Autoregressive EHR Foundation Models with Multimodal Inputs

Yuxuan "Edison" Liu [*1 2] **Joshua Placidi** [*1 2] **Jinpei Han** [*1] **Alfred John Balston** [1 2 3] **Marek Rei** [1]
A. Aldo Faisal [1 2 4 5]

Abstract

Autoregressive foundation models trained on tokenized electronic health records (EHRs) can support zero-shot clinical prediction, yet most operate on structured event codes alone, and do not incorporate multiple modalities in a principled way. We present a framework for conditioning such models on auxiliary clinical modalities, including ECG waveforms, chest X-ray images, and clinical notes, using modality-specific latent compression and gated cross-attention with temporal alignment. We investigate two key design choices: (1) how to compress long per-modality sequences (e.g., ECG time series) before they enter the multi-modal cross-attention. This feature may be essential to reduce compute overheads and may be beneficial for generalization; (2) how the choice of pretrained encoder for each modality impacts downstream performance. Through controlled ablations on MIMIC-IV, we show that the best latent-compression configurations outperforms both uncompressed cross-attention and mean pooling. Encoder choice has a clear within-modality effect, with stronger pretrained encoders consistently outperforming weaker alternatives. We further show that merely adding auxiliary modalities does not guarantee improvement on ICU mortality prediction over an EHR-only baseline. This implies that careful design of the fusion architecture and an appropriate evaluation in the clinical context are required.

*Equal contribution. [1]Dept. of Computing, Imperial College London, London, United Kingdom [2]UKRI Centres in AI for Health, United Kingdom [3]School of Public Health, Imperial College London, London, United Kingdom [4]Dept. of Bioengineering, Imperial College London, London, United Kingdom [5]Chair in Digital Health, Universität Bayreuth, Bayreuth, Germany. Correspondence to: Joshua Placidi <joshua.placidi25@imperial.ac.uk>, Yuxuan "Edison" Liu <edison.liu22@imperial.ac.uk>, A. Aldo Faisal <aldo.faisal@imperial.ac.uk>.

Proceedings of the Workshop on Structured Data for Health at the 43rd International Conference on Machine Learning, Seoul, South Korea. Copyright 2026 by the author(s).

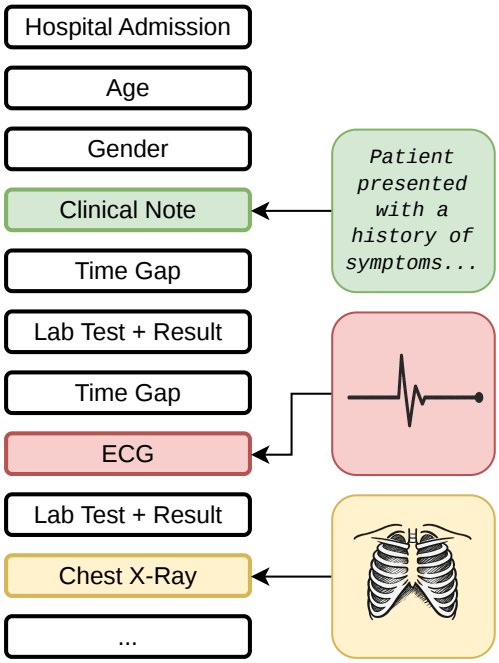

Figure 1. Electronic health records are represented as sequences of clinical events spanning multiple modalities.

1. Introduction

Recent work has shown that autoregressive transformers trained on tokenized electronic health records (EHRs) can learn reusable patient representations that support zero-shot clinical prediction (Renc et al., 2024; Waxler et al., 2025). These models linearize longitudinal EHR events (i.e. diagnoses, procedures, medications, vitals) into discrete token sequences and apply next-token prediction, analogous to language modeling (Pang et al., 2021). However, most EHR foundation models operate on a single modality of structured event codes, even though clinical decision-making routinely draws on complementary multimodal evidence as shown in Fig. 1. Electrocardiograms (ECG) provide direct measurements of cardiac electrical activity and support the detection of arrhythmias, conduction abnormalities, and ischemic patterns (Gu et al., 2025). Chest X-ray (CXR) images offer rapid evidence of pulmonary and cardiac pro-

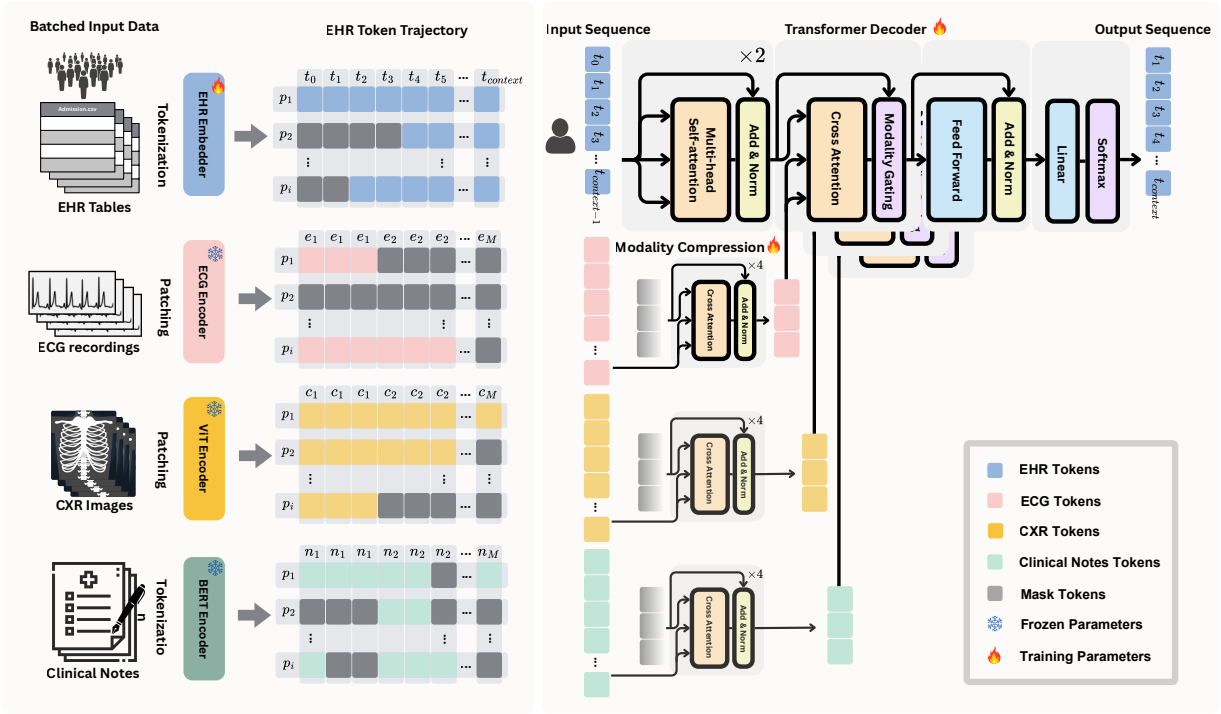

Figure 2. Multimodal EHR trajectory model. Auxiliary modalities (ECG, CXR, notes) are encoded, compressed into latent tokens, and fused into a GPT-2-style decoder via gated cross-attention. Temporal masking ensures each EHR token attends only to past modality events.

cesses such as consolidation, edema, cardiomegaly, and pleural effusions. Clinical notes capture information that is not well represented in coded tables, including presenting symptoms, differential diagnoses, and clinician assessments (Wang et al., 2024). Integrating these modalities into a generative EHR model is nontrivial.

Each modality differs in dimensionality, sampling rate, and temporal availability; missingness is clinically informative rather than random; and naive fusion risks temporal leakage when auxiliary observations acquired after an EHR event are allowed to influence its prediction (Spathis & Kawsar, 2024). Prior multimodal clinical work has largely adopted late fusion—aggregating predictions from independently trained unimodal models—which cannot capture fine-grained cross-modal dependencies (Stahlschmidt et al., 2022; Jorf & Shamout, 2025). Intermediate fusion via cross-attention, as popularized by Flamingo (Alayrac et al., 2022) and adapted for medical tasks (Moor et al., 2023; Tu et al., 2024; Amar et al., 2025), offers a more expressive alternative but has not been systematically evaluated for autoregressive EHR trajectory modeling.

In this work, we augment a decoder-only EHR backbone to use ECG, CXR, and clinical notes with latent compression and cross-attention. We propose a compression module, inspired by Perceiver (Jaegle et al., 2021), that compresses modality information into a small number of latents, improving model efficiency and performance. Through controlled ablations on MIMIC-IV, we show that latent compression is essential for effective and efficient fusion of long modality sequences, and that encoder choice has a clear within-modality effect. However, we also find that simply adding auxiliary modalities does not guarantee improvement on ICU mortality: only one bimodal configuration (notes with BioMedBERT) marginally exceeds the EHR-only baseline, and a unified three-modality model does not surpass it either. This points to a need for more flexible fusion architectures and clinically contextual evaluation that can reveal where each modality genuinely contributes.

2. Method

2.1. Problem Setup and Data

We study autoregressive modeling of patient trajectories from MIMIC-IV v3.1 (Johnson et al., 2024). Following ETHOS (Renc et al., 2024), we convert MIMIC-IV and MIMIC-IV-ED (Johnson et al., 2023a) into the Medical Event Data Standard (MEDS) format (Arnrich et al., 2024) and tokenize each patient's record into a chronologically ordered sequence $x_i = (x_{i,1}, \ldots, x_{i,T_i})$ over a vocabulary V of medical-event tokens, with event timestamps $\tau_{i,t}$.

In addition to the EHR stream, each patient p_i may have auxiliary observations from a modality set $\mathcal{M} = \{\text{ecg}, \text{cxr}, \text{note}\}$, with available modalities $\mathcal{M}_i \subseteq \mathcal{M}$ varying across patients. Specifically, patient p_i may have ECG recordings $E_i = \{(e_{i,r}, \tau^{\text{ecg}}_{i,r})\}_{r=1}^{R_i}$, where $e_{i,r} \in \mathbb{R}^{12 \times L_{i,r}}$ is a 12-lead waveform and $\tau^{\text{ecg}}_{i,r}$ is the acquisition time; chest radiographs $C_i = \{(c_{i,j}, \tau^{\text{cxr}}_{i,j})\}_{j=1}^{J_i}$; and clinical notes $N_i = \{(n_{i,k}, \tau^{\text{note}}_{i,k})\}_{k=1}^{K_i}$, where each note $n_{i,k}$ is tokenized into a sequence $u_{i,k} = (u_{i,k,1}, \ldots, u_{i,k,L_{i,k}})$ and truncated or padded to the encoder context length $L_{\max}$. All auxiliary data are sourced from MIMIC-IV-ECG (Gow et al., 2023), MIMIC-CXR-JPG (Johnson et al., 2019), and MIMIC-IV-Note (Johnson et al., 2023b) via PhysioNet (Goldberger et al., 2000). Missing modalities are represented via explicit masks rather than imputation. We use an 80/10/10 subject-level train/validation/test split.

2.2. Multimodal Architecture

Our Multimodal EHR trajectory model (see Fig. 2) reads an EHR and auxiliary modalities (here ECG time series, CXR images, and clinical notes text) token sequence that are embedded via a learned lookup table and processed by a GPT-2-style decoder (Radford et al., 2019) with latent dimension D. Each auxiliary modality is encoded by a frozen pretrained, modality-specific encoder. For ECG, each recording is partitioned into contiguous temporal patches and projected into patch embeddings. For CXR, images are split into 2D patches following the standard vision-transformer approach. For clinical notes, each document is tokenized and encoded by a BERT-style model, with truncation or padding to a fixed context length.

Latent compression. Each modality encoder produces a sequence of embeddings whose length depends on the input (e.g., a clinical note of N tokens yields an (N, d_{encoder}) tensor, where d_{encoder} is the encoder output dimension). Naive cross-attention from the EHR sequence of length T to such a modality sequence requires $\mathcal{O}(NT)$ operations, which quickly becomes expensive. We propose a latent compression module that compresses each modality sequence into k latents with $k \ll N$. The latent queries are learned parameters initialized randomly and iteratively refined through l layers of cross-attention to the modality sequence, reducing the effective cross-attention cost to $\mathcal{O}(Nk + Tk)$.

Gated cross-attention. The decoder conditions on compressed modality latents via cross-attention blocks interleaved with self-attention layers. Following Flamingo (Alayrac et al., 2022), each cross-attention block is modulated by a learnable scalar gate $g^l = \sigma(\beta^l)$ that controls the contribution of the modality update at layer l, stabilizing training and enabling depth-dependent modality usage.

			ICU Mortality		
Compression	k	l	AUROC	AUPRC	F1
None	–	–	0.8565	0.4357	0.7061
Mean Pooling	–	–	0.8642	0.4274	0.7037
Latent	8	1	0.8662	0.4400	0.7082
Latent	8	2	0.8709	0.4444	**0.7131**
Latent	8	4	0.8751	**0.4599**	0.7107
Latent	16	1	0.8640	0.4426	0.7121
Latent	16	2	**0.8771**	0.4486	0.7082
Latent	16	4	0.8580	0.4235	0.6957

Table 1. Analysis of different compression methods in a bimodal model trained with EHR and clinical notes. Clinical notes were encoded using BioMedBERT (Gu et al., 2020) as the note encoder. Performance is reported on the ICU Mortality task.

Temporal alignment masking. To ensure that predictions do not condition on future auxiliary observations, we enforce a past-only cross-attention rule: an EHR token at time τ_t may only attend to modality events with acquisition time strictly before τ_t (equal timestamps are also masked). Modality instances are retrieved by subject ID within the EHR window's time span, with at most K_m events per modality. This causal mask is combined with event-availability and padding masks to handle partial or absent modality availability.

Training objective. We train with standard next-token cross-entropy loss on the EHR token stream: $\mathcal{L}_{\text{CE}}(\theta) = -\sum_u \log p_\theta(x_{i,u+1} \mid x_{i,\leq u}, E_i, C_i, N_i)$, where auxiliary modalities serve only as conditioning inputs and the model generates EHR tokens only. The loss excludes padded and demographic positions.

2.3. Modality Encoders

We evaluate multiple pretrained encoders per modality to assess the impact of encoder choice on downstream performance, while keeping the EHR backbone and fusion mechanism fixed. For **ECG**, we compare CSFM (Gu et al., 2026), a multi-dataset cardiac foundation model, with ECG-FM (McKeen et al., 2025), an open electrocardiogram foundation model; both produce patch-level embeddings from 12-lead waveforms. For **CXR**, we compare BioMedCLIP, a vision–language model pretrained on biomedical image–text pairs, with a ViT-MAE pretrained via masked autoencoding on natural images. For **clinical notes**, we compare BERT (Devlin et al., 2019) with BioMedBERT (Gu et al., 2020), a domain-adapted variant pretrained on PubMed abstracts. All encoders are frozen during training; the latent compression module maps their outputs into k latent vectors of dimension D, matching the backbone hidden dimension.

Table 2. Effect of modality and encoder choice on ICU mortality. Bimodal rows pair EHR with one auxiliary modality; the final row reports a unified model conditioned on all three modalities using the best-performing encoder per modality (BioMedCLIP, CSFM, BioMedBERT). All multimodal variants use $k = 8$ and $l = 4$.

Modalities				Modality Encoder	ICU Mortality		
EHR	CXR	ECG	Notes		AUROC	AUPRC	F1
✓	×	×	×	-	0.8651	0.4561	0.7182
✓	✓	×	×	BioMedCLIP ViT (Zhang et al., 2023)	**0.8624**	**0.4471**	**0.7075**
✓	✓	×	×	ViT-MAE (He et al., 2022)	0.8505	0.4315	0.6755
✓	×	✓	×	CSFM (Gu et al., 2026)	**0.8703**	**0.4263**	**0.7024**
✓	×	✓	×	ECG-FM (McKeen et al., 2025)	0.8680	0.4072	0.7022
✓	×	×	✓	BERT (Devlin et al., 2019)	0.8579	0.4164	0.7066
✓	×	×	✓	BioMedBERT (Gu et al., 2020)	**0.8751**	**0.4599**	**0.7107**
✓	✓	✓	✓	Best per modality	0.8460	0.4330	0.6997

2.4. Zero-Shot Evaluation

We focus on zero-shot ICU mortality prediction. Following ETHOS (Renc et al., 2024), evaluation conditions on the last L tokens of patient history ending at the task start event (e.g., the ICU admission token for ICU mortality) and samples K=20 stochastic futures; the predicted probability is the fraction of rollouts containing the target token within the task horizon. For multimodal models, auxiliary encodings are computed once from the observed window and kept fixed during rollout. Further details of the rollout-based evaluation pipeline, including termination criteria and metric computation, are provided in Appendix A.

3. Results & Discussion

Table 1 reports a compression ablation in a bimodal EHR + clinical notes model on ICU mortality. Mean pooling improves AUROC over no compression but slightly lowers AUPRC and F1, whereas best latent compression configuration improves all three. Sweeping $k \in \{8, 16\}$ and $l \in \{1, 2, 4\}$, we select k=8, l=4, which gives the best AUPRC (0.4599) and competitive AUROC; we adopt it for all subsequent experiments. We note that no single setting dominates across all three metrics, so this choice prioritises AUPRC given the class imbalance of the mortality task.

Multimodal encoder ablation. Table 2 evaluates pretrained encoder choice in bimodal settings on ICU mortality, with the EHR-only baseline at 0.8651/0.4561/0.7182 (AUROC/AUPRC/F1). Adding an auxiliary modality does not uniformly improve over this baseline: CXR with either encoder performs below the baseline across all three metrics; ECG with CSFM (Gu et al., 2026) improves AUROC (0.8703 vs. 0.8651) but lowers AUPRC and F1; clinical notes with BioMedBERT (Gu et al., 2020) is the only configuration to exceed the baseline on both AUROC and AUPRC (0.8751/0.4599), while still trailing on F1. Encoder choice nonetheless has a clear within-modality effect: BioMedCLIP (Zhang et al., 2023) outperforms ViT-MAE (He et al., 2022) on CXR; CSFM outperforms ECG-FM (McKeen et al., 2025) on ECG with a notable AUPRC gap (0.426 vs. 0.407); and BioMedBERT substantially outperforms BERT (Devlin et al., 2019) on notes across all metrics. For CXR and notes, the advantage comes from domain-adapted pretraining over general-purpose models; for ECG, CSFM benefits from training on a larger and more diverse cardiac corpus. The unified model conditioned on all three modalities (Table 2, final row; BioMedCLIP for CXR, CSFM for ECG, and BioMedBERT for notes) also fails to surpass the EHR-only baseline, indicating that simply combining modalities does not yield additive gains.

Discussion. Our results show that these design choices matter substantially, yet they are often left uninvestigated. We note, however, that the rollout-based evaluation is stochastic and can be noisy, and due to computational constraints we were unable to repeat all experiments across multiple random seeds. These results should therefore be interpreted as evidence of broad trends rather than definitive rankings between closely performing configurations. Three takeaways emerge from these results. First, learned latent compression setting substantially outperforms both uncompressed cross-attention and mean pooling (Table 1), while reducing the attention cost from $\mathcal{O}(NT)$ to $\mathcal{O}(Nk + Tk)$. This is consistent with the Perceiver and Flamingo line of work (Jaegle et al., 2021; Alayrac et al., 2022), in which a small set of learned latents acts as a bottleneck that summarises a long modality sequence: here the bottleneck appears to distil the clinically informative content of each modality more effectively than either attending over the full sequence or naively averaging it, with the added benefit of bounded compute regardless of modality sequence length.

Second, encoder choice has a clear within-modality effect (Table 2): for every modality, the domain-adapted or more

broadly pretrained encoder outperforms the general-purpose alternative, holding the EHR backbone and fusion mechanism fixed. This indicates that, in our setting, the quality of the frozen modality representation is at least as consequential as the fusion architecture, echoing observations from medical foundation-model work that domain-specific pretraining materially improves transfer (Gu et al., 2020; Zhang et al., 2023). A practical implication of this finding is that effort spent selecting or adapting the modality encoder may yield larger returns than additional fusion-side complexity.

Third, and most importantly, adding modalities does not guarantee improvement on ICU mortality: only notes with BioMedBERT exceed the baseline on AUROC and AUPRC, and the unified three-modality model underperforms the EHR-only baseline. This is not unique to clinical data – a consistent finding across multimodal learning is that the best unimodal network can match or beat a jointly trained multimodal one, because joint optimisation tends to be *greedy*—the model comes to rely on whichever modality is fastest to learn from, leaving the others underfit (Wang et al., 2020; Wu et al., 2022), a phenomenon also framed as *modality competition* under joint training (Huang et al., 2022). In our case the EHR token stream is a strong, dense, and temporally complete signal for mortality, whereas the auxiliary modalities are sparse, intermittently available, and masked to past-only events; under a single next-token objective the backbone has little incentive to exploit them. This reframes our negative result: the deficit is plausibly an optimisation and signal-balance problem rather than evidence that the modalities lack information.

The CXR result deserves an extra comment, as it is a negative result beyond non-additivity; it falls below EHR-only baseline on all metrics (AUROC, AUPRC, F1) with both evaluated encoders. CXR appears to be the temporally sparsest modality in our training set. We attribute this primarily to temporal availability: radiographs are sparse relative to the dense EHR window, and after past-only masking many prediction points have no admissible CXR study to attend to, so the cross-attention pathway contributes mostly noise while still consuming model capacity. With our data we cannot fully separate this limited availability from low intrinsic mortality signal in CXR. ECG, in contrast, though also sparse, carries a more directly mortality-relevant physiological signal and at least improves AUROC.

4. Conclusion

We presented a framework for conditioning autoregressive EHR foundation models on ECG, CXR, and clinical notes via latent compression and gated cross-attention. Controlled ablations on MIMIC-IV show that latent compression outperforms simpler aggregation while also improving efficiency ($\mathcal{O}(Nk + Tk)$ attention cost), encoder choice has a clear within-modality effect, and simply adding modalities does not consistently improve ICU mortality over a strong EHR-only baseline. Future work includes more flexible fusion and modality-aware training, and cohort-stratified evaluation that can reveal where each modality contributes most.

Acknowledgements

YL is supported by UK Research and Innovation (UKRI) Centre for Doctoral Training in AI for Healthcare grant number (EP/S023283/1); JP is supported by the UKRI AI Centre for Doctoral Training in Digital Healthcare grant number (EP/Y030974/1), and by a philanthropic Bupa PhD Scholarship. AJB is supported by the UKRI AI Centre for Doctoral Training in Digital Healthcare grant number (EP/Y030974/1), and by The Lee Family Faculty of Medicine Scholarship. AAF is supported by the UKRI Turing AI Fellowship grant number (EP/V025449/1). AAF and MR acknowledge support of the UKRI AI programme and the Engineering and Physical Sciences Research Council (EPSRC) for the AI Hub in Generative Models (EP/Y028805/1). AAF & MR acknowledge the use of resources provided by the Isambard-AI National AI Research Resource (AIRR). Isambard-AI is operated by the University of Bristol and is funded by the UK Government's Department for Science, Innovation and Technology (DSIT) via UK Research and Innovation; and the Science and Technology Facilities Council (ST/AIRR/I-A-I/1023).

References

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., and Simonyan, K. Flamingo: a visual language model for few-shot learning. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=EbMuimAbPbs.

Amar, J., Liu, E., Breschi, A., Zhang, L., Kheradpour, P., Li, S., Lehmann, L. S., Giulianelli, A., Edwards, M., Jia, Y., et al. Integrating genomics into multimodal ehr foundation models. *arXiv preprint arXiv:2510.23639*, 2025.

Arnrich, B., Choi, E., Fries, J., McDermott, M., Oh, J., Pollard, T., Shah, N., Steinberg, E., Wornow, M., and van de Water, R. Medical event data standard (meds): Facilitating machine learning for health. In *Workshop on Time Series Learning for Health (TS4H) at ICLR*, 2024. URL https://iclr.cc/virtual/2024/23574.

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

Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., and Stanley, H. E. Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000. RRID:SCR_007345.

Gow, B., Pollard, T., Nathanson, L. A., Johnson, A., Moody, B., Fernandes, C., Greenbaum, N., Waks, J. W., Eslami, P., Carbonati, T., et al. Mimic-iv-ecg: Diagnostic electrocardiogram matched subset. *Type: dataset*, 6:13–14, 2023.

Gu, X., Shu, Y., Han, J., Liu, Y., Liu, Z., Anibal, J., Sangha, V., Phillips, E., Segal, B., Yuan, H., et al. Foundation models for biosignals: A survey. *Authorea Preprints*, 2025.

Gu, X., Tang, W., Han, J., Sangha, V., Liu, F., Gowda, S. N., Ribeiro, A. H., Schwab, P., Branson, K., Clifton, L., et al. Cardiac health assessment across scenarios and devices using a multimodal foundation model pretrained on data from 1.7 million individuals. *Nature Machine Intelligence*, 8(2):220–233, 2026.

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., and Poon, H. Domain-specific language model pretraining for biomedical natural language processing, 2020.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16000–16009, 2022.

Huang, Y., Lin, J., Zhou, C., Yang, H., and Huang, L. Modality competition: What makes joint training of multimodal network fail in deep learning? (provably). In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, pp. 9226–9259, 2022.

Jaegle, A., Gimeno, F., Brock, A., Vinyals, O., Zisserman, A., and Carreira, J. Perceiver: General perception with iterative attention. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4651–4664. PMLR, 2021. URL https://proceedings.mlr.press/v139/jaegle21a.html.

Johnson, A., Bulgarelli, L., Pollard, T., Celi, L. A., Mark, R., and Horng, S. Mimic-iv-ed (version 2.2), 2023a. URL https://doi.org/10.13026/5ntk-km72. RRID:SCR_007345.

Johnson, A., Pollard, T., Horng, S., Celi, L., and Mark, R. Mimic-iv-note: Deidentified free-text clinical notes. physionet, 2023b.

Johnson, A., Bulgarelli, L., Pollard, T., Gow, B., Moody, B., Horng, S., Celi, L. A., and Mark, R. Mimic-iv (version 3.1), 2024. URL https://doi.org/10.13026/kpb9-mt58. RRID:SCR_007345.

Johnson, A. E., Pollard, T. J., Greenbaum, N. R., Lungren, M. P., Deng, C.-y., Peng, Y., Lu, Z., Mark, R. G., Berkowitz, S. J., and Horng, S. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019.

Jorf, B. A. and Shamout, F. Medpatch: Confidence-guided multi-stage fusion for multimodal clinical data. *arXiv preprint arXiv:2508.09182*, 2025.

McKeen, K., Masood, S., Toma, A., Rubin, B., and Wang, B. Ecg-fm: An open electrocardiogram foundation model. *JAMIA open*, 8(5):ooaf122, 2025.

Moor, M., Huang, Q., Wu, S., Yasunaga, M., Dalmia, Y., Leskovec, J., Zakka, C., Reis, E. P., and Rajpurkar, P. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pp. 353–367. PMLR, 2023.

Pang, C., Jiang, X., Kalluri, K. S., Spotnitz, M., Chen, R., Perotte, A., and Natarajan, K. Cehr-bert: Incorporating temporal information from structured ehr data to improve prediction tasks. In *Machine Learning for Health*, pp. 239–260. PMLR, 2021.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. *OpenAI Technical Report*, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Renc, P., Jia, Y., Samir, A. E., Was, J., Li, Q., Bates, D. W., and Sitek, A. Zero-shot health trajectory prediction using transformers. *npj Digital Medicine*, 7(1):256, 2024. doi: 10.1038/s41746-024-01235-0. URL https://doi.org/10.1038/s41746-024-01235-0.

Spathis, D. and Kawsar, F. The first step is the hardest: Pitfalls of representing and tokenizing temporal data for large language models. *Journal of the American Medical Informatics Association*, 31(9):2151–2158, 2024.

Stahlschmidt, S. R., Ulfenborg, B., and Synnergren, J. Multimodal deep learning for biomedical data fusion: a review. *Briefings in bioinformatics*, 23(2):bbab569, 2022.

Tu, T., Azizi, S., Driess, D., Schaekermann, M., Amin, M., Chang, P.-C., Carroll, A., Lau, C., Tanno, R., Ktena, I., et al. Towards generalist biomedical ai. *Nejm Ai*, 1(3): AIoa2300138, 2024.

Wang, W., Tran, D., and Feiszli, M. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12695–12705, 2020.

Wang, Y., Yin, C., and Zhang, P. Multimodal risk prediction with physiological signals, medical images and clinical notes. *Heliyon*, 10(5), 2024.

Waxler, S., Blazek, P., White, D., Sneider, D., Chung, K., Nagarathnam, M., Williams, P., Voeller, H., Wong, K., Swanhorst, M., et al. Generative medical event models improve with scale. *arXiv preprint arXiv:2508.12104*, 2025.

Wu, N., Jastrzebski, S., Cho, K., and Geras, K. J. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162, pp. 24043–24055, 2022.

Zhang, S., Xu, Y., Usuyama, N., Bagber, H., Cliff, R., Crandall, D., Wong, C., Naumann, T., and Poon, H. Biomedclip: A multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023.

Component	Setting
Embedding dimension	768
Transformer layers	6 (unless stated otherwise)
Attention heads	12
Context length	2048
Dropout	0.1
Batch size	128
Optimizer	AdamW
Learning rate	3×10^{-4}
Weight decay	0.1
Validation interval	every 10K steps

Table 3. Model and training hyperparameters used in our experiments.

A. Model, Training, and Evaluation Framework

Model. We use a GPT-2 style architecture. The model consists of a learned token embedding layer, a learned positional embedding layer, a stack of transformer blocks with causal self-attention and MLP sublayers, a final layer normalization, and a linear language-modeling head. The model is trained with a standard next-token prediction objective over tokenized EHR trajectories. A "reduce learning rate on plateau" scheduler is used to multiply the learning rate by 0.5 when the validation loss plateaus for more than 3 validation intervals in a row. Table 3 summarizes the core model and optimization hyperparameters used in our experiments.

Evaluation framework. Evaluation is performed using a rollout-based benchmark framework as proposed in (Renc et al., 2024). For the ICU Mortality tasks we prompt the model with patient trajectories up to and including the presence of an `ICU_ADMISSION` token. The model then generates the next token auto-regressively until it predicts either the `MEDS_DEATH` or `ICU_DISCHARGE` tokens, or until a maximum generation length of 4096 tokens is reached. For each prompt, we run 20 stochastic rollouts. A scalar score is computed as the fraction of rollouts that terminate with the different end tokens, and this score is used as the model output for binary classification. Simulations that exceed the 4096-token limit are terminated and excluded from benchmark calculations. AUROC and AUPRC are then computed from the resulting per-prompt scores.