MULTIMODAL IRREGULAR TIME-SERIES ANALYSIS ON CLINICAL DATABASES

Anonymous authors
Paper under double-blind review

ABSTRACT

Clinical datasets are very useful to drive patient outcome prediction. State-of-the-art approaches typically aggregate patient-wise time-series from such databases and build analysis models. However clinical time-series analysis poses unique challenges due to their irregular inputs under interleaved multi-modalites. Current methods are either unimodal, or fail to capture cross-modal interactions across the interleaved modalities. To bridge this gap, we propose to achieve seamless cross-modal interaction, that is, allowing bidirectional information flow across interleaved multimodal inputs. Our model, MIST (Multimodal Irregular Seamless Transformer), serves as a novel clinical time-series model that supports seamless cross-modal interaction under a unified token-stream representation. We evaluate MIST across four tasks on two real-world databases, showing that MIST outperforms state-of-the-art baselines, with 8.8% and 5.4% relative AUROC gains on adverse-outcome prediction, a 4.8% AUROC gain on MIMIC phenotyping, 12.3% and 18.5% MSE reductions on length-of-stay, and 0.7% and 0.8% MSE reductions on vital-sign forecasting averaged over all horizons, rising above 10% at short horizons. We further demonstrate the multimodal gains through component ablations and case studies. [1]

1 INTRODUCTION

Clinical databases of electronic health records (EHRs) underpin a wide range of patient-outcome prediction tasks, including vital-sign forecasting (Hyland et al., 2020; Zhang et al., 2023), length-of-stay regression (Harutyunyan et al., 2019; Rocheteau et al., 2021), phenotype classification (Lipton et al., 2015; Harutyunyan et al., 2019), and adverse-outcome prediction (Thorsen-Meyer et al., 2020; Choi et al., 2016). Recent state-of-the-art approaches to these tasks usually organize each patient's records as a time series and apply time-series models (Zhang et al., 2023; Yalavarthi et al., 2024; Zhang et al., 2022b; 2024). Developing accurate clinical time-series models is therefore central to fully utilize these databases.

However, the analysis of clinical time series poses unique challenges. First, clinical records are often logged sporadically (Lipton et al., 2016; Che et al., 2018), introducing irregularity and sparsity. Second, alongside numerical records of vital signs and lab measurements, clinical databases also contain textual records such as patient activities (Van Aken et al., 2021); modeling such data requires capturing correlations both within and between modalities. Third, medical time series typically carry implicit medical context that requires additional effort to interpret (Alsentzer et al., 2019; Lee et al., 2020). These challenges become more pronounced on specialized cohorts such as pediatric databases (Zeng et al., 2020; Rivera et al., 2022; Digitale et al., 2025), where irregularity is more extensive, textual records are more frequently maintained, and the medical context behind patients are more specialized.

While many clinical time-series models have been proposed, each addresses only part of the challenge. Most prior work optimizes for irregularity (Tipirneni & Reddy, 2022; Shukla & Marlin, 2021; Zhang et al., 2024; Yalavarthi et al., 2024; Zhang et al., 2022b) within a single modality, leaving informative textual data unused. Few approaches tried to handle heterogeneous modalities, but mostly under aligned, regular sampling assumptions (Khadanga et al., 2019; Zhang et al., 2022a; Soenksen

[1] Code available at `https://anonymous.4open.science/r/MIST-8104`

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.

et al., 2022). Only a few recent methods (Zhang et al., 2023; Dai et al., 2025) attempt both axes simultaneously, yet they process text and numerical streams in parallel pipelines, missing the complex cross-modal interactions that occur when notes and vitals interleave in time.

To address these limitations, we argue that the key is to achieve *seamless* cross-modal interaction: numerical measurements and textual notes should be processable as arbitrarily-interleaved events in a single stream, with information flowing in both directions across the entire temporal range. We propose Multimodal Irregular Seamless Transformer (MIST), a novel clinical time-series model that comprises three components: (1) a unified token-stream representation that is scalable on irregular multimodal data, and (2) a bidirectional gated cross-modal attention that enables seamless interactions. MIST further integrates unimodal time-series encoders to strengthen vital-sign representations and boost downstream performance.

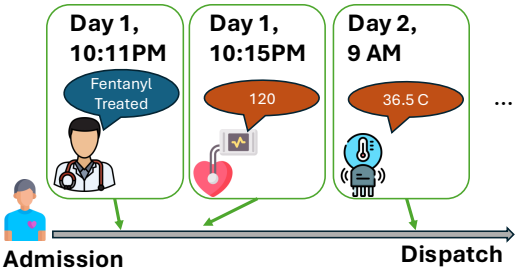

Figure 1: Example of clinical databases. Patient records are collected in either modality irregularly.

We validate our approach on four complementary patient-outcome prediction tasks on MIMIC-III (Johnson et al., 2016; Harutyunyan et al., 2019) as a general-purpose database. We also used a Children's Hospital of Atlanta (CHoA) cohort which is a pediatric-focused database. MIST significantly outperforms state-of-the-art clinical and general-purpose time-series models, relatively improving adverse-outcome AUROC by 8.8% and 5.4% on MIMIC and CHoA respectively, reducing length-of-stay MSE by 12.3% and 18.5%, lifting phenotype AUROC by 4.8%, and achieving 0.7% and 0.8% vital forecasting MSE reduction averaged over all horizon while over 10% on near-horizon. We further provide ablation studies on model architecture and stratified analyses across forecast horizons and vital channels to localize where multimodal gains originate.

Our contributions can be summarized as:

- We propose MIST, a multimodal clinical time-series model that unifies irregular numerical and textual data under a single token-stream representation and supports *seamless* cross-modal interaction, tackling the unaddressed challenge of multimodal irregular EHR modeling.
- We conduct a comprehensive evaluation across four patient-outcome prediction tasks (vital-sign forecasting, mortality classification, length-of-stay regression, and phenotyping) on two clinical cohorts (MIMIC-III and CHoA), emphasizing superior performance against state-of-the-art baselines.
- We provide ablation studies that localize where multimodal lift originates, showing that text-derived gains concentrate in specific time horizons and vital channels.

2 Related Work

2.1 Clinical and General-purpose Time-Series Models

A rich body of work has built on these benchmarks to model patient-level outcomes from clinical time-series, featured by the benchmarking efforts on MIMIC (Harutyunyan et al., 2019). State-of-the-art approaches model patient-wise time-series with a particular focus on irregularity and sparsity, via recurrent architectures (Che et al., 2018), attention-based mechanisms (Horn et al., 2020; Shukla & Marlin, 2021; Tipirneni & Reddy, 2022), or graph neural networks (Zhang et al., 2022b; Yalavarthi et al., 2024; Zhang et al., 2024). A separate line of work has explored the potential of heterogeneous modalities, but largely on regularly sampled inputs such as chest imaging (Hayat et al., 2022; Soenksen et al., 2022), bucketed EHRs (Zhang et al., 2022a), or free-text notes (Van Aken et al., 2021). Only a small number of recent works (Zhang et al., 2023; Dai et al., 2025) reach the intersection of multimodal and irregular inputs, and even those route each modality through a separate pipeline followed by a late fusion, underestimating the cross-modal interactions present in the interleaved data stream.

Beyond the clinical setting, time-series modeling has seen substantial development in domains such as electricity demand (Zhou et al., 2021), epidemiology (Rodriguez et al., 2024), and finance (Li et al., 2024). Classical time-series models target evenly sampled, unimodal inputs through end-to-end pretraining (Nie et al., 2023; Liu et al., 2024c; Zeng et al., 2023) or large-scale pretraining (Ansari et al., 2024; Kamarthi, 2024; Liu et al., 2025) and cannot ingest an irregular stream of both numeric and textual events. More recent work has begun to study multimodal extensions (Liu et al., 2024b; Chang et al., 2025; Zhang et al., 2026). However, these pipelines carry similar concerns as clinical models that they rely on late fusion and do not model interactions between interleaved modalities.

2.2 Multimodal Machine Learning

Beyond time-series, multimodal learning has been extensively studied in adjacent domains such as vision (Radford et al., 2021; Alayrac et al., 2022; Liu et al., 2024a) and audio (Akbari et al., 2021; Girdhar et al., 2023), evolving from early contrastive-learning pipelines (Radford et al., 2021; Li et al., 2022) to later cross-attention designs (Alayrac et al., 2022; Liu et al., 2023). Direct application to clinical time-series, however, is nontrivial: the available training-data scale is much smaller, established cross-modality alignments (image–text, audio–text) do not exist for vital–event data, and most of these methods assume dense, temporally aligned input streams that do not hold in irregular clinical records.

3 Problem Definition

We consider a clinical time-series dataset $\mathcal{D} = \{X_j\}_{j\in\{1,\dots,N\}}$ in which the record of patient j consists of two kinds of entries collected over the admission. Numeric records $X_j^{\text{num}} = \{(v_i, e_i, t_i)\}$ pair a scalar value $v_i \in \mathbb{R}$ with a measured variable $e_i \in \mathcal{E}_{\text{num}}$ (a vital sign, a laboratory test) at timestamp t_i. Textual records $X_j^{\text{text}} = \{(s_i, \tau_i, t_i)\}$ pair a short context phrase s_i (e.g., "Atrial Flutter", "escalating") with an event type $\tau_i \in \mathcal{E}_{\text{text}}$ given by a natural-language name (e.g., "Heart rhythm", "Neosynephrine drip") at timestamp t_i. Textual records in this work are such short, timestamped chart entries, such as medication orders, care observations, and diagnosis tags, rather than long-form clinical notes. The full record is $X_j = X_j^{\text{num}} \cup X_j^{\text{text}}$, and we make no assumption on how records of either kind are distributed along the temporal axis.

Let y_j denote the patient-outcome prediction target for patient j. Our goal is to learn a clinical time-series model $f_\theta \colon X_j \mapsto \hat{y}_j$ parameterized by θ that minimizes

$$\min_\theta \ \mathbb{E}_{X_j}\big[\mathcal{L}\big(y_j, f_\theta(X_j)\big)\big], \tag{1}$$

where y_j can be a regression target such as length-of-stay or a classification target such as in-hospital mortality, and $\mathcal{L}$ is correspondingly aggregated squared error or cross-entropy.

4 Method

Section 3 describes a clinical record as two kinds of entries, numeric and textual. Existing multimodal clinical models consume them through a dual-stream design, maintaining two seperate streams followed by a late fusion module. We instead argue in this section that this view discards the property of clinical records that carries most of their multimodal signal, and we build our model from that argument. Section 4.1 makes the observation, states what it requires of a model, and derives a unified token stream that meets the requirement. Section 4.2 gives the attention mechanism that operates on the stream; Section 4.3 assembles the full model.

4.1 Irregular, Interleaved Inputs and a Unified Token Stream

Although stored as two record types, the numeric and textual entries of a patient share one timeline. A numeric observation may be followed by another numeric observation or by a textual event, at an arbitrary gap, and the modality alternates irregularly throughout the stay. We argue that this *interleaving* is important to capture in clincal databases. For example, "10:15 pm; heart rate; 120" followed by "10:16 pm; patient activity; chest pain" suggests a cardiac event that neither record indicates alone, while the same two records in the opposite order or hours apart would not.

We therefore argue that the key to clinical data is what we call *seamless* interaction: every token sits in a single pipeline in which it can attend to any other token, within or across modality boundaries and with relative timing preserved, so that information flows in both directions along the whole temporal range.

Achieving such a seamless advantage under clinical databases is non-trivial and not covered by existing efforts. Irregular time-series encoders (Horn et al., 2020; Shukla & Marlin, 2021; Yalavarthi et al., 2024) represent each record as a triplet (v_i, e_i, t_i) with a scalar value and an integer-indexed variable as type, which cannot fits textual records since a high-dimensional contextual embedding barely share a space with the projection of a scalar. Interleaved vision–language models (Lu et al., 2019; Alayrac et al., 2022), conversely, assume that each unit of the stream is self-contained and carries enough meaning on its own, so that a secondary modality can be attached easily. Yet, a single clinical record is close to uninformative in isolation, while meaning emerges only from many readings over time.

To bridge this gap, we construct a unified stream in which each modality is moved part-way toward the other, so that all three channels of a token are shared. For each token, we consider three components:

- *Time*, where we encode identically for both modalities by Time2Vec (Kazemi et al., 2019).
- *Type*, where we assign every type from either modality with a context value by passing its natural-language name through ClinicalBERT (Alsentzer et al., 2019) and linearly projecting the result.
- *Value*, we maintain the original numeric values. For textual records, we embed the context phrase s_i with ClinicalBERT and take its projection onto the first principal component of all training-set context embedding.

After this pipeline, records of either modality are now under triplets (v_i, e_i, t_i) of one form, placed in a single sequence ordered by time. This representation gives a textual record a presence while we retain the full context embedding as a side sequence c_i. Since every textual record now has a token at its own position, that token can serve as the query that helps pull its full content c_i back into the stream, which leads to an ideal seamless cross-modality interaction. We further give the follow-up mechanism in Section 4.2.

4.2 Establishing Seamless Attention on Unified Token Stream

Based on the previous designs, we are able to introduce the seamless cross-modality interactions based on a cross-attention-based procedure inspired by vision–language model architectures (Lu et al., 2019; Alayrac et al., 2022; Moor et al., 2023). Since attention-mechanism effectively captures the influence between sequential data flow, we expect self-attention and cross-modality attention to capture interactions inside/between modalities across the interleaved inputs, correspondingly, while a learnable gate with skip lines progressively integrates cross-modal context into the numeric stream. Eventually, the numeric stream would be fused with textual context.

Specifically, after the encoding stage, we obtain a pair of aligned embedding sequences $\{(z_i, c_i, m_i)\}_{i=1}^{S}$, where z_i is initially the encoded triplet token under aforementioned token stream, c_i is the text-side embedding (zero-padded if the token is from numeric records), and $m_i \in \{0, 1\}$ are binary mask indicating whether the token is oriented from numeric ($m_i = 0$) or textual records at each position. Each layer of the fusion module consists of two blocks applied to the numeric stream:

1. **Gated cross-modality attention.** The stream attends over the text embeddings, with $\{z_i\}$ as queries and $\{c_i\}$ as keys and values; positions with $m_i = 0$ are excluded from the key set, so only textual records contribute. The output is scaled by a learnable gate and added back to the stream as a residual. This step is deliberately one-directional, injecting the semantic content of each note into the unified stream.
2. **Self-attention**: the numeric/triplet embeddings $\{z_i\}$ are processed by multi-head self-attention (MHA) with no modality mask, so every token, numeric or textual, can attend to every other token regardless of modality or position in time.

Let $Z^{(k)}$ denote the numeric embedding sequence entering the k-th layer. The procedure is then formalized as

$$Z' = Z^{(k)} + \tanh(\alpha) \cdot \mathrm{MHA}\Big(Q{=}Z^{(k)},\ K{=}C, V{=}C\Big)\,, \tag{2}$$

$$Z^{(k+1)} = Z' + \mathrm{MHA}(Z') \tag{3}$$

where $\mathrm{MHA}(Q, K{=}V)$ denotes multi-head attention with the indicated query and key/value tensors, $C = \{c_i\}_{i=1}^{S}$, and $\mathrm{MHA}(Z^{(k)})$ notes the self-attention on $Z^{(k)}$. The division of labour is what we mean by seamless: cross-attention grounds the stream in textual semantics, while self-attention lets information flow in both directions between textual tokens and numeric tokens, and across time gaps. Stacking the two blocks N times lets a numeric token influence how a later layer reads a textual token and vice versa, which a single late-fusion step cannot do.

4.3 Multimodal Irregular Seamless Transformer

Combining the components above, we instantiate **MIST** (Multimodal Irregular Seamless Transformer), a clinical time-series model designed to address multimodal and irregular clinical contexts. An overview is shown in Figure 2.

After the cross-modality attention stack of Section 4.2, MIST produces an embedding sequence that fuses textual and numeric event information. We merge this representation with the output of a backbone model that specializes in vital signs. The intuition is that vital signs are densely sampled and carry well-structured physiological context; existing clinical time-series methods have been extensively optimized to encode them, and we wish to reuse rather than replace this capability.

We adopt an attention-based fusion module to combine the two streams, estimating the interactions between multimodal context and vital context through attention mechanism. Let $f(x) \in \mathbb{R}^{T\times d}$ denote the output of the backbone time-series model and let $Z \in \mathbb{R}^{S\times d_Z}$ denote the fused embedding sequence produced by the cross-modality module of Section 4.2. The fusion procedure is

$$\begin{aligned} \tilde{z}_{pool} &= \mathbf{1}_T \otimes \mathrm{MHA}(q,\, W_e Z,\, W_e Z) \in \mathbb{R}^{T\times d}, \\ g &= \sigma\big(W_g\big[\, f(x) \,\|\, \tilde{z}_{pool} \,\big]\big) \in \mathbb{R}^{T\times d}, \\ \hat{y} &= f(x) + \alpha \cdot \left(g \odot f(x) + (1-g) \odot \tilde{z}_{pool}\right), \end{aligned}$$

where q, W_e, W_g, and α are trainable parameters, σ is the element-wise sigmoid, $\odot$ is the Hadamard (element-wise) product, $\|$ denotes concatenation along the channel dimension, and $\mathbf{1}_T\otimes$ broadcasts the pooled vector across the T time steps of the backbone output. The output $\hat{y}$ is then passed to a task-specific readout module for each downstream task. The computational complexity is approximately $O(S^2)$ in the sequence length S; we show in the appendix that this remains affordable on clinical cohorts.

5 Experiments

We evaluate MIST on four patient-outcome prediction tasks against state-of-the-art clinical and general-purpose time-series models drawn from a range of forecasting and classification settings. We further provide ablation studies that localize the performance gain to specific architectural choices.

5.1 Settings

Datasets We use two clinical datasets, CHoA and MIMIC-III. MIMIC-III (Johnson et al., 2016) is a public adult-ICU dataset maintained by the MIT Laboratory for Computational Physiology. CHoA is an access-restricted pediatric database maintained by Children's Healthcare of Atlanta. Cohort-construction details as well as summary statistics are provided in the Appendix A.

Tasks We consider four downstream tasks that are generally recognized as a complementary set of patient-outcome prediction problems (Harutyunyan et al., 2019). Implementation details are shown in the Appendix B.

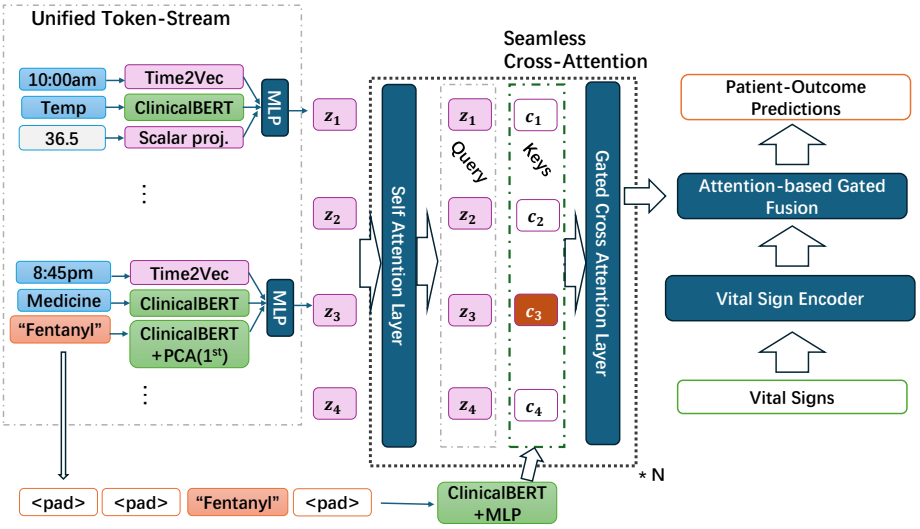

Figure 2: An overview of the proposed Multimodal Irregular Seamless Transformer (MIST) method. Given inputs from clinical time-series, MIST first encodes numeric and textual records via an irregularity-aware unified token-stream backed with clinical contexts. Subsequently, a seamless cross-attention module integrates textual semantics with the numerical representations. Finally, MIST fuses the extracted context with the outputs of a vital-sign encoder to produce the final prediction.

- **Adverse-outcome Prediction.** On MIMIC-III we take the ICU expire flag as the positive label, following (Harutyunyan et al., 2019). On CHoA we take "Transfer to ICU" and "Deceased" as positive labels. Input records are truncated to 24h, following (Tipirneni & Reddy, 2022).
- **Phenotype Classification.** We predict 25-class ICD-9 phenotypes on MIMIC, following (Harutyunyan et al., 2019; Tipirneni & Reddy, 2022; Zhang et al., 2022b), with a 48h input window. CHoA is omitted from this task because corresponding phenotype labels are not available.
- **Length-of-Stay Regression.** Following (Rajkomar et al., 2018; Khadanga et al., 2019; Wang et al., 2020), we predict length-of-stay on the log scale, $\log(1 + \text{LOS}_{\text{hours}})$ (capped at 30 days), from the first 24h of input.
- **Vital-sign Forecasting.** Following (Harutyunyan et al., 2019; Zhang et al., 2024; Yalavarthi et al., 2024; Zhang et al., 2023), we use the first 24h of patient records as input and predict the next 24h of vital signs, bucketed into 96 timestamps.

Baselines We compare against three families of time-series models: (i) general-purpose time-series models, including DLinear (Zeng et al., 2023), PatchTST (Nie et al., 2023), iTransformer (Liu et al., 2024c), and the recent multimodal irregular model IMM-TSF (Chang et al., 2025); (ii) state-of-the art clinical time-series models designed for forecasting, including t-PatchGNN (Zhang et al., 2024), GraFITi (Yalavarthi et al., 2024), and UTDE (Zhang et al., 2023); and (iii) state-of-the-art clinical time-series models designed for classification, including GRU-D (Che et al., 2018), mTAND (Shukla & Marlin, 2021), and RAINDROP (Zhang et al., 2022b). Implementation details are provided in the Appendix.

Metrics For forecasting and length-of-stay prediction, we report mean squared error (MSE) and mean absolute error (MAE). For classification tasks (Phenotyping, Adverse-outcome), we report AUROC and AUPRC. Specifically, for phenotyping we use macro-averaged scores across the 25 classes. All reported metrics are averaged across three random seeds.

Vital Encoder We use GraFITi as the vital encoder across all tasks. An ablation over alternative encoders is provided in Appendix C.

Table 1: Phenotype and Adverse-outcome classification tasks. MIMIC-III-IHM: in-hospital mortality. MIMIC-III-PHE: 25-class ICD-9 phenotyping (macro-averaged). CHoA-ESC: ICU escalation. Higher is better (↑). Best in **bold**, second-best underlined.

Method	MIMIC-III-IHM		MIMIC-III-PHE		CHoA-ESC	
	AUROC ↑	AUPRC ↑	AUROC ↑	AUPRC ↑	AUROC ↑	AUPRC ↑
General Time-Series Models						
DLinear	.662 ± .010	.294 ± .007	.621 ± .013	.290 ± .010	.721 ± .007	.054 ± .006
PatchTST	.612 ± .013	.221 ± .010	.605 ± .005	.280 ± .004	.660 ± .003	.037 ± .005
iTransformer	.659 ± .001	.264 ± .006	.615 ± .011	.286 ± .008	.727 ± .005	.051 ± .003
IMM-TSF	.704 ± .002	.297 ± .007	.667 ± .004	.320 ± .006	.732 ± .009	.054 ± .007
Clinical Forecasting Models (classification-adapted)						
t-PatchGNN	.699 ± .011	.291 ± .009	.676 ± .002	.332 ± .005	.731 ± .004	.061 ± .005
GraFITi	.701 ± .003	.289 ± .011	.692 ± .001	.349 ± .003	.739 ± .032	.053 ± .008
UTDE	.714 ± .003	.301 ± .004	.689 ± .001	.341 ± .002	.739 ± .001	.055 ± .006
Clinical Classification Models						
GRU-D	.725 ± .004	.313 ± .009	.692 ± .003	.345 ± .005	.748 ± .003	.068 ± .004
mTAND	.710 ± .002	.303 ± .005	.673 ± .001	.322 ± .000	.746 ± .011	.056 ± .002
RAINDROP	.646 ± .037	.227 ± .023	.639 ± .002	.295 ± .002	.764 ± .004	.064 ± .006
MIST (Ours)	**.789** ± .011	**.398** ± .020	**.725** ± .004	**.380** ± .002	**.805** ± .035	**.073** ± .021

Table 2: Length-of-stay regression on MIMIC-III and CHoA (predicted log-LOS). Lower is better (↓). Best in **bold**, second-best underlined.

Method	MIMIC-III		CHoA	
	MSE ↓	MAE ↓	MSE ↓	MAE ↓
General Time-Series Models				
DLinear	.517 ± .005	.592 ± .003	.724 ± .001	.584 ± .001
PatchTST	.612 ± .006	.621 ± .006	.738 ± .003	.595 ± .001
iTransformer	.504 ± .003	.580 ± .003	.723 ± .004	.586 ± .005
IMM-TSF	.496 ± .001	.581 ± .003	.707 ± .002	.580 ± .002
Clinical Forecasting Models				
t-PatchGNN	.493 ± .005	.575 ± .005	.710 ± .001	.582 ± .002
GraFITi	.470 ± .004	.556 ± .003	.701 ± .004	.578 ± .000
UTDE	.481 ± .002	.558 ± .001	.696 ± .002	.576 ± .001
Clinical Classification Models				
GRU-D	.474 ± .003	.562 ± .002	.692 ± .002	.577 ± .000
mTAND	.482 ± .001	.562 ± .001	.702 ± .001	.579 ± .000
RAINDROP	.610 ± .007	.603 ± .002	.695 ± .003	.578 ± .005
MIST (Ours)	**.412** ± .003	**.504** ± .001	**.564** ± .008	**.526** ± .006

5.2 RESULTS

The performance of MIST and the baseline models is reported in Table 1 (adverse-outcome and phenotyping classification), Table 2 (length-of-stay regression), and Table 3 (vital-sign forecasting). MIST outperforms every baseline on every task. Relative to the strongest baseline for each task, MIST improves adverse-outcome AUROC from 0.725 to 0.789 ($\approx 8.8\%$ relative), improves phenotyping AUROC from 0.692 to 0.725 ($\approx 4.8\%$ relative), and reduces length-of-stay MSE from 0.470 to 0.412 ($\approx 12.3\%$ relative). Similar gains hold on CHoA, with relative improvements of 5.4% in adverse-outcome AUROC and 18.5% in length-of-stay MSE. These gaps clearly illustrate the contribution

Table 3: 24h-ahead vital sign forecasting on MIMIC-III and CHoA. Lower is better (↓). Best in **bold**, second-best underlined.

Method	MIMIC-III		CHoA	
	MSE ↓	MAE ↓	MSE ↓	MAE ↓
General Time-Series Models				
DLinear (Zeng et al., 2023)	.623 ± .002	.550 ± .001	.530 ± .000	.501 ± .001
PatchTST (Nie et al., 2023)	.613 ± .000	.545 ± .001	.510 ± .000	.487 ± .000
iTransformer (Liu et al., 2024c)	.609 ± .001	.541 ± .000	.509 ± .000	.488 ± .000
IMM-TSF (Chang et al., 2025)	.599 ± .001	**.526** ± .000	.483 ± .001	.475 ± .001
Clinical Forecasting Models				
t-PatchGNN (Zhang et al., 2024)	.597 ± .002	.537 ± .003	.494 ± .001	.480 ± .001
GraFITi (Yalavarthi et al., 2024)	<u>.590</u> ± .001	.533 ± .002	<u>.482</u> ± .000	<u>.472</u> ± .001
UTDE (Zhang et al., 2023)	.596 ± .001	.540 ± .002	.482 ± .000	.474 ± .000
Clinical Classification Models (forecasting-adapted)				
GRU-D (Che et al., 2018)	.591 ± .001	.535 ± .001	.482 ± .000	.472 ± .000
mTAND (Shukla & Marlin, 2021)	.592 ± .000	.536 ± .001	.484 ± .000	.475 ± .000
RAINDROP (Zhang et al., 2022b)	.618 ± .005	.551 ± .004	.492 ± .001	.481 ± .001
MIST (Ours)	**.586** ± .001	<u>.530</u> ± .000	**.478** ± .000	**.471** ± .000

Table 4: Architecture ablations on MIMIC. Each row reports the change in test metric from the full MIST model under one ablation: (1) replacing the seamless cross-attention fusion with naive concatenation, (2) zero-masking all type-name encodings, and (3) zero-masking text-context values.

Model Variant	Forecast MSE ↓	Mortality AUROC ↑	LOS MSE ↓	Phenotype AUROC ↑
MIST	.586	.789	.412	.725
- seam. cross-attn	+.004	-.092	+.063	-.032
- type-enc	+.001	-.011	+.008	-.010
- text-value	.000	-.005	+.003	-.004

of MIST: prior clinical time-series baselines either cannot consume the categorical event stream at all, or apply late fusion without modeling how events interleave with the numerical signals. MIST addresses both limitations and substantially improves patient-outcome prediction.

The forecasting gain in Table 3 is smaller, which is expected: unlike patient-level outcome targets, vital-sign forecasting requires a dense per-step prediction sequence and is intrinsically harder to improve in aggregate. To examine where the gain comes from, we decompose the MIMIC forecasting improvement over the GraFITi baseline by vital channel and by forecast horizon in Figure 3. The improvement is heavily concentrated at short horizons, and is markedly larger on "Temp" and "DBP". This aligns with the observation that the majority of MIMIC text events are temperature- or hemodynamics-related, and such events tend to exert short-term influence on patient state.

5.3 Ablation Studies

We perform three architecture ablations: (1) removing the cross-attention and adapt a dual-stream pipeline followed by late-fusion mechanism (detailed in Appendix B.6), to isolate the contribution of the cross-attention mechanism; (2) replacing the type-name encodings with all-zeros, to isolate the contribution of introducing meta medical contexts; (3) replacing the textual context values with zeros, to isolate the contribution of text-conditioned context in the unified token stream. As shown in Table 4, all three ablations degrade performance, with the cross-attention ablation producing by far the largest drop on every task, confirming that seamless cross-attention is the primary architectural contributor to MIST's gain. The text-value ablation degrades performance less, which is expected since a lot of the type names already carries most of the clinical meanings.

Table 5: **Case study on synthetic patients: how cross-modal records support the multimodal lift.** "h +n" denotes hours since admission. (a) GraFITi misses an unconscious patient whose numeric signs appear stable, whereas MIST draws on the textual records and correctly flips the prediction. (b) GraFITi raises a false alarm for a patient with large numeric fluctuations, whereas MIST recognizes from the text that these fluctuations stem from an emergent but successful intervention, after which the patient reawakens, and correctly predicts low risk.

Case A. Outcome: **Deceased**
GraFITi $p = 0.23$; **MIST** $p = 0.63$.

Numeric Records		
[vit]	HR	$75 \rightarrow 96$
[vit]	RR	$22 \rightarrow 20$
[vit]	SBP	$109 \rightarrow 150s \rightarrow 90$
[vit]	O_2 sat	$96 \rightarrow 100$
[vit]	Temp	$34.8 \rightarrow 35.6$
[lab]	h +16	pH $7.50 \rightarrow 7.25$
Textual Records		
[care]	h +3	Riker-SAS: Unarousable
[care]	h +3	GCS E1 M1 V1 $= 3/15$
[med]	h +2	Neosynephrine drip, escalating
[care]	h +16	Intubated (peri-arrest)

(a) Part of records from synthetic patient A who ended up with Deceased in ICU.

Case B. Outcome: **Survived**
GraFITi $p = 0.65$; **MIST** $p = 0.42$.

Numeric Records		
[vit]	HR	$40 \rightarrow 26(h\ +8) \rightarrow 80$
[vit]	SBP	$74 \rightarrow 160$
[vit]	O_2 sat	$47 \rightarrow 100$
Textual Records		
[care]	h +0	Level of consciousness: Alert
[care]	h +1	Heart rhythm: Atrial Flutter
[care]	h +8	Heart rhythm: Complete Heart Block
[care]	h +8	Heart rhythm: Ventricular Tachycardia
[care]	h +8	Heart rhythm: V Paced
[care]	h +13	Arouse to Stimulation
[care]	h +16	Re-awakens

(b) Part of records from synthetic patient B who ended up with Survived in ICU.

5.4 Case Studies

To investigate how MIST benefits downstream patient-outcome prediction, we conduct case studies on in-hospital mortality using synthetic patients generated from MIMIC-III records (details in Appendix B.5). As shown in Tables 5a and 5b, MIST corrects the prediction in cases where the unimodal GraFITi baseline fails. For Patient A, the numeric record alone shows a stable pattern, yet the accompanying clinical notes reveal substantial risk: the patient remains unarousable and is receiving an unusual Neo-Synephrine treatment, indicating that the seemingly stable numeric signs in fact mask a high-risk condition. For Patient B, the vital signs alone appear concerning, with large fluctuations around the 8th hour; however, the notes attribute these changes to a heart block that is successfully resolved, as confirmed by the patient's subsequent reawakening, and MIST lowers its risk prediction accordingly.

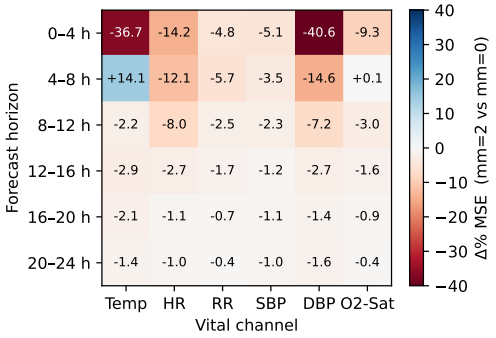

Figure 3: Forecasting gain of MIST on MIMIC, decomposed across vital channels and forecast horizons. Although the aggregate gap is small, MIST yields a substantial improvement at short horizons.

6 Conclusion

In this paper, we have discussed the difficulties in clinical database analysis and downstream patient-outcome prediction, introduced by the complexity inside the clinical records, represented by irregularity, multimodality, and inherent medical context. We propose that the key to addressing this challenge is to achieve seamless inter-modality interactions, defined as the capability to achieve bidirectional relationships from interleaved multimodal inputs. Accordingly, we design a novel clinical time-series model, Multimodal Irregular Seamless Time-series Transformer (MIST). Our experiments on clinical databases demonstrate the strong performance of MIST. Furthermore, our case study results serve as strong proof that MIST captures the inner connections between the multimodal information. The limitation of our paper is that our experiments are limited to clinical databases and only tested on patient-outcome predictions. Still, we believe this paper can be expanded to more domains that share similar challenges.

7 AI USE STATEMENT

During this study, we used generative AI tools for: create or modify scientific figures or images, suggest experimental parameters, create or edit software code, sourcing/searching for information, edit a research paper to improve readability, clean and reformat dataset. We have not used generative AI tools for: draft parts of a research paper, summarize or analyse existing literature, discover research topics or identify gaps, format references, suggest a structure for a research paper, propose a title or keywords for a research paper, generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, interpret results, design or provide feedback on methodology or experiments, identify relevant literature. And these tasks do not apply for this paper: Formulate questions for surveys or interviews, creation of artifacts, transcribe recordings of research material, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, assist with translation.

REFERENCES

Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in neural information processing systems*, 34:24206–24221, 2021.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. In *Proceedings of the 2nd clinical natural language processing workshop*, pp. 72–78, 2019.

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*, 2024.

Ching Chang, Jeehyun Hwang, Yidan Shi, Haixin Wang, Wen-Chih Peng, Tien-Fu Chen, and Wei Wang. Time-imm: A dataset and benchmark for irregular multimodal multivariate time series. *arXiv preprint arXiv:2506.10412*, 2025.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8(1):6085, 2018.

Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. *Advances in neural information processing systems*, 29, 2016.

Wei Dai, Peilin Chen, Malinda Lu, Daniel A Li, Haowen Wei, Hejie Cui, and Paul Pu Liang. Climb: Data foundations for large scale multimodal clinical foundation models. In *International Conference on Machine Learning*, pp. 11904–11953. PMLR, 2025.

Elizabeth R DeLong, David M DeLong, and Daniel L Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, pp. 837–845, 1988.

Jean Digitale, Deborah Franzon, Mark J Pletcher, Charles E McCulloch, and Efstathios D Gennatas. Methods for addressing missingness in electronic health record data for clinical prediction models: comparative evaluation. *JMIR Medical Informatics*, 13(1):e79307, 2025.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15180–15190, 2023.

Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific data*, 6(1):96, 2019.

Nasir Hayat, Krzysztof J Geras, and Farah E Shamout. Medfuse: Multi-modal fusion with clinical time-series data and chest x-ray images. In *Machine Learning for Healthcare Conference*, pp. 479–503. PMLR, 2022.

Max Horn, Michael Moor, Christian Bock, Bastian Rieck, and Karsten Borgwardt. Set functions for time series. In *International Conference on Machine Learning*, pp. 4353–4363. PMLR, 2020.

Stephanie L Hyland, Martin Faltys, Matthias Hüser, Xinrui Lyu, Thomas Gumbsch, Cristóbal Esteban, Christian Bock, Max Horn, Michael Moor, Bastian Rieck, et al. Early prediction of circulatory failure in the intensive care unit using machine learning. *Nature medicine*, 26(3):364–373, 2020.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

Harshavardhan Kamarthi. Large pre-trained time series models for cross-domain time series analysis tasks. *Advances in Neural Information Processing Systems*, 37:56190–56214, 2024.

Seyed Mehran Kazemi, Rishab Goel, Sepehr Eghbali, Janahan Ramanan, Jaspreet Sahota, Sanjay Thakur, Stella Wu, Cathal Smyth, Pascal Poupart, and Marcus Brubaker. Time2vec: Learning a vector representation of time. *arXiv preprint arXiv:1907.05321*, 2019.

Swaraj Khadanga, Karan Aggarwal, Shafiq Joty, and Jaideep Srivastava. Using clinical notes with time series data for icu management. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 6432–6437, 2019.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.

Tong Li, Zhaoyang Liu, Yanyan Shen, Xue Wang, Haokun Chen, and Sen Huang. Master: Market-guided stock transformer for stock price forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 162–170, 2024.

Zachary C Lipton, David C Kale, Charles Elkan, and Randall Wetzel. Learning to diagnose with lstm recurrent neural networks. *arXiv preprint arXiv:1511.03677*, 2015.

Zachary C Lipton, David C Kale, Randall Wetzel, et al. Modeling missing data in clinical time series with rnns. *Machine Learning for Healthcare*, 56(56):253–270, 2016.

Chenghao Liu, Taha Aksu, Juncheng Liu, Xu Liu, Hanshu Yan, Quang Pham, Silvio Savarese, Doyen Sahoo, Caiming Xiong, and Junnan Li. Moirai 2.0: When less is more for time series forecasting. *arXiv preprint arXiv:2511.11698*, 2025.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024a.

Haoxin Liu, Shangqing Xu, Zhiyuan Zhao, Lingkai Kong, Harshavardhan Prabhakar Kamarthi, Aditya Sasanur, Megha Sharma, Jiaming Cui, Qingsong Wen, Chao Zhang, et al. Time-mmd: Multi-domain multimodal dataset for time series analysis. *Advances in Neural Information Processing Systems*, 37:77888–77933, 2024b.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. In *The Twelfth International Conference on Learning Representations*, 2024c. URL https://openreview.net/forum?id=JePfAI8fah.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pp. 353–367. PMLR, 2023.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. Scalable and accurate deep learning with electronic health records. *NPJ digital medicine*, 1(1):18, 2018.

Eduardo A Trujillo Rivera, James M Chamberlain, Anita K Patel, Hiroki Morizono, Julia A Heneghan, and Murray M Pollack. Dynamic mortality risk predictions for children in icus: development and validation of machine learning models. *Pediatric Critical Care Medicine*, 23(5):344–352, 2022.

Emma Rocheteau, Pietro Liò, and Stephanie Hyland. Temporal pointwise convolutional networks for length of stay prediction in the intensive care unit. In *Proceedings of the conference on health, inference, and learning*, pp. 58–68, 2021.

Alexander Rodriguez, Harshavardhan Kamarthi, Pulak Agarwal, Javen Ho, Mira Patel, Suchet Sapre, and B Aditya Prakash. Machine learning for data-centric epidemic forecasting. *Nature Machine Intelligence*, 6(10):1122–1131, 2024.

Satya Narayan Shukla and Benjamin Marlin. Multi-time attention networks for irregularly sampled time series. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=4c0J6lwQ4_`.

Luis R Soenksen, Yu Ma, Cynthia Zeng, Leonard Boussioux, Kimberly Villalobos Carballo, Liangyuan Na, Holly M Wiberg, Michael L Li, Ignacio Fuentes, and Dimitris Bertsimas. Integrated multimodal artificial intelligence framework for healthcare applications. *NPJ digital medicine*, 5(1):149, 2022.

Hans-Christian Thorsen-Meyer, Annelaura B Nielsen, Anna P Nielsen, Benjamin Skov Kaas-Hansen, Palle Toft, Jens Schierbeck, Thomas Strøm, Piotr J Chmura, Marc Heimann, Lars Dybdahl, et al. Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records. *The Lancet Digital Health*, 2(4):e179–e191, 2020.

Sindhu Tipirneni and Chandan K Reddy. Self-supervised transformer for sparse and irregularly sampled multivariate clinical time-series. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 16(6):1–17, 2022.

Betty Van Aken, Jens-Michalis Papaioannou, Manuel Mayrdorfer, Klemens Budde, Felix Gers, and Alexander Loeser. Clinical outcome prediction from admission notes using self-supervised knowledge integration. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 881–893, 2021.

Shirly Wang, Matthew BA McDermott, Geeticka Chauhan, Marzyeh Ghassemi, Michael C Hughes, and Tristan Naumann. Mimic-extract: A data extraction, preprocessing, and representation pipeline for mimic-iii. In *Proceedings of the ACM conference on health, inference, and learning*, pp. 222–235, 2020.

Frank Wilcoxon. Individual comparisons by ranking methods. In *Breakthroughs in statistics: Methodology and distribution*, pp. 196–202. Springer, 1992.

Vijaya Krishna Yalavarthi, Kiran Madhusudhanan, Randolf Scholz, Nourhan Ahmed, Johannes Burchert, Shayan Jawed, Stefan Born, and Lars Schmidt-Thieme. Grafiti: Graphs for forecasting irregularly sampled time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 16255–16263, 2024.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 11121–11128, 2023.

Xian Zeng, Gang Yu, Yang Lu, Linhua Tan, Xiujing Wu, Shanshan Shi, Huilong Duan, Qiang Shu, and Haomin Li. Pic, a paediatric-specific intensive care database. *Scientific data*, 7(1):14, 2020.

Chaohe Zhang, Xu Chu, Liantao Ma, Yinghao Zhu, Yasha Wang, Jiangtao Wang, and Junfeng Zhao. M3care: Learning with missing modalities in multimodal healthcare data. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 2418–2428, 2022a.

Weijia Zhang, Chenlong Yin, Hao Liu, Xiaofang Zhou, and Hui Xiong. Irregular multivariate time series forecasting: A transformable patching graph neural networks approach. In *Forty-first International Conference on Machine Learning*, 2024.

Xiang Zhang, Marko Zeman, Theodoros Tsiligkaridis, and Marinka Zitnik. Graph-guided network for irregularly sampled multivariate time series. In *International Conference on Learning Representations*, 2022b. URL https://openreview.net/forum?id=Kwm8I7dU-l5.

Xinlu Zhang, Shiyang Li, Zhiyu Chen, Xifeng Yan, and Linda Ruth Petzold. Improving medical predictions by irregular multimodal electronic health records modeling. In *International conference on machine learning*, pp. 41300–41313. PMLR, 2023.

Yuwei Zhang, Kumar Ayush, Siyuan Qiao, A. Ali Heydari, Girish Narayanswamy, Maxwell A Xu, Ahmed Metwally, Jinhua Xu, Jake Garrison, Xuhai Xu, Tim Althoff, Yun Liu, Pushmeet Kohli, Jiening Zhan, Mark Malhotra, Shwetak Patel, Cecilia Mascolo, Xin Liu, Daniel McDuff, and Yuzhe Yang. SensorLM: Learning the language of wearable sensors. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=TrHeq0yFhv.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 11106–11115, 2021.

A Dataset Details

A.1 Cohort Establishment

CHoA This dataset consists of electronic health records maintained by Children's Healthcare of Atlanta. We take all records up to August 2025 and retain admissions that (1) have a length of stay longer than 48 hours and (2) have valid prediction targets and at least 20 numeric records in the first 24 hours; from these, we keep the 50,000 most recent admissions.

MIMIC We take the CHARTEVENTS table from the MIMIC-III database and preprocess it as follows. We first merge and clean features following the STraTS pipeline (Tipirneni & Reddy, 2022), and select 'Temp', 'SpO2', 'DBP', 'SBP', 'Resp', and 'Pulse' as the vital signs. We then follow the selection criteria of prior clinical time-series work (Tipirneni & Reddy, 2022; Zhang et al., 2023; Yalavarthi et al., 2024) to build a cohort of patients who (1) have a length of stay longer than 48 hours, (2) have no prior ICU admission (to prevent patient-level leakage), (3) have records for at least 4 vital signs between hours 24 and 48, and (4) have at least 20 numeric records in the first 24 hours. As mentioned in Section 3, we only retain short, time-stamped free-text events rather than free-text documents such as discharge summaries, which also removes the primary source of outcome leakage in MIMIC-III.

Statistics of the selected cohorts over the first 48 hours are given in Table 6. In addition, Table 7 compares the outcome distributions of retained and held-out MIMIC patients, showing that our selection criteria do not introduce a meaningful selection bias.

Table 6: Cohort statistics: (1) cohort size (patient stays), (2–3) average number of numeric and textual records per patient, (4) modality-mix ratio (percentage of records immediately followed by a record of a different modality), (5) adverse-outcome rate (%), (6) mean log length-of-stay, and (7) average phenotype-label prevalence (%). Both cohorts are irregular and multimodal, and the effect is more pronounced on CHoA owing to its pediatric setting.

Dataset	Cohort Size	Avg. Num. Records	Avg. Text Records	Modality Mix Ratio	Adverse Rate	Mean log -LOS	Phenotype Prevalence
CHoA	50000	512.8	290.9	20.8	1.1	3.5	-
MIMIC	12197	485.1	397.5	10.8	14.4	4.8	20.3

Table 7: Outcome distributions of retained and held-out MIMIC patients. The two groups are closely matched, indicating that our selection criteria do not introduce a meaningful selection bias.

	Retained	**Held-out**	**All patients**
Stays	12,197	7,886	20,083
Age, median (IQR)	66 (54–78)	68 (54–80)	67 (54–79)
LOS hours, median (IQR)	101.9 (68.8–197.0)	82.9 (62.7–138.0)	94.3 (66.1–170.4)
In-hospital mortality, percentage	14.4	12.2	13.6
Phenotype labels per stay, mean	5.07	4.67	4.91

B Implementation Details

B.1 Hardware Specifications

Experiments on CHoA were run on an NVIDIA H100 NVL GPU with an Intel Xeon Gold CPU. Experiments on MIMIC-III were run on an NVIDIA H100 (HBM3) GPU with Intel Xeon Platinum 8580 CPUs.

B.2 Dataset Division

For both MIMIC and CHoA, we split the data 70:30 into training and test sets by patient ID. A further 15% of the training data is held out as a validation set, which is used to select the best checkpoint of each run.

B.3 MIST

Output projection design For forecasting, we directly project the output embedding into the vital-sign space. For adverse-outcome prediction, phenotyping, and length-of-stay regression, we apply timestep-wise pooling followed by a two-layer MLP as the prediction head for each task, with task-specific output dimensions.

Hyperparameters For all MIST experiments, we use an encoder dimension of 128, 8 attention heads, and an attention depth of 4. The fusion gate is initialized to 0.1 and is active from the first step, and as a regularizer we mask out the text branch on 30% of training batches. We train with a learning rate of 10^{-4} for up to 100 epochs, with early stopping after 20 steps without improvement.

B.4 Baseline Models

By default, we configure the baselines with the same encoder dimension, number of attention heads, learning rate, and attention depth as MIST, unless the original papers suggest otherwise. All remaining hyperparameters (e.g., the patch configuration of PatchTST and the GNN configurations of GraFITi and RAINDROP) follow the original papers. For IMM-TSF, we use ClinBERT as the language encoder and GraFITi as the time-series encoder.

Among the baselines, only UTDE and IMM-TSF accept multimodal inputs, so they receive the full records. General-purpose time-series models are restricted to vital signs, since they cannot handle a variable number of input features. For the clinical time-series models other than UTDE, since feature selection can affect their performance, we report the better result under two settings: (1) using only vital signs as input features, and (2) using all available numeric records in the dataset. Some baselines are implemented only for certain tasks; for the remaining tasks, we keep their main model structure and replace only the output projection head as a straightforward adaptation.

For IMM-TSF, since it's capable of using a variety of time-series and language backbones, we instantiate its language and time-series encoders with ClinicalBERT and GraFITi so that its backbones match MIST's, isolating the effect of the fusion design.

B.5 Case Study Generation

The patients used in our case studies come from clinical databases with restricted access, so we are unable to present real patient records. Instead, we generate synthetic patients by perturbing and rounding both the recorded values and the recorded timestamps of the original records, and we further mask rare diagnoses and medications that could pose a re-identification risk.

B.6 Late-fusion Ablation Design

To isolate the contribution of our seamless attention under a fair comparison to the previous late-fusion mechanism, we established a variant of MIST where numeric and textual events are routed into two seperate streams, each with corresponding input tokens of MIST and its own stacked self-attention layers yet no interleaving and no cross-modal attention layers. The two streams are only interacted by a late cross-modal fusion module using the same gated cross-attention. We ensure all the hyperparameters the same as MIST, including the depth of stacks.

C Supplementary Experiments

C.1 Computational Cost of MIST

We analyze the computational cost of MIST. Table 8 reports training time, inference time, and peak memory usage at batch size 64 for various input lengths, using the same hardware and model configuration as above. We emphasize that the length-2048 setting is a stress test only; as Table 6 shows, such lengths are rarely reached in real-world patient cohorts.

Table 8: Time and memory usage of MIST on MIMIC at batch size 64.

Input Length	64	128	256	512	1024	2048
Train (ms)	56.5	58.9	66.0	98.9	208.5	597.0
Inference (ms)	20.5	20.7	23.3	33.7	71.4	251.3
Peak GPU Memory (GB)	4.8	5.0	6.6	11.3	26.4	79.0

C.2 Paired Significance Test

To verify that MIST's improvements in patient-outcome prediction are not attributable to seed-level randomness, we perform paired significance tests on MIMIC between MIST and the three strongest baselines: GRU-D, GraFITi, and UTDE. Specifically, we apply a paired DeLong test (DeLong et al., 1988) on the ROC curves for in-hospital mortality and phenotyping, and a Wilcoxon signed-rank test (Wilcoxon, 1992) for length-of-stay regression and forecasting. Since phenotyping is a 25-label classification task, we combine the per-label p-values using Fisher's method. As shown in Table 9, all p-values are small, confirming that the performance gaps between MIST and the baselines are statistically significant.

Table 9: p-values of paired significance tests between MIST and the strongest baselines on MIMIC.

	GraFITi	GRU-D	UTDE
In-hospital mortality	4.80E-22	4.20E-12	2.80E-14
Phenotyping	4.90E-157	1.10E-159	7.10E-132
Length of stay	4.10E-33	9.00E-26	1.30E-20
Forecasting	0.005	1.90E-11	5.70E-14

C.3 Choice of Vital Encoder

We replace the vital encoder of MIST (GraFITi) with GRU-D, mTAND, t-PatchGNN, and RAINDROP. As shown in Table 10, MIST consistently improves over the corresponding unimodal baseline regardless of the encoder choice. On length-of-stay regression, MIST with a GRU-D or RAINDROP encoder even slightly outperforms MIST with GraFITi. This demonstrates that MIST can robustly build on a wide range of clinical time-series architectures and improve their performance by incorporating multimodal information.

C.4 Performance on the Held-out Cohort

To further address the concern of selection bias from MIMIC cohort filtering, we conduct additional experiments in which models are trained on the full cohort and evaluated on the held-out cohort. We note that the held-out patients have limited observations, which makes the evaluation results considerably less reliable; this is why they are not included in the main results. Nevertheless, Tables 11–12 show that MIST still outperforms the baselines that performed best on the retained cohort, suggesting that the gains reported in Tables 1–3 are not an artifact of selection bias.

Table 10: **Ablation of vital sign encoder.** Each row shows MIST's performance with a different baseline as the vital-sign encoder, and the gap relative to that baseline alone.

	Mortality (AUROC↑)		**LOS** (MSE↓)		**Phenotype** (AUROC↑)	
Vital Encoder	MIST	Δ	MIST	Δ	MIST	Δ
GRU-D	$.775 \pm .006$	+.050	$.404 \pm .007$	-.070	$.700 \pm .001$	+.008
mTAND	$.771 \pm .008$	+.061	$.409 \pm .001$	-.073	$.720 \pm .006$	+.047
t-PatchGNN	$.713 \pm .013$	+.014	$.476 \pm .005$	-.017	$.706 \pm .005$	+.030
RAINDROP	$.709 \pm .004$	+.063	$.406 \pm .009$	-.204	$.683 \pm .002$	+.044

Table 11: Phenotyping and mortality prediction performance of MIST and the strongest baselines, trained on the full cohort and evaluated on the held-out cohort.

	Phenotyping		**Mortality**	
Method	AUROC↑	AUPRC↑	AUROC↑	AUPRC↑
GraFITi	0.693 ± 0.004	0.286 ± 0.008	0.694 ± 0.001	0.348 ± 0.000
GRU-D	0.723 ± 0.002	0.304 ± 0.004	0.699 ± 0.002	0.354 ± 0.002
UTDE	0.706 ± 0.002	0.290 ± 0.003	0.698 ± 0.004	0.351 ± 0.005
IMM-TSF	0.690 ± 0.005	0.276 ± 0.015	0.667 ± 0.003	0.317 ± 0.003
MIST (ours)	**0.786 ± 0.019**	**0.381 ± 0.034**	**0.725 ± 0.005**	**0.384 ± 0.007**

Table 12: Length-of-stay regression and vital-sign forecasting performance of MIST and the strongest baselines, trained on the full cohort and evaluated on the held-out cohort.

	Length-of-Stay		**Forecasting**	
Method	MSE↓	MAE↓	MSE↓	MAE↓
GraFITi	0.428 ± 0.001	0.520 ± 0.000	0.586 ± 0.000	0.528 ± 0.000
GRU-D	0.434 ± 0.000	0.528 ± 0.001	0.590 ± 0.000	0.532 ± 0.000
UTDE	0.429 ± 0.001	0.522 ± 0.001	0.590 ± 0.001	0.532 ± 0.000
IMM-TSF	0.452 ± 0.002	0.541 ± 0.002	0.600 ± 0.000	0.535 ± 0.000
MIST (ours)	**0.387 ± 0.001**	**0.484 ± 0.000**	**0.585 ± 0.000**	**0.527 ± 0.000**