Autoregressive EHR Foundation Models with Multimodal Inputs

Anonymous Authors[1]

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. 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. Through controlled ablations on MIMIC-IV, we study two key design choices for multimodal EHR fusion: how to compress long modality sequences before cross-attention, and whether the choice of pretrained modality encoder matters for downstream performance. We show that latent compression substantially outperforms both uncompressed cross-attention and mean pooling, and that encoder choice has a clear within-modality effect, with stronger pretrained encoders consistently outperforming weaker alternatives. We further find that, under our current architecture, simply adding auxiliary modalities does not guarantee improvement on aggregate ICU mortality prediction over a strong EHR-only baseline, motivating future work on more flexible fusion architectures and clinically contextual evaluation.

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 Figure 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 processes 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.

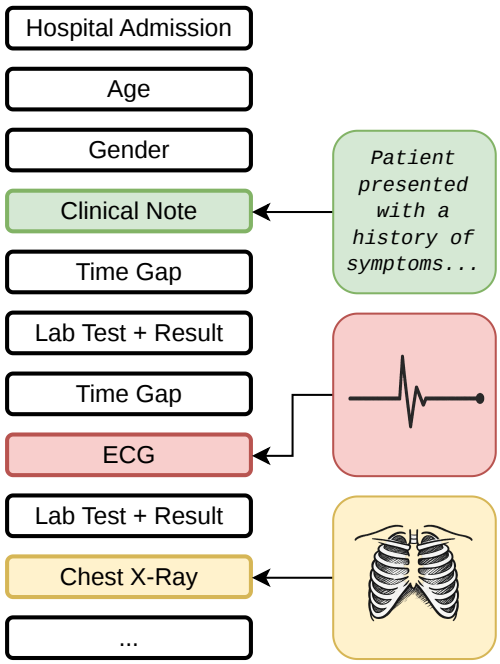

Figure 1. Electronic health records can be represented as sequences of clinical events spanning multiple modalities.

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

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

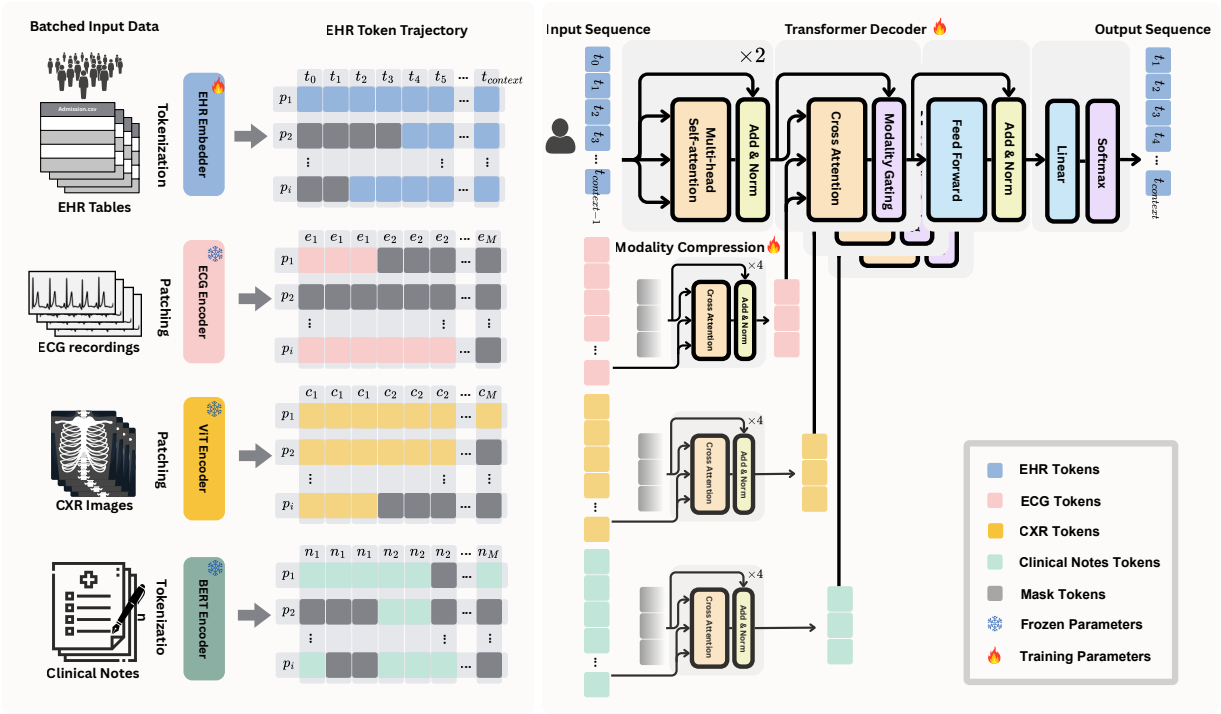

Figure 2. Multimodal EHR trajectory model. Auxiliary modalities (ECG, CXR, notes) are encoded, compressed into latent tokens, and fused into a GPT-2-style decoder via gated cross-attention. Temporal masking ensures each EHR token attends only to past modality events.

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 aggregate 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

Figure 2 summarizes the architecture. The EHR token sequence is embedded via a learned lookup table and processed by a GPT-2-style decoder (Radford et al., 2019) with hidden dimension D. Each auxiliary modality is encoded by a pretrained, modality-specific encoder (frozen or finetuned). 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 latents are initialized as random noise 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.

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: $\mathcal{L}_{\text{CE}}(\theta) = -\sum_u \log p_\theta(x_{i,u+1} \mid x_{i,\le u}, E_i, C_i, N_i)$, where auxiliary modalities serve only as conditioning inputs and the model generates EHR tokens only. The loss excludes padded and demographic positions.

Compression	k	l	ICU Mortality 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.

2.3. Modality Encoders

We experiment with 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.

2.4. Zero-Shot Evaluation

We evaluate on four clinical benchmarks: hospital mortality, ICU mortality, ICU admission, and ICU readmission. Following ETHOS (Renc et al., 2024), zero-shot 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.

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 encoder per modality.

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

3. Results & Discussion

Table 1 reports a compression ablation in a bimodal EHR + clinical notes model on ICU mortality. Mean pooling slightly improves over no compression, while latent compression yields the largest gains. Sweeping $k \in \{8, 16\}$ and $l \in \{1, 2, 4\}$, we find k=8, l=4 gives the strongest overall performance and we adopt it for subsequent experiments.

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 marginally improve AUROC and AUPRC (0.8751/0.4599) while still trailing the baseline on F1. Encoder choice nonetheless has a clear within-modality effect: BioMedCLIP outperforms ViT-MAE 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 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 dataset. The unified model conditioned on all three modalities also fails to surpass the EHR-only baseline, indicating that simply combining modalities does not yield additive gains.

Discussion. Three takeaways emerge from these results. First, learned latent compression substantially outperforms uncompressed attention and mean pooling (Table 1), suggesting that a bottleneck distills modality information more effectively than the alternatives. Second, encoder choice has a clear within-modality effect (Table 2), suggesting that the quality of the initial representation is at least as important as the fusion mechanism. Third, adding modalities does not guarantee improvement on aggregate metrics: only notes with BioMedBERT marginally exceed the baseline, and the unified three-modality model also fails to surpass it. This motivates two directions for future work. Architecturally, more flexible fusion and modality-aware training (e.g., per-modality loss weighting or curriculum over availability) may be needed for additive gains. Methodologically, aggregate metrics likely obscure modality-specific value, for example ECG is most informative for arrhythmia or ischemia cohorts, CXR for pulmonary cohorts, motivating clinically contextual evaluation that can reveal where each modality contributes.

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 aggregate 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.

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.

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., Lungren, M., Peng, Y., Lu, Z., Mark, R., Berkowitz, S., and Horng, S. Mimic-cxr-jpg-chest radiographs with structured labels. *PhysioNet*, 101:215–220, 2019.

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.

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, 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.

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.