrieves population-level patterns to support report generation. Experimental results on the ADNI dataset demonstrate significant improvement of the proposed method in both disease forecasting and result interpretation compared with state-of-the-art methods.

Keywords: Longitudinal Forecasting · Graph Representation Learning · Population Memory · Multimodal Language Model.

1 Introduction

Alzheimer's disease (AD) is a neurodegenerative disorder and a leading cause of dementia, typically evolving from cognitively normal (NC) to mild cognitive impairment (MCI) and AD [10]. Accurate and interpretable forecasting is critical for early intervention and efficient clinical workflows [14].

† Yuheng Fan and Hongyi Qin contributed equally to this work.

Early longitudinal forecasting methods [12] focused on single-modality temporal modelling, showing that incorporating multiple visits improves disease prediction. Recent works [17, 19, 23] have extended this paradigm to multi-modality, leveraging clinical and imaging information to enhance performance. However, missing data remains an unavoidable challenge in real-world settings. Approaches address incomplete inputs through loss masking [18] or adaptive temporal modelling [3], or learnable padding for missing modalities [2]. Graph-based methods further model heterogeneous variables under missingness, such as GRAPE [26] with bipartite graphs, GraFITi [24] via graph edge prediction. Despite these advances, interpretability for clinical decision support remains limited. Recent generative models offer a new avenue for interpretation by synthesizing future images to visualize disease progression [25, 20, 5]. While visually intuitive, these outputs require expert interpretation and lack structured clinical reasoning. Meanwhile, large language models (LLMs) [15, 1, 11] provide flexible explanations and demonstrate the potential to support longitudinal clinical reasoning. However, existing LLM-based approaches are typically limited to single time-point data and lack hierarchical structural priors to guide explanation generation.

Therefore, we proposed a novel MLM-based disease forecasting framework named ENIGMA, which **EN**ables **I**nterpretable multimodal longitudinal modeling with bipartite **G**raph encoding and population **M**emory-**A**ugmented generation. ENIGMA is further enhanced by proposed hierarchical interpretations, improving forecasting performance while providing more meaningful and clinically relevant explanations. The contributions are summarized as follows: 1) To our best knowledge, this is the first MLM-based longitudinal forecasting framework for interpretable disease prediction with hierarchical explanation generation; 2) We design a bipartite graph encoding (BGE) module to encode incomplete multimodal longitudinal records into semantically aligned representations for better MLM understanding; 3) We propose a Dual Population Memory (DPM) module to capture multimodal disease evolution with textual impression, enabling population-aware forecasting and enhancing clinical relevant explanations.

2 Methods

We consider the longitudinal dataset $\mathcal{D} = \{(R_m^{1:T}, l_m^{1:T})\}_{m=1}^{M}$ containing T visits from M participants, where $R_m^{1:T}$ denotes the clinical records and $l_m^{1:T}$ represents three disease categories (0:NC, 1:MCI, 2:AD). Each record on i-th visit contains 3D MRI scans x_{img}^i, time-variant clinical variables x_{tvv}^i like cognitive assessment, time-invariant variables x_{tiv} such as demographic and genetic information, and time interval t^i between the i-th and the first visit. This work aims to leverage the data from $T-1$ visits to predict disease status at visit T and generate corresponding explanatory outputs. Fig. 1 provides an overview of the proposed framework, where our longitudinal forecasting model shown at the bottom and the dataset summary generation for its training at the top. Specifically, our model consists of three main components: the bipartite graph module (BGE) for multimodal data encoding; the dual population memory module (DPM) to

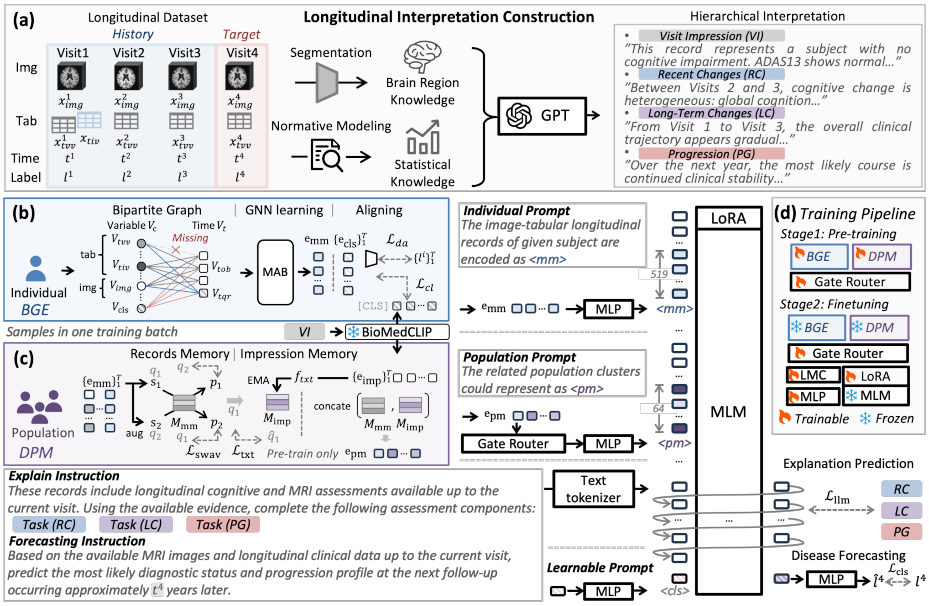

Fig. 1: Overview of proposed framework. (a) Workflow to construct interpretation from regional imaging and statistical knowledge; (b)Bipartite graph-based incomplete multimodal records encoding; (c) Dual-memory module for population-augmented language interpretation; (d) Pre-training and finetuning pipeline.

capture population-level evolutions; and a multimodal language model (MLM) that integrates all learned representations to generate disease forecasting and explanatory outputs. Each component is introduced in the following subsections.

2.1 Longitudinal interpretation construction

In clinical practice, disease progression is grounded in temporally documented evidence to ensure continuity and auditability [4]. Statistical approaches, such as normative modeling [16], quantify a normative reference from healthy populations and compare individualized deviations to indicate disease patterns. Specifically for Alzheimer's disease, progression knowledge manifests as longitudinal changes in brain regional structures and clinical assessments. To convert such evidence into language supervision, we extract brain volume quantiles from MRI using FreeSurfer [6] and estimate both structural and clinical normative deviations. Specifically, each deviation score is computed as the covariate-adjusted residual divided by the standard error of the estimate, thereby reflecting departures from normative trajectories. We then organize deviation measures into four different timescales: visit impression **(VI)** that describe normative status for each visit, recent changes **(RC)** that reflect the most recent evolution toward forecasting target, long-term changes **(LC)** that review historical pattern starting from

initial visits, and progression status **(PG)** which consider evolution between the most recent record and the target visit. To avoid manual summarization, we provide GPT-5.2 with the four-visit longitudinal records, including tabular data and diagnoses, along with pre-visit and inter-visit normative results. This enables GPT-5.2 to summarize textual explanations at each timescale, which are then used as supervision signals to train our MLM for explanation generation.

2.2 Bipartite Graph for Multimodal Records Encoding

BGE encodes individual records and infers an embedding $e_{mm} \in \mathbb{R}^{1\times N_m \times d_m}$ with $N_m = 519$ tokens and $d_m = 512$ dimension at query time t^T. Inspired by [24], we represent records on a bipartite graph $G = (V,E)$ as shown in Fig. 1(b), where nodes V are concept placeholders, the connectivity between nodes reflects missingness, edges E store observed values and we predict the future visit as edge embeddings embeddings using an attention-based graph neural network(GNN).

In details, the node set V contains variable nodes V_c and time nodes V_t. Time nodes consist of observed nodes V_{tob} and one query node V_{tqr}, which indicate the observed time and query time, respectively. Variable nodes consist of time-variant V_{tvv} and invariant tabular nodes V_{tiv}, image patch nodes V_{img} and a learnable class node V_{cls}. All edges are connected variable nodes and time nodes. Specifically, we disconnect edges if no value is observed and we add edges between variable nodes and the query node, thus the network can aggregate all observations to represent the target visit. For node initial embeddings, we employ a learnable lookup table to encode variable-node indices to distinguish each variable, and we represent the all visit times as d_m dimension vectors using sinusoidal encoding. For query time edges, we initialize them with learnable parameters. For the initialization of observed edges, we use the 3D ViT-B/16 to encode MRI scans into $T-1$ groups of patch embeddings, which are then taken as edge embeddings between image patch nodes and observed time nodes. The observed variables are encoded into dense feature in $\mathbb{R}^{d_m}$ via 2-layer MLP.

These embeddings further pass along the graph G using 4 layers of multi-head attention blocks(MAB)[24]. Since time-invariant variables have interacted with image and time variant observations at each layer, for encoding efficiency, we collect the embedding from image and time-variant query edges after the last GNN layer, resulting 519 output tokens e_{mm} at query time t^T. The encoding of records at all time are denoted as $\{e_{mm}\}_1^T$. We additionally summarize each visit's status by inserting a single class node V_{cls} between the observed and query time nodes. These class edges are learnable initialized and their output embeddings summarize each visit's status. The query-time class-edge embedding e_{cls} is also used to initialize the learnable prompt `<cls>` for MLM-based forecasting.

$$\mathcal{L}_{\text{BGE}} = \frac{1}{T}\sum_{i=1}^{T}\left(\alpha_{\text{cl}}\mathcal{L}_{\text{cl}}^{i} + \alpha_{\text{da}}\mathcal{L}_{\text{da}}^{i}\right) \tag{1}$$

BGE aligns textual representations and disease status across visits. For each visit i, we encode the visit impression **VI** into text embedding e_{imp}^{i} using the frozen

text encoder of BioMedCLIP [28]. We then align BioMedCLIP's `cls` embedding and the embedding of the corresponding class edge using CLIP loss $\mathcal{L}_{\text{cl}}^{i}$ [21]. In addition, a two-layer classifier is trained with the cross-entropy loss $\mathcal{L}_{\text{da}}^{i}$ between the i-th prediction and label l^{i}. The overall pre-training objective is given in Eq. 1, with $\alpha_{\text{cl}} = 2.0$ and $\alpha_{\text{ca}} = 1.0$ set empirically.

2.3 Dual Population Memory for Population-aware forecasting

The DPM module maintains a record memory $M_{\text{mm}} \in \mathbb{R}^{P \times d_m}$ and an impression memory $M_{\text{imp}} \in \mathbb{R}^{P \times d_m}$. It captures P representative trajectories and their memory-level impressions from record embeddings $\{e_{\text{mm}}\}_1^T$ and text embeddings $\{e_{\text{imp}}\}_1^T$ across training batches. As shown in Fig. 1(c), we update M_{mm} using the swap strategy similar to [5]: we create an augmented view by applying dropout and noise to the record embeddings, and then project the two views to $s_1, s_2 \in \mathbb{R}^{d_m}$ for swap prediction. M_{mm} is trained with the SwAV loss in Eq. 2, wherer B is batch size, the predicted assignment of the given record is $p_* = \text{softmax}(s_*^{\top} \cdot M_{\text{mm}})$, and q_* is obtained using the Sinkhorn–Knopp algorithm.

$$\mathcal{L}_{\text{swav}}(s_1, s_2) = -\frac{1}{2}\mathbb{E}_B(q_1 \log p_2 + q_2 \log p_1) \tag{2}$$

$$\mathcal{L}_{\text{txt}} = 1 - \mathbb{E}_B \left[\cos(r,\, q_1 M_{\text{imp}}) + q_1^{\top} \log \hat{q}_1\right] \tag{3}$$

Record memory captures population dynamics in multimodal trajectory space, but interpretation also needs language-level semantics. We therefore maintain the impression memory that aligns with each memory prototype in M_{mm}. We first project text embeddings $\{e_{\text{imp}}\}_1^T$ into impression vectors $r \in \mathbb{R}^{d_m}$ with a MLP f_{txt}. Impression vectors are then collected based on assignments q_1 to update impression memory using EMA for stable semantic meanings. We further add disentanglement and assignment rematching losses as Eq. 3, to reduce semantic redundancy and assure each column in M_{imp} matches the corresponding record memory. Here, $\hat{q}_1 = \text{softmax}(r^{\top} \cdot M_{\text{imp}})$ represents the text-induced assignment.

$$\mathcal{L}_{\text{DPM}} = \mathcal{L}_{\text{swav}} + \mathcal{L}_{\text{txt}} \tag{4}$$

The DPM loss is defined in Eq. 4. We concatenate the records memory M_{mm} and the impression memory M_{imp} to form the dual memory $e_{\text{pm}} \in \mathbb{R}^{P \times 2d_m}$.

2.4 Multimodal language model forecasting

As shown in Fig. 1, our MLM first uses two-layer MLPs to adapt the individual embedding e_{mm} and the dual-population memory e_{pm} to the language-model embedding space. Under task instructions `Instr` to let MLM forecast recent changes, long-term changes, and progression status, the model generates the textual explanation $\hat{y}^{txt}$ and predicts the disease label $\hat{l}^T$ at the target time t^T. For disease status prediction, we follow language-action alignment [22] by using a learnable prompt `<cls>` as model's inputs and a two-layer classifier head on

Table 1: Quantitative comparison of different methods for category forecasting. Each metric reports its macro value.

	Image		Tabular		Image + Tabular					
Metric	GLIMNet [9]	LaTiM [27]	GraFITi [24]	TimeCHEAT [13]	UniLMMV [2]	DTGPT [15]	NoteLM [1]	ViT-B/16-LM	ResNet50-LM	Ours
ACC ↑	0.751	0.697	0.734	0.709	0.744	0.764	0.737	0.677	0.680	**0.842**
SEN ↑	0.714	0.697	0.719	0.721	0.708	0.723	0.709	0.621	0.644	**0.799**
SPE ↑	0.856	0.834	0.866	0.859	0.863	0.873	0.864	0.866	0.871	**0.912**
F1 ↑	0.746	0.678	0.717	0.730	0.710	0.726	0.702	0.622	0.642	**0.810**
AUC ↑	0.821	0.838	0.890	0.863	0.874	–	–	–	–	**0.893**

the last-token representation. To better exploit population patterns, we introduce a gated router jointly trained with MLM, where a sparsity regularization encourages the router to activate only 25% of the memories during forecasting. The MLM is trained to jointly minimize the language generation loss for **RC**/**LC**/**PG** explanations and focal loss with l^T disease prediction:

$$\mathcal{L}_{\text{MLM}} = \mathbb{E}_{\mathcal{D}}[-\log p_{\text{MLM}}(\hat{y}^{txt}, \hat{l}^T | \texttt{<mm>}, \texttt{<pm>}, \texttt{Instr}, \texttt{<cls>})] \tag{5}$$

Overall, we optimize the framework in two stages as shown in Fig. 1(d). We first pre-train BGE, DPM, and the gate router using Eq. 1 and Eq. 4. We then finetune the MLM while keeping BGE and DPM fixed with Eq. 5 to learn structured interpretations. We also perform 200 iterations of language-generation warm-up before joint optimization for stable finetuning.

3 Experiments

3.1 Setup Details

Dataset. Following the setting of [9], we split dataset in participant level and constructed 2342 clips (train/val/test:1764/281/297) containing 4-visit from 399 participants in the ADNI dataset [10]. Clips converting to AD before the fourth visit are excluded, as the purpose is to predict before any conversion. Each visit comprises T1 MRI scans along with six cognitive scores and demographic information. MRIs were processed with FreeSurfer [6] and registered to MNI152 space [7], then normalized and resampled to 128× 128×128. For MLM backbones and all baseline methods, `Qwen3-4B-Instruct` is used as the default model. AdamW optimizer is utilized to update the parameters, and LoRA [8] (rank 64) is applied to all LLMs.[6]

Metrics and Baselines The disease forecasting performance is evaluated using the 4-th visit diagnosis as ground-truth, measured by accuracy (ACC), sensitivity (SEN), specificity (SPE), F1-score, and AUC. Explanations are evaluated against the generated summary described in Sec. 2.1 using BLEU, METEOR, ROUGE-L, and BERTScore. Furthermore, the clinical correctness is assessed via LLM-based judge [29] with GPT-5.2 and ratings from two radiology clinicians. The clinical correctness is evaluated from six dimensions following [11],

[6] The detailed implementation is available at https://github.com/yhf42/ENIGMA.

Table 2: Comprehensive comparison of explanation generation quality.

Model	NLG Metrics				LLM Judge Metrics					
	BERT↑	BLEU↑	MET↑	RL↑	CS ↑	HF ↑	PF ↑	SS ↑	HP ↑	HL ↑
GPT (GT)	–	–	–	–	4.367	4.347	4.125	4.360	4.040	4.956
DTGPT	0.750	28.328	0.472	0.331	3.400	3.306	3.134	4.265	3.370	3.548
NoteLM	**0.755**	28.577	**0.475**	**0.345**	3.000	**3.484**	2.764	**4.336**	3.047	3.111
ResNet-LM	0.737	26.471	0.452	0.305	3.320	3.026	3.051	4.131	3.333	3.391
ViT-LM	0.741	27.112	0.454	0.311	3.309	3.158	3.064	4.122	3.340	3.374
Ours	0.747	**28.735**	0.471	0.329	**3.640**	3.221	**3.387**	4.215	**3.593**	**3.845**

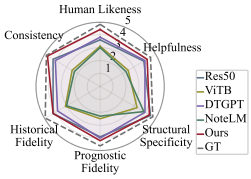

Fig. 2: Clinician-judged scores for methods and GT summary.

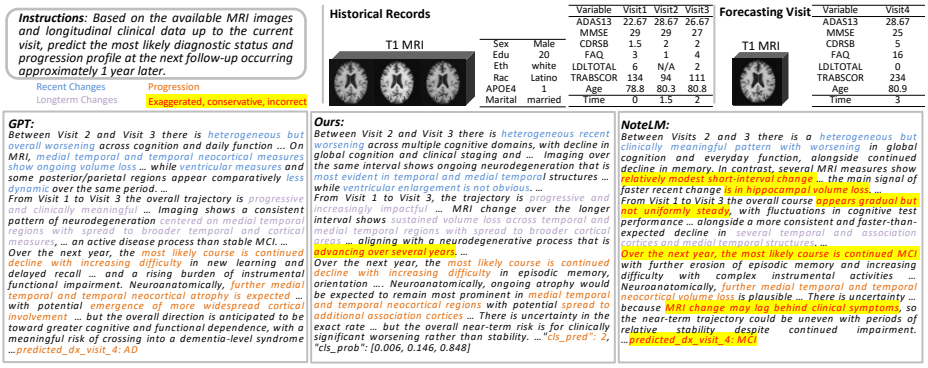

Fig. 3: Representative case study comparing explanatory outputs from our model and NoteLM against the reference explanation and next-visit diagnosis.

including: *Consistency* (CS), internal coherence and evidence consistency; *Historical Fidelity* (HF), accurate trajectory reflection and direction change in key cognitive and anatomical domains; *Prognostic Fidelity* (PF), clinically progression judgment aligned with the expected diagnostic direction; *Structural Specificity* (SS), system-level descriptions grounded in domains, functions, and regions; *Helpfulness* (HP), clear and actionable reasoning for diagnostic conclusion; and *Human-likeness* (HL), natural clinical writing style. Nine baselines are compared, including single-modality, multi-modality and VLMs.

3.2 Experimental Results

Quantitative results for disease forecasting and explanation. Table 1 presents the forecasting results. Our approach achieves the best performance across all metrics. The accuracy gains are significant ($p < 0.05$, McNemar's tests). The gains over multimodal and LLM/MLM-based methods highlight the importance of modelling longitudinal dynamics and population-level patterns for disease forecasting. Table 2 evaluates explanation generation quality using both NLG metrics and LLM-based judgement. Notably, DTGPT and NoteLM are unable to process MRI images, instead ground-truth volumetric measurements are directly provided as input. In contrast, other methods must extract this infor-

Table 3: Ablation study results on modality contribution, interpretation summary analysis, and MLM component design. BGE: BGE results with different input modalities. Full Pipeline: full pipeline ablation results.

	Setting	Image	Tabular	ACC↑	SEN↑	SPE↑	F1↑	AUC↑
BGE	Modality	✓	✗	0.478	0.524	0.753	0.459	0.702
		✗	✓	0.744	0.733	0.871	0.726	0.915
		✓	✓	0.822	0.784	0.900	0.793	0.932

	Setting	RC	LC	PG	ACC↑	SEN↑	SPE↑	F1↑	AUC↑
Full Pipeline	Interpret	✓	✗	✗	0.801	0.761	0.888	0.777	0.893
		✓	✓	✗	0.832	0.795	0.907	0.803	0.892

	Setting	$\mathcal{L}_{da}$	$\mathcal{L}_{cl}$	<cls>	M_{mm}	M_{imp}	ACC↑	SEN↑	SPE↑	F1↑	AUC↑	RL↑	MET↑	BLEU↑	BERT↑
Full Pipeline	MLM	✓	✗	✗	✗	✗	0.626	0.650	0.854	0.634	–	0.314	0.462	27.665	0.742
		✓	✓	✗	✗	✗	0.737	0.728	0.898	0.721	–	0.325	0.468	28.469	0.747
		✓	✓	✓	✗	✗	0.825	0.790	0.902	0.801	**0.905**	0.325	0.465	28.335	0.747
		✓	✓	✓	✓	✗	0.835	0.788	0.904	0.806	0.897	0.325	0.468	28.371	0.746
		✓	✓	✓	✓	✓	**0.842**	**0.799**	**0.912**	**0.810**	0.893	**0.329**	**0.471**	**28.735**	**0.747**

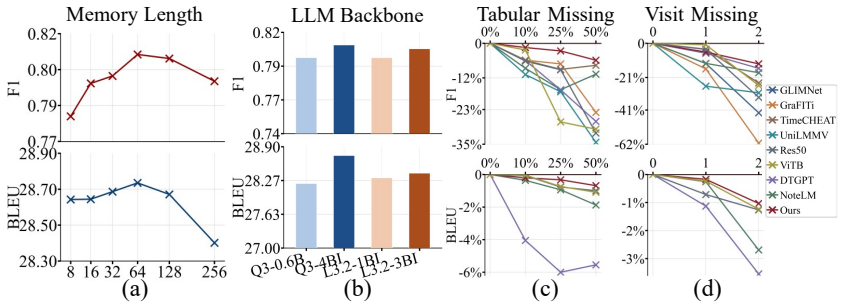

Fig. 4: (a) Memory length ablation; (b) LLM backbone scaling; (c) performance under missing tabular variables; (d) performance under missing-visit scenarios.

mation from imaging data. Consequently, the superior performance of NoteLM on some metrics is expected. Compared with image processing-based MLM, our approach achieves the best performance. Two clinicians further independently assessed interpretability of ten generated explanations and GT summaries (Fig. 2). Evidence fidelity is used as the primary criterion, and two clinicians achieve high agreement (Cohen's κ 0.79). Our approach demonstrates clear advantages.

Qualitative results for explanation generation. Fig. 3 compares reference explanations, our results align more closely with structural evidence and clearly distinguish recent changes from long-term changes, while the NoteLM (strongest baseline) tends to conflate them. For regional descriptions, such as medial temporal and temporal neocortical, our approach provides a more accurate degeneration description. Moreover, NoteLM places a narrow structural emphasis on the hippocampus, whereas our approach captures broader patterns of regional degeneration. Notably, although NoteLM identifies clinically meaningful deterioration, its final prediction is inconsistent.

Ablation study. See Table 3 and Fig. 4. The top panel shows that multimodal inputs are essential for disease forecasting. The middle panel further analyzes

structured interpretation summaries, demonstrating that multi-scale structured explanations improve forecasting performance. Additionally, the encoder trained with the generated VI by $\mathcal{L}_{da}$ achieves better results than training with CLIP loss $\mathcal{L}_{cl}$ alone. The bottom panel examines the effects of `<cls>` and DPM. Fig. 4(a) evaluates the memory module hyper-parameter, based on which a memory length of 64 is selected. Fig. 4(b) compares different MLM backbones, showing improved performance with stronger models. Robustness tests in (c) and (d) demonstrate our model stability under missing tabular variables and visit scenarios.

4 Conclusion

We present ENIGMA, an MLM-based framework for longitudinal disease forecasting with explanatory outputs. It integrates multimodal encoding and dual population memory to capture heterogeneous data and cohort-level patterns for enhanced predictive reasoning. Trained with automatically generated interpretation summaries, ENIGMA improves scalability without expert annotation. By producing structured explanations alongside predictions, it mitigates black-box limitations and enhances interpretability. Experiments on ADNI demonstrate improved forecasting accuracy and explanation quality.

Acknowledgments. This work is funded by the St Paul's Eye Research Foundation (CF349 25.26) and the Royal Society Fund (IES/R2/252401).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Alba, C., Xue, B., Abraham, J., Kannampallil, T., Lu, C.: The foundational capabilities of large language models in predicting postoperative risks using clinical notes. npj Digital Medicine **8**(1), 95 (2025)
2. Chen, B., Oliva, J., Niethammer, M.: A unified model for longitudinal multi-modal multi-view prediction with missingness. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 410–420. Springer (2024)
3. Cho, H., Ahn, H., Wu, G., Kim, W.H.: Adaptive adversarial data augmentation with trajectory constraint for alzheimer's disease conversion prediction. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 13–23. Springer (2025)
4. Cui, H., Unell, A., Chen, B., Fries, J.A., Alsentzer, E., Koyejo, S., Shah, N.H.: Timer: Temporal instruction modeling and evaluation for longitudinal clinical records. npj Digital Medicine **8**(1), 577 (2025)
5. Fan, Y., Xie, J., Luo, Y., Meng, Y., Madhusudhan, S., Lip, G.Y., Cheng, L., Zheng, Y., Zhao, H.: tHPM-LDM: Integrating individual historical record with population memory in latent diffusion-based glaucoma forecasting. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 619–629. Springer (2025)

6. Fischl, B.: Freesurfer. Neuroimage **62**(2), 774–781 (2012)
7. Fonov, V., Evans, A.C., Botteron, K., Almli, C.R., McKinstry, R.C., Collins, D.L., Group, B.D.C., et al.: Unbiased average age-appropriate atlases for pediatric studies. Neuroimage **54**(1), 313–327 (2011)
8. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. Iclr **1**(2), 3 (2022)
9. Hu, X., Zhang, L.X., Gao, L., Dai, W., Han, X., Lai, Y.K., Chen, Y.: Glim-net: chronic glaucoma forecast transformer for irregularly sampled sequential fundus images. IEEE Transactions on Medical Imaging **42**(6), 1875–1884 (2023)
10. Jack Jr, C.R., Bernstein, M.A., Fox, N.C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P.J., L. Whitwell, J., Ward, C., et al.: The alzheimer's disease neuroimaging initiative (ADNI): MRI methods. Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine **27**(4), 685–691 (2008)
11. Kwon, T., Ong, K.T.i., Kang, D., Moon, S., Lee, J.R., Hwang, D., Sohn, B., Sim, Y., Lee, D., Yeo, J.: Large language models are clinical reasoners: Reasoning-aware diagnosis framework with prompt-generated rationales. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 18417–18425 (2024)
12. Li, H., Fan, Y.: Early prediction of alzheimer's disease dementia based on baseline hippocampal MRI and 1-year follow-up cognitive measures using deep recurrent neural networks. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019). pp. 368–371 (2019)
13. Liu, J., Cao, M., Chen, S.: Timecheat: A channel harmony strategy for irregularly sampled multivariate time series analysis. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 18861–18869 (2025)
14. Maheux, E., Koval, I., Ortholand, J., Birkenbihl, C., Archetti, D., Bouteloup, V., Epelbaum, S., Dufouil, C., Hofmann-Apitius, M., Durrleman, S.: Forecasting individual progression trajectories in alzheimer's disease. Nature Communications **14**(1), 761 (2023)
15. Makarov, N., Bordukova, M., Quengdaeng, P., Garger, D., Rodriguez-Esteban, R., Schmich, F., Menden, M.P.: Large language models forecast patient health trajectories enabling digital twins. npj Digital Medicine **8**(1), 588 (2025)
16. Marquand, A.F., Kia, S.M., Zabihi, M., Wolfers, T., Buitelaar, J.K., Beckmann, C.F.: Conceptualizing mental disorders as deviations from normative functioning. Molecular psychiatry **24**(10), 1415–1424 (2019)
17. Muksimova, S., Umirzakova, S., Baltayev, J., Cho, Y.I.: Multi-modal fusion and longitudinal analysis for alzheimer's disease classification using deep learning. Diagnostics **15**(6), 717 (2025)
18. Nguyen, H.H., Blaschko, M.B., Saarakkala, S., Tiulpin, A.: Clinically-inspired multi-agent transformers for disease trajectory forecasting from multimodal data. IEEE transactions on medical imaging **43**(1), 529–541 (2023)
19. Ouyang, J., Zhao, Q., Adeli, E., Zaharchuk, G., Pohl, K.M.: SOM2LM: Self-organized multi-modal longitudinal maps. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 400–410. Springer (2024)
20. Puglisi, L., Alexander, D.C., Ravì, D., Initiative, A.D.N., et al.: Brain latent progression: Individual-based spatiotemporal disease progression on 3D brain MRIs via latent diffusion. Medical image analysis p. 103734 (2025)
21. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from

natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
22. Renz, K., Chen, L., Arani, E., Sinavski, O.: Simlingo: Vision-only closed-loop autonomous driving with language-action alignment. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 11993–12003 (2025)
23. Xu, L., Wu, H., He, C., Wang, J., Zhang, C., Nie, F., Chen, L.: Multi-modal sequence learning for alzheimer's disease progression prediction with incomplete variable-length longitudinal data. Medical image analysis **82**, 102643 (2022)
24. Yalavarthi, V.K., Madhusudhanan, K., Scholz, R., Ahmed, N., Burchert, J., Jawed, S., Born, S., Schmidt-Thieme, L.: GraFITi: Graphs for forecasting irregularly sampled time series. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 16255–16263 (2024)
25. Yoon, J.S., Zhang, C., Suk, H.I., Guo, J., Li, X.: SADM: Sequence-aware diffusion model for longitudinal medical image generation. In: International Conference on Information Processing in Medical Imaging. pp. 388–400. Springer (2023)
26. You, J., Ma, X., Ding, Y., Kochenderfer, M.J., Leskovec, J.: Handling missing data with graph representation learning. Advances in Neural Information Processing Systems **33**, 19075–19087 (2020)
27. Zeghlache, R., Conze, P.H., El Habib Daho, M., Li, Y., Le Boité, H., Tadayoni, R., Massin, P., Cochener, B., Rezaei, A., Brahim, I., et al.: LaTiM: Longitudinal representation learning in continuous-time models to predict disease progression. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 404–414. Springer (2024)
28. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., et al.: Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv preprint arXiv:2303.00915 (2023)
29. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al.: Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. Advances in neural information processing systems **36**, 46595–46623 (2023)