This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Supplementary Material for PolyAge-DT: Missingness-Aware Multimodal Representation Learning for Aging Digital Twins

Jiaxin Liu[1,2], Zheng Wei[2,3], Avery Shuo Wang[1,2], Yiwei Dong[2,4], Wansi Deng[2,4], Xianwen Shang[1,2], Xin Zhou[2,3,5,6], Xueli Zhang[7], Mingguang He[1,2,8,9], and Jing Zhang[1,2]

[1] School of Optometry, The Hong Kong Polytechnic University, Hong Kong, China
mingguang.he@polyu.edu.hk
jing-jan.zhang@polyu.edu.hk
[2] PolyU-Stanford Joint Research Centre for Longitudinal Deep Omics, The Hong Kong Polytechnic University, Hong Kong, China
[3] Center on Phenomics and Precision Medicine, Intelligent Medicine Institute, Fudan University, Shanghai, China
[4] Department of Health Technology and Informatics, The Hong Kong Polytechnic University, Hong Kong, China
[5] Department of Genetics, Stanford University School of Medicine, Stanford, CA, USA
[6] Department of Periodontics, Shanghai Stomatological Hospital, Fudan University, Shanghai, China
[7] Institute of Scientific Instrumentation, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China
[8] Research Centre for SHARP Vision, The Hong Kong Polytechnic University, Hong Kong, China
[9] Centre for Eye and Vision Research, 17W Hong Kong Science Park, Hong Kong, China

Supplementary Figures

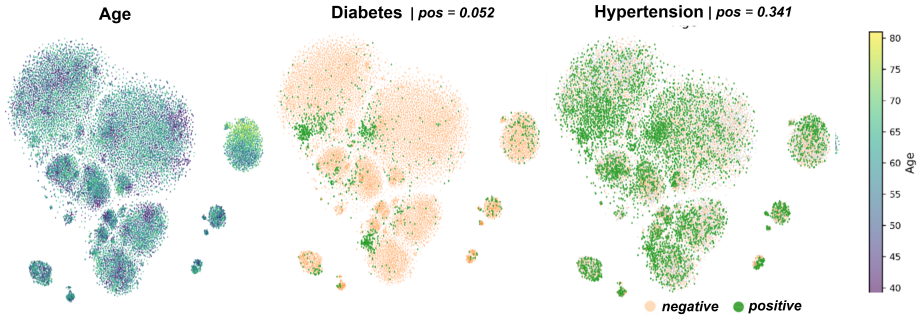

Fig. S1. Latent-space visualization of the fused PolyAge representation. t-SNE projection of test-set fused PolyAge representations annotated by chronological age and selected EHR-derived clinical labels, including diabetes and hypertension. The visualization illustrates aging-related and disease-relevant structure captured by the unified multimodal representation.

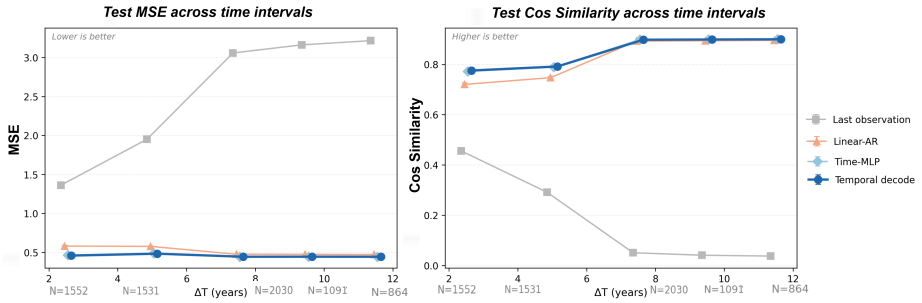

Fig. S2. Temporal prediction performance across age-delta horizons. Performance of the temporal decoder and baseline models across different intervals between historical and target visits. Compared with the last-observation baseline, learned temporal models maintained lower prediction error and higher cosine similarity across longer prediction horizons.

Supplementary Tables

Table S1. Comparison of fusion strategies across modality-availability conditions. Performance of PolyAge-DT fusion variants across participants grouped by the number of naturally available modalities. Age prediction is evaluated using mean absolute error (MAE), and disease classification is reported as the mean AUROC across 16 binary clinical endpoint classification tasks.

	Age prediction (MAE ↓)			Disease classification (AUROC ↑)		
Available modalities	Attention	Concatenation	Mean	Attention	Concatenation	Mean
Complete input	**3.3891**	4.0105	4.8659	**0.8123**	0.7978	0.7585
Three modalities ($k = 3$)	**3.1867**	3.7843	5.1419	**0.8337**	0.8190	0.7497
Two modalities ($k = 2$)	**3.4043**	3.9925	4.9319	**0.8127**	0.7979	0.7600
One modality ($k = 1$)	**4.2967**	4.9518	5.1923	**0.7447**	0.7349	0.7192

Table S2. Temporal future health-state prediction. Performance on the participant-held-out test set. The temporal decoder was compared with the last-observation baseline, Linear-AR, and Time-MLP.

Method	Parameters	MSE ↓	Cosine similarity ↑
Last observation	0	2.4817	0.1890
Linear-AR	18K	0.5212	0.8240
Time-MLP	73K	0.4577	0.8476
Temporal decoder (ours)	2.97M	**0.4537**	**0.8491**

Supplementary Equations

The following equations provide additional implementation details omitted from the main manuscript. Symbols are defined locally for clarity.

Artificial Modality Masking

$$\mathcal{V}_i = \mathcal{O}_i \setminus \mathcal{M}_i, \qquad |\mathcal{V}_i| \geq 1. \tag{S1}$$

where $\mathcal{O}_i$ denotes the originally available modality set for patient–visit i, $\mathcal{M}_i$ denotes the artificially masked subset during training, and $\mathcal{V}_i$ represents the remaining visible modalities provided to the model.

Reconstruction Objectives

$$\mathcal{L}_{\text{masked}} = \frac{\sum_i \sum_{m \in \mathcal{M}_i} \|\widehat{\mathbf{h}}_i^{(m)} - \mathbf{h}_i^{(m)}\|_2^2}{\sum_i |\mathcal{M}_i|} \tag{S2}$$

where m indexes modality types, $\mathbf{h}_i^{(m)}$ is the latent representation of modality m, and $\widehat{\mathbf{h}}_i^{(m)}$ is the reconstructed latent representation.

$$\mathcal{L}_{\text{visible}} = \frac{\sum_i \sum_{m \in \mathcal{V}_i} \|\widehat{\mathbf{h}}_i^{(m)} - \mathbf{h}_i^{(m)}\|_2^2}{\sum_i |\mathcal{V}_i|} \tag{S3}$$

The overall representation objective is:

$$\mathcal{L}_{\text{representation}} = \lambda_{\text{masked}} \mathcal{L}_{\text{masked}} + \lambda_{\text{visible}} \mathcal{L}_{\text{visible}} \tag{S4}$$

where λ_{masked} and λ_{visible} control the relative contributions of cross-modal inference and visible-modality information preservation, respectively.

Downstream Evaluation

$$\widehat{y}_i^{\text{age}} = f_{\text{age}}(\mathbf{z}_i) \tag{S5}$$

where $\mathbf{z}_i$ denotes the fused PolyAge representation and f_{age} is the age regression head.

$$\widehat{p}_{i,c} = \sigma(f_c(\mathbf{z}_i)) \tag{S6}$$

where c indexes clinical endpoints, f_c denotes the classification head, and $\sigma(\cdot)$ is the sigmoid activation function.

Fusion Baseline