PIKE: A Multimodal Framework Fusing Physiological Priors and Knowledge Graphs for Short-Sequence EHR Prediction

Tuan-Cuong Vuong, Trang Xuan Mai, Son Thai Mai, Tung Duong Ta,
Vu-Duc Ngo, Tien-Cuong Nguyen, Huan Vu, and Thien Van Luong

***Abstract*—Clinical outcome prediction from Electronic Health Records (EHR) in early-stage ICU settings faces three key challenges. First, restricted observation windows combined with irregular sampling result in *short-sequence, sparse* data where temporal dynamics are difficult to learn reliably. Second, existing knowledge graph integration methods often either lack patient-specific subgraph construction or rely on probabilistic extraction that may introduce spurious relations. Third, multimodal fusion must account for uneven modality reliability and control parameter budget under data scarcity. We propose PIKE, a novel framework tailored for short-sequence, sparse clinical prediction. We design a Hybrid Set-Sequence Encoder that incorporates variable-name embeddings and initializes decay rates using physiological half-lives. We construct deterministic patient knowledge graphs through controlled multi-hop expansion in PrimeKG with evidence-based relation scoring. We propose Bidirectional Low-Rank Fusion (BLRF), a parameter-efficient dual-pathway architecture balancing stable optimization with expressive cross-modal interactions. Experiments on MIMIC-III and MIMIC-IV demonstrate that PIKE achieves 90.85% and 93.86% AUROC for mortality prediction, corresponding to absolute improvements of 11.03% and 17.45% over state-of-the-art baselines. Our source code is available at https://anonymous.4open.science/r/pike.**

***Index Terms*—Electronic Health Records, Multimodal Learning, Knowledge Graph, Clinical Time-Series, Data Sparsity, Multimodal Fusion.**

I. Introduction

Electronic Health Records (EHR) record multimodal patient information, including time-series measurements, clinical notes, and diagnostic codes [1]. ICU predictive models leverage these data to support critical clinical tasks such as mortality prediction and readmission forecasting [2]. Early-stage prediction operates under restricted observation windows [3], where benchmark discretization pipelines divide continuous recordings into time steps within a fixed window. However, clinical data exhibits irregular sampling and high missingness [4]. This leads to *short-sequence, sparse* data that challenges predictive models relying on sufficient temporal context.

Prior work has addressed various aspects of EHR prediction through multiple approaches. Temporal modeling methods have employed recurrent architectures [5] and attention mechanisms designed for irregular time series [6]. Clinical text has been leveraged through pretrained language models [7]. Motivated by complementary signals across data sources, multimodal fusion frameworks have been developed to combine time-series, notes, and codes [2, 8]. More recently, knowledge-enhanced methods have linked clinical codes to external biomedical knowledge graphs to augment model understanding [2, 9–11]. However, these approaches are typically evaluated under settings with longer temporal horizons and richer context. The challenges posed by *short-sequence, sparse* data in early-stage ICU prediction have received limited attention.

First, temporal modeling in early-stage ICU faces fundamental challenges due to *short-sequence, sparse* observations. In our setting, restricted observation windows [3] combined with benchmark discretization pipelines result in 4–6 timesteps, while irregular sampling and high missingness [4] further reduce usable temporal evidence. Sequential models such as LSTM [12] and RNN [5] typically benefit from longer temporal context; under short horizons with high missingness, temporal dynamics become under-determined. Time-aware methods such as GRU-D [4] address irregularity through elapsed-time-based exponential decay, but decay dynamics are typically learned from data without explicit physiological constraints and may become under-determined in short-sequence, sparse settings. Attention-based models for irregular time series [6] exist but remain data-driven and may require more context to learn reliable interactions under extremely short horizons. When sequences are short, strong predictive signals often reside in cross-variable correlations among multiple labs and vitals within the same time step. Set functions for time series have been explored [13, 14] to capture such correlations through permutation-invariant aggregation. However, in standard benchmark formulations [3], variables are represented as channels in matrix form, leaving feature semantics (e.g., lab test names) implicit; under short-sequence regimes, learning correlations purely from data becomes more difficult.

Second, when temporal context is limited, knowledge graphs can provide additional semantic and biomedical context to complement sparse observations [9, 11]. Prior knowledge-enhanced methods can be grouped into three categories. Ontology-based methods such as GRAM [9] and KAME [10] leverage medical ontologies to regularize or augment code representations through graph attention. However, they do not explicitly formulate patient-specific subgraph construction as a module for external structured knowledge in early-stage prediction. Moreover, these methods often emphasize code-level ontology structure and may offer limited mechanism-level connections from high-level phenotypes to underlying

T.-C. Vuong and T. X. Mai are with A2I Lab, School of Computing, Phenikaa University, Ha Noi, Viet Nam (e-mail: cuong.vuongtuan@phenikaa-uni.edu.vn; trang.maixuan@phenikaa-uni.edu.vn).

Thai Son Mai is with Queen's University Belfast, Belfast, UK (email: thaison.mai@qub.ac.uk).

T. V. Luong and H. Vu are with Business AI Lab, Faculty of DS&AI, College of Technology, National Economics University, Ha Noi, Viet Nam (e-mail: thienlv@neu.edu.vn, huanv@neu.edu.vn).

biomedical pathways. LLM-based methods such as GraphCare [11] generate personalized patient graphs through probabilistic extraction and linking. Retrieval-augmented methods such as EMERGE [2] integrate knowledge through entity linking and retrieval. Both approaches provide richer context but may introduce spurious relations that are difficult to verify. Small errors in extraction or linking may propagate through downstream reasoning. Path-based and local-global knowledge augmentation methods [15, 16] demonstrate that multi-hop reasoning can be beneficial. However, unconstrained multi-hop expansion risks noise amplification and scalability issues, while overly restrictive constraints limit connectivity between phenotypes and mechanisms. Therefore, there is a need for deterministic, patient-specific knowledge graph construction grounded in curated biomedical resources such as PrimeKG [17], with controlled multi-hop expansion to bridge high-level phenotypes and underlying mechanisms under short-sequence, sparse conditions.

Third, multimodal fusion in early-stage ICU faces challenges due to uneven modality reliability and data scarcity [1, 8]. In early-stage settings, modalities exhibit varying quality: time-series suffer from short-sequence and missingness [4], clinical notes may be sparse or unavailable, and diagnostic codes are coarse, phenotype-level summaries that may be temporally misaligned with early physiological changes. Fusion mechanisms should account for these reliability differences rather than uniformly combining modalities. Cross-attention-based fusion methods [2] enable rich interactions but do not inherently enforce balanced modality contributions; learned attention weights may be sensitive to modality noise and availability under sparse data [8]. Moreover, simple fusion strategies such as concatenation or gating [18] can be competitive with complex attention mechanisms, indicating that high-capacity interactions may not be well-supported by limited early-stage observations. Dataset shift between MIMIC-III [19] and MIMIC-IV [20] may expose modality bias, as differences in recording practices and schema can alter modality usefulness; fusion may be brittle under such dataset shift. Additionally, when knowledge graphs or retrieval-augmented summaries are integrated [2], these knowledge-enhanced modalities may dominate, risking over-reliance and neglect of time-series evidence. Therefore, fusion should constrain interaction complexity, preserve unimodal predictive paths (e.g., residual connections), and be robust to modality imbalance and shift.

To address these gaps, we propose PIKE, a framework tailored for early-stage clinical prediction. Our contributions are:

1) **Hybrid Set-Sequence Encoder.** We combine a permutation-invariant Feature-Set Encoder that incorporates variable-name embeddings to capture cross-variable correlations, with a Sequence Encoder whose decay rates are initialized using physiological half-lives [21] and refined via learnable residuals. This design addresses under-determined temporal dynamics in short-sequence, sparse settings.
2) **Deterministic Patient Knowledge Graph Construction.** We design a three-phase pipeline grounded in PrimeKG [17]: semantic matching to link diagnostic codes to ontology nodes, controlled multi-hop expansion to connect phenotypes with underlying mechanisms, and GRADE-inspired relation scoring [22] to prune noisy nodes. This provides verifiable biomedical context without probabilistic extraction.
3) **Bidirectional Low-Rank Fusion (BLRF).** We propose a dual-pathway residual architecture: a base concatenation pathway for stable optimization, and an interaction pathway combining bidirectional cross-attention with low-rank bilinear pooling [23]. A learnable scaling parameter β (initialized to zero) balances expressiveness and parameter efficiency under data scarcity.

We conduct extensive experiments on MIMIC-III [19] and MIMIC-IV [20], including ablation studies, sensitivity analyses, and comparisons with state-of-the-art temporal, multimodal, and knowledge-enhanced baselines, demonstrating consistent gains in both AUROC and AUPRC under short-sequence constraints.

II. Related Work

A. Multimodal EHR Learning

Advances in deep learning have enabled the analysis of heterogeneous EHR modalities. Early research primarily targeted individual data types, such as recurrent architectures for time-series data [4] and transformer-based models for clinical text [7]. For irregular clinical time-series, STraTS [6] introduced time-aware attention mechanisms, while Raindrop [24] employed graph-guided networks to model inter-variable dependencies. To capture complementary information across sources, subsequent studies introduced multimodal fusion strategies: M3Care [8] addressed missing modalities via modality-adaptive similarity, and VecoCare [25] utilized contrastive learning between visit sequences and clinical narratives. Fusion architectures range from simple concatenation to gated mechanisms [18] and cross-attention [2], with varying robustness to modality imbalance and dataset shift. However, these works are typically evaluated with richer temporal context and do not explicitly target extremely short-horizon early-stage windows where modality reliability varies significantly.

B. Knowledge Graph Enhanced EHR Prediction

Knowledge graphs (KGs) provide a structured framework for incorporating external medical expertise into EHR predictive models. Early research focused on hierarchical ontologies, with GRAM [9] and KAME [10] utilizing graph attention and ontology embeddings to refine medical concept representations. Subsequent models like MedPath [15] and MedRetriever [26] integrated high-order connectivity and disease-specific documentation to enhance interpretability. Recently, LLMs have been used for knowledge extraction and linking; GraphCare [11] constructs personalized KGs via prompting, while EMERGE [2] leverages RAG to align EHR entities with PrimeKG [17]. Specialized approaches such as CGL [27] and KerPrint [16] further address patient-disease interactions and local-global KG signals. However, existing KG integration often focuses on representation augmentation

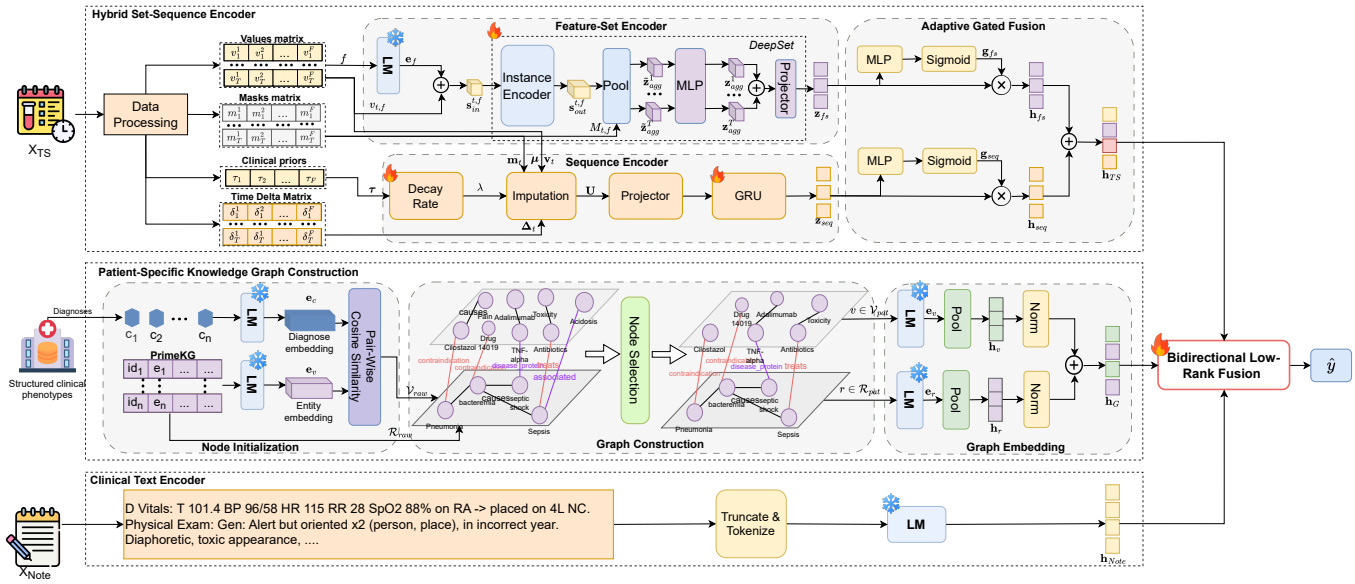

Fig. 1. Overview of the PIKE framework. It consists of three key components: (1) Hybrid Set-Sequence Encoder, (2) Patient Knowledge Graph Construction, and (3) Bidirectional Low-Rank Fusion (BLRF). "LM" denotes Language Model.

or probabilistic extraction rather than deterministic patient-specific multi-hop construction under short-sequence, sparse settings.

III. Proposed Framework

Figure 1 illustrates the overall architecture of PIKE, which comprises three core components addressing the problems identified in Section I. We address the problem of learning predictive models from multimodal sources under sparse, short-sequence observations. For a given patient, we define three primary input modalities $\mathcal{X} = \{\mathcal{X}_{TS}, \mathcal{X}_{Note}, \mathcal{X}_G\}$:

Time-Series Modality ($\mathcal{X}_{TS}$). We represent the physiological state as a multivariate time-series matrix $\mathbf{X}_{TS} \in \mathbb{R}^{T\times F}$, where T denotes the number of discretized time steps and F refers to the number of clinical features.

Clinical Notes Modality ($\mathcal{X}_{Note}$). Unstructured notes document patient status and assessments within the observation period.

Knowledge Graph Modality ($\mathcal{X}_G$). To augment limited observations, we construct a patient-specific knowledge graph $\mathcal{G}_{pat}$ from structured clinical phenotypes, specifically diagnosis codes $\mathcal{D}$. We leverage PrimeKG [17], a curated biomedical knowledge graph $\mathcal{G}_{KG} = (\mathcal{V}_{KG}, \mathcal{R}_{KG})$, where $\mathcal{V}_{KG}$ is the set of biomedical entities and $\mathcal{R}_{KG}$ is the set of relation types. The constructed patient knowledge graph $\mathcal{G}_{pat} = (\mathcal{V}_{pat}, \mathcal{R}_{pat})$ serves as the graph object for this modality.

Throughout the model, we use h to denote the hidden dimension and employ pre-trained LMs to generate semantic embeddings with dimension d.

A. Hybrid Set-Sequence Encoder

1) Feature-Set Encoder: When sequences are short, strong predictive signals often reside in cross-variable correlations among multiple labs and vitals within the same timestep, but standard benchmark formulations leave feature semantics (e.g., lab test names) implicit. To address this, we model each timestep as a *permutation-invariant set* [14] and incorporate variable-name embeddings to make feature semantics explicit.

Given the values matrix $\mathbf{V} \in \mathbb{R}^{T\times F}$ containing measurements across T timesteps and F clinical features, we extract feature names (e.g., "Heart Rate", "Mean Arterial Pressure") and encode them via a pretrained LM to obtain semantic embeddings $\mathbf{e}_f \in \mathbb{R}^d$ for each feature f. At each timestep t, we concatenate the scalar value with its semantic embedding:

$$\mathbf{s}_{in}^{t,f} = \text{Concat}(v_{t,f}, \mathbf{e}_f) \in \mathbb{R}^{1+d}. \tag{1}$$

Instance Encoding. Each element embedding $\mathbf{s}_{in}^{t,f}$ is independently transformed via a shared two-layer MLP $\Psi : \mathbb{R}^{1+d} \to \mathbb{R}^h$:

$$\mathbf{s}_{out}^{t,f} = \text{ReLU}(\mathbf{W}_{\psi,2} \cdot \text{ReLU}(\mathbf{W}_{\psi,1}\mathbf{s}_{in}^{t,f} + \mathbf{b}_{\psi,1}) + \mathbf{b}_{\psi,2}) \in \mathbb{R}^h. \tag{2}$$

where $\mathbf{W}_{\psi,1} \in \mathbb{R}^{h\times(1+d)}$, $\mathbf{W}_{\psi,2} \in \mathbb{R}^{h\times h}$. This step maps heterogeneous feature-value pairs into a shared latent space.

Given the observation mask matrix $\mathbf{M} \in \{0,1\}^{T\times F}$ where $M_{t,f} = 1$ indicates that feature f was observed (i.e., not NaN) at timestep t, we aggregate transformed features using masked summation to ensure only observed features contribute without introducing ordering or missingness bias:

$$\tilde{\mathbf{z}}_{agg}^t = \sum_{f=1}^{F} M_{t,f} \cdot \mathbf{s}_{out}^{t,f} \in \mathbb{R}^h. \tag{3}$$

Following the DeepSets framework [14], the pooled representation is refined through a two-layer MLP Φ:

$$\mathbf{z}_{agg}^t = \mathbf{W}_{\phi,2} \cdot \text{ReLU}(\mathbf{W}_{\phi,1}\tilde{\mathbf{z}}_{agg}^t + \mathbf{b}_{\phi,1}) + \mathbf{b}_{\phi,2} \in \mathbb{R}^h. \tag{4}$$

Temporal Aggregation. To obtain the final feature-set representation, we concatenate timestep-wise set representations and apply a linear projection to summarize set-level temporal dynamics:

$$\mathbf{z}_{fs} = \mathbf{W}_{temp}[\mathbf{z}_{agg}^1; \ldots; \mathbf{z}_{agg}^T] + \mathbf{b}_{temp} \in \mathbb{R}^h. \tag{5}$$

2) Sequence Encoder: Under short-sequence, sparse settings, temporal dynamics become under-determined, and existing time-aware methods such as GRU-D [4] learn decay dynamics from data without explicit physiological constraints. To address this, we initialize decay rates using physiological half-lives and refine them via learnable residuals.

We initialize decay rates $\boldsymbol{\tau} \in \mathbb{R}^F$ where τ_f denotes the physiological half-life (in hours) of feature f, representing clinically motivated time scales for information decay (e.g., 0.5 hours for vital signs, 24 hours for laboratory values; detailed in Appendix A). These priors are refined via learnable residuals $\delta_f \in \mathbb{R}$ (initialized to zero) to adapt to data-specific dynamics:

$$\tau'_f = \tau_f \cdot \exp(\delta_f), \quad \lambda_f = \frac{\ln(2)}{\tau'_f}, \tag{6}$$

where $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_F] \in \mathbb{R}^F$ are the effective decay rates controlling how fast information fades.

Decay Imputation. To handle missing values, we first compute the global empirical mean $\boldsymbol{\mu} \in \mathbb{R}^F$ for each feature from observed training data:

$$\mu_f = \frac{\sum_{t=1}^{T} m_{t,f} \cdot v_{t,f}}{\sum_{t=1}^{T} m_{t,f}}, \tag{7}$$

where $m_{t,f}$ is the observation mask and $v_{t,f}$ is the value at timestep t. The mean μ_f is computed globally across all patients and timesteps in the training set where feature f is observed.

Let $\boldsymbol{\Delta}_t \in \mathbb{R}^F$ denote the time gaps since the last observation for each feature. Specifically, $\Delta_{t,f} = 0$ if feature f is observed at timestep t (i.e., $m_{t,f} = 1$); otherwise, $\Delta_{t,f}$ equals the time elapsed since the most recent timestep where $m_{t',f} = 1$ for $t' < t$. For each timestep t, let $\mathbf{v}_t, \mathbf{m}_t, \boldsymbol{\Delta}_t \in \mathbb{R}^F$ denote row vectors extracted from their respective matrices. We maintain a carry-forward state $\mathbf{x}_t \in \mathbb{R}^F$ initialized to $\mathbf{x}_0 = \boldsymbol{\mu}$:

$$\mathbf{x}_t = \mathbf{m}_t \odot \mathbf{v}_t + (\mathbf{1} - \mathbf{m}_t) \odot \mathbf{x}_{t-1}, \tag{8}$$

where $\odot$ denotes element-wise multiplication and $\mathbf{1} \in \mathbb{R}^F$ is the all-ones vector. We compute decay coefficients $\boldsymbol{\gamma}_t = \exp(-\boldsymbol{\lambda} \odot \boldsymbol{\Delta}_t) \in \mathbb{R}^F$, and impute missing values:

$$\mathbf{u}_t = \mathbf{m}_t \odot \mathbf{v}_t + (\mathbf{1} - \mathbf{m}_t) \odot (\boldsymbol{\gamma}_t \odot \mathbf{x}_{t-1} + (\mathbf{1} - \boldsymbol{\gamma}_t) \odot \boldsymbol{\mu}) \in \mathbb{R}^F. \tag{9}$$

Stacking row vectors yields the imputed matrix $\mathbf{U} = [\mathbf{u}_1; \ldots; \mathbf{u}_T] \in \mathbb{R}^{T \times F}$. We project the imputed sequence to hidden dimension and process with a GRU [28], retaining only the final hidden state:

$$\tilde{\mathbf{X}} = \mathbf{U}\mathbf{W}_{seq} + \mathbf{b}_{seq}^{\top} \in \mathbb{R}^{T \times h}, \quad \mathbf{z}_{seq} = \text{GRU}(\tilde{\mathbf{X}})_T \in \mathbb{R}^h. \tag{10}$$

3) Adaptive Gated Fusion: Unlike scalar gating methods that typically constrain modalities via a simplex (e.g., softmax normalization) [18], we employ independent element-wise gates, allowing each latent dimension to adaptively emphasize set-level or temporal information:

$$\mathbf{g}_{fs} = \sigma(\text{MLP}_{fs}(\mathbf{z}_{fs})), \quad \mathbf{g}_{seq} = \sigma(\text{MLP}_{seq}(\mathbf{z}_{seq})) \tag{11}$$

where MLP_{fs} and MLP_{seq} are single-layer networks (linear layers): $\text{MLP}_{fs}(\mathbf{z}) = \mathbf{W}_{fs}\mathbf{z} + \mathbf{b}_{fs}$ and similarly for MLP_{seq}. The final hybrid representation $\mathbf{h}_{TS}$ is a weighted sum:

$$\mathbf{h}_{TS} = \mathbf{g}_{fs} \odot \mathbf{z}_{fs} + \mathbf{g}_{seq} \odot \mathbf{z}_{seq} \in \mathbb{R}^h. \tag{12}$$

B. Patient Knowledge Graph Construction

When temporal context is limited, knowledge graphs can provide additional semantic and biomedical context, but existing methods either lack patient-specific subgraph construction or rely on probabilistic extraction that may introduce spurious relations, while unconstrained multi-hop expansion risks noise amplification. To address these limitations, we construct a patient knowledge graph $\mathcal{G}_{pat} = (\mathcal{V}_{pat}, \mathcal{R}_{pat})$ grounded in PrimeKG [17], a curated biomedical knowledge graph $\mathcal{G}_{KG} = (\mathcal{V}_{KG}, \mathcal{R}_{KG})$, through deterministic semantic matching and controlled multi-hop expansion with evidence-based relation scoring.

1) Node Initialization: We align the patient's clinical codes with PrimeKG entities through semantic matching. Given the patient's diagnosis set $\mathcal{D} = \{c_1, c_2, \ldots, c_N\}$ (ICD-9/10 codes), we extract the textual title of each diagnosis from the ICD dictionary and encode it via a pre-trained language model:

$$\mathbf{e}_{c_n} = \text{LM}(\text{title}(c_n)) \in \mathbb{R}^d, \quad n \in \{1, \ldots, N\}, \tag{13}$$

where $\text{title}(c_n)$ retrieves the long description (e.g., "Acute respiratory failure") from the ICD-9/10 dictionary. Similarly, for each entity $v \in \mathcal{V}_{KG}$, we encode its MONDO name from PrimeKG [17]:

$$\mathbf{e}_v = \text{LM}(\text{name}(v)) \in \mathbb{R}^d, \tag{14}$$

where $\text{name}(v)$ is the disease name field in PrimeKG (e.g., "respiratory failure"). We perform pairwise cosine similarity:

$$s(c_j, v) = \frac{\mathbf{e}_{c_j} \cdot \mathbf{e}_v}{\|\mathbf{e}_{c_j}\| \|\mathbf{e}_v\|}. \tag{15}$$

For each diagnosis, we select the top-M entities with similarity above threshold θ:

$$\mathcal{S}_j = \{v \in \mathcal{V}_{KG} \mid s(c_j, v) > \theta\}_{\text{top-}M}. \tag{16}$$

The initial node set aggregates matches across all diagnoses:

$$\mathcal{V}_0 = \bigcup_{j=1}^{N} \mathcal{S}_j. \tag{17}$$

2) Graph Construction: Starting from $\mathcal{V}_0$, we construct the patient knowledge graph through multi-hop expansion followed by node selection.

Multi-Hop Expansion. Let $\mathcal{N}_{KG}(v)$ denote the set of all neighbor entities directly connected to v via any relation in $\mathcal{R}_{KG}$. We extend this to a set of nodes: $\mathcal{N}_{KG}(\mathcal{V}) = \bigcup_{v \in \mathcal{V}} \mathcal{N}_{KG}(v)$. Starting from $\mathcal{V}_0$, we iteratively expand the node set for H hops:

$$\mathcal{V}_h = \mathcal{V}_{h-1} \cup \mathcal{N}_{KG}(\mathcal{V}_{h-1}), \quad h \in \{1, \ldots, H\}. \tag{18}$$

The raw expanded node set is:

$$\mathcal{V}_{raw} = \bigcup_{h=0}^{H} \mathcal{V}_h. \tag{19}$$

We then induce all relations from PrimeKG connecting pairs of nodes in $\mathcal{V}_{raw}$:

$$\mathcal{R}_{raw} = \{r \in \mathcal{R}_{KG}\}, \tag{20}$$

where r connects two nodes in $\mathcal{V}_H$. This yields the raw graph $\mathcal{G}_{raw} = (\mathcal{V}_{raw}, \mathcal{R}_{raw})$ before refinement.

Node Selection. To control the graph size while preserving clinically relevant nodes, we employ a relation scoring function $\omega : \mathcal{R}_{KG} \to \{1, 2, 3\}$ inspired by the GRADE evidence hierarchy [22]: interventional relations ($\omega = 3$), mechanistic relations ($\omega = 2$), and associative relations ($\omega = 1$) (detailed mapping is in Appendix A).

For each node $v \in \mathcal{V}_{raw}$, we compute its importance score by summing the weights of its connected relations with r connects v:

$$\sigma(v) = \sum_{r \in \mathcal{R}_{raw}} \omega(r). \tag{21}$$

We rank all nodes by $\sigma(v)$ in descending order and select the top N_{max} nodes:

$$\mathcal{V}_{pat} = \{v \in \mathcal{V}_{raw}\}^{\sigma}_{\text{top-}N_{max}}. \tag{22}$$

Finally, we induce relations among selected nodes with r connects two entities in $\mathcal{V}_{pat}$:

$$\mathcal{R}_{pat} = \{r \in \mathcal{R}_{raw}\}. \tag{23}$$

The final patient knowledge graph is $\mathcal{G}_{pat} = (\mathcal{V}_{pat}, \mathcal{R}_{pat})$. Figure 1 illustrates this process.

3) Graph Embedding: We compute graph embeddings by separately aggregating node and relation representations. This disentangled approach reduces model parameters compared to learnable graph neural networks [29].

Node Representation. For each entity $v \in \mathcal{V}_{pat}$, we obtain its semantic embedding $\mathbf{e}_v \in \mathbb{R}^d$ from the language model. The aggregated node representation is computed via mean pooling:

$$\mathbf{h}_v = \frac{1}{|\mathcal{V}_{pat}|} \sum_{v \in \mathcal{V}_{pat}} \mathbf{e}_v \in \mathbb{R}^d. \tag{24}$$

Relation Representation. For each relation type $r \in \mathcal{R}_{pat}$, we compute its embedding $\mathbf{e}_r \in \mathbb{R}^d$. The aggregated relation representation is:

$$\mathbf{h}_r = \frac{1}{|\mathcal{R}_{pat}|} \sum_{r \in \mathcal{R}_{pat}} \mathbf{e}_r \in \mathbb{R}^d. \tag{25}$$

The final graph representation concatenates L2-normalized node and relation embeddings:

$$\mathbf{h}_G = \left[\frac{\mathbf{h}_v}{\|\mathbf{h}_v\|} \,\middle\|\, \frac{\mathbf{h}_r}{\|\mathbf{h}_r\|} \right] \in \mathbb{R}^{2d}. \tag{26}$$

C. Multimodal Clinical Context Fusion

1) Clinical Text Encoding: Admission notes provide rich unstructured context about patient symptoms, clinical reasoning, and treatment plans. Given note text $\mathcal{N}$, we truncate to maximum length L_{max} tokens and encode:

$$\mathbf{h}_{Note} = \text{LM}(\text{Tokenize}(\mathcal{N})) \in \mathbb{R}^d, \tag{27}$$

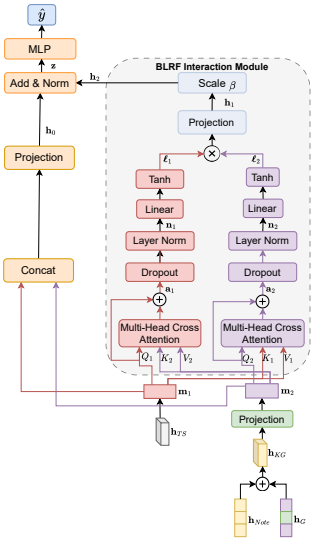

Fig. 2. Bidirectional Low-Rank Fusion (BLRF) architecture.

where the [CLS] token embedding serves as the note representation. To create the unified knowledge graph context, we concatenate graph and note embeddings:

$$\mathbf{h}_{KG} = \text{Concat}(\mathbf{h}_G, \mathbf{h}_{Note}) \in \mathbb{R}^{3d}, \tag{28}$$

where $\mathbf{h}_G$ from Section III-B captures explicit biomedical knowledge (graph structure), while $\mathbf{h}_{Note}$ captures implicit clinical reasoning.

2) Bidirectional Low-Rank Fusion: Multimodal fusion in early-stage ICU faces challenges due to uneven modality reliability and data scarcity: cross-attention-based methods do not inherently enforce balanced modality contributions, simple fusion strategies can be competitive indicating that high-capacity interactions may not be well-supported, and knowledge-enhanced modalities may dominate risking over-reliance. To address these challenges, we propose Bidirectional Low-Rank Fusion (BLRF), a dual-pathway architecture that constrains interaction complexity, preserves unimodal predictive paths through residual connections, and balances expressiveness with parameter efficiency.

Given time-series representation $\mathbf{h}_{TS} \in \mathbb{R}^h$ from Section III-A and knowledge graph representation $\mathbf{h}_{KG} \in \mathbb{R}^{3d}$, we first align both modalities to dimension h. Since $\mathbf{h}_{TS}$ is already in dimension h, we directly use:

$$\mathbf{m}_1 = \mathbf{h}_{TS} \in \mathbb{R}^h. \tag{29}$$

For the knowledge graph, projection is required with a linear layer:

$$\mathbf{m}_2 = \mathbf{W}_{KG}\mathbf{h}_{KG} + \mathbf{b}_{KG} \in \mathbb{R}^h. \tag{30}$$

BLRF employs two parallel pathways (Figure 2). The *base pathway* (left) establishes a stable optimization baseline through direct concatenation:

$$\mathbf{h}_0 = \mathbf{W}_0 \text{Concat}(\mathbf{m}_1, \mathbf{m}_2) + \mathbf{b}_0 \in \mathbb{R}^h, \tag{31}$$

where $\mathbf{W}_0 \in \mathbb{R}^{d \times 2d}$. Following residual learning principles [30], this baseline ensures the model performs no worse than simple concatenation during early training.

The *interaction pathway* (right in Figure 2) learns complex cross-modal dependencies through a two-stage mechanism: bidirectional attention followed by low-rank bilinear pooling.

Stage 1: Bidirectional Cross-Attention. Unlike standard unidirectional cross-attention, we enable mutual information exchange between modalities. We employ multi-head attention [31] with A heads. For brevity, we present the single-head formulation; the full implementation concatenates outputs from all heads.

For one attention head, $\mathbf{m}_1$ queries information from $\mathbf{m}_2$:

$$\mathbf{a}_1 = \text{softmax}\left(\frac{\mathbf{Q}_1(\mathbf{K}_2)^\top}{\sqrt{h/A}}\right)\mathbf{V}_2 \in \mathbb{R}^{h/A}, \tag{32}$$

where $\mathbf{Q}_1 = \mathbf{W}_{Q1}\mathbf{m}_1$, $\mathbf{K}_2 = \mathbf{W}_{K2}\mathbf{m}_2$, $\mathbf{V}_2 = \mathbf{W}_{V2}\mathbf{m}_2$, with $\mathbf{W}_{Q1}$, $\mathbf{W}_{K2}, \mathbf{W}_{V2} \in \mathbb{R}^{(h/A)\times h}$. Symmetrically, $\mathbf{m}_2$ queries $\mathbf{m}_1$:

$$\mathbf{a}_2 = \text{softmax}\left(\frac{\mathbf{Q}_2(\mathbf{K}_1)^\top}{\sqrt{h/A}}\right)\mathbf{V}_1 \in \mathbb{R}^{h/A}, \tag{33}$$

where $\mathbf{Q}_2 = \mathbf{W}_{Q2}\mathbf{m}_2$, $\mathbf{K}_1 = \mathbf{W}_{K1}\mathbf{m}_1$, $\mathbf{V}_1 = \mathbf{W}_{V1}\mathbf{m}_1$, with $\mathbf{W}_{Q2}$, $\mathbf{W}_{K1}, \mathbf{W}_{V1} \in \mathbb{R}^{(h/A)\times h}$.

After concatenating multi-head outputs, we apply residual connections and layer normalization:

$$\mathbf{n}_1 = \text{LayerNorm}(\mathbf{m}_1 + \text{Dropout}(\text{Concat}_{i=1}^{A}\mathbf{a}_1^{(i)})) \in \mathbb{R}^h, \tag{34}$$

$$\mathbf{n}_2 = \text{LayerNorm}(\mathbf{m}_2 + \text{Dropout}(\text{Concat}_{i=1}^{A}\mathbf{a}_2^{(i)})) \in \mathbb{R}^h. \tag{35}$$

At this stage, both $\mathbf{n}_1$ and $\mathbf{n}_2$ contain bidirectional information from both modalities through mutual attention exchange.

Stage 2: Low-Rank Bilinear Pooling. Given the attention-enriched representations $\mathbf{n}_1$ and $\mathbf{n}_2$, we capture second-order cross-modal interactions via low-rank bilinear pooling [23]. We project both vectors to low-rank space $r \ll h$:

$$\boldsymbol{\ell}_1 = \tanh(\mathbf{W}_1^{LR}\mathbf{n}_1 + \mathbf{b}_1^{LR}) \in \mathbb{R}^r, \tag{36}$$

$$\boldsymbol{\ell}_2 = \tanh(\mathbf{W}_2^{LR}\mathbf{n}_2 + \mathbf{b}_2^{LR}) \in \mathbb{R}^r, \tag{37}$$

where $\mathbf{W}_1^{LR}, \mathbf{W}_2^{LR} \in \mathbb{R}^{r\times h}$. The $\tanh$ activation bounds values to $[-1, 1]$ for gradient stability. We compute element-wise product and project back to dimension h:

$$\mathbf{h}_1 = \mathbf{W}_1(\boldsymbol{\ell}_1 \odot \boldsymbol{\ell}_2) + \mathbf{b}_1 \in \mathbb{R}^h, \tag{38}$$

where $\mathbf{W}_1 \in \mathbb{R}^{h\times r}$ and $\odot$ denotes element-wise multiplication. This low-rank factorization reduces parameters from $O(h^2)$ to $O(rh)$, providing essential regularization for limited ICU data.

The final fused representation combines the base pathway $\mathbf{h}_0$ with scaled interaction pathway output $\mathbf{h}_1$:

$$\mathbf{h}_2 = \beta \cdot \mathbf{h}_1, \tag{39}$$

where $\beta \in \mathbb{R}$ is a learnable scaling parameter initialized to zero. We apply additive fusion with normalization:

$$\mathbf{z} = \text{LayerNorm}(\mathbf{h}_0 + \text{Dropout}(\mathbf{h}_2)) \in \mathbb{R}^h. \tag{40}$$

This residual design ensures that when $\beta \approx 0$ during early training, the model defaults to the concatenation baseline $\mathbf{h}_0$, guaranteeing stable optimization.

The fused representation $\mathbf{z}$ is passed through a two-layer feedforward network for binary outcome prediction:

$$\mathbf{h}_{pred} = \text{ReLU}(\mathbf{W}_1^{pred}\mathbf{z} + \mathbf{b}_1^{pred}) \in \mathbb{R}^{h/2}, \tag{41}$$

$$\hat{y} = \sigma(\mathbf{W}_2^{pred}\text{Dropout}(\mathbf{h}_{pred}) + \mathbf{b}_2^{pred}) \in [0, 1], \tag{42}$$

where σ is the sigmoid function and $\mathbf{W}_1^{pred} \in \mathbb{R}^{(h/2)\times h}$, $\mathbf{W}_2^{pred} \in \mathbb{R}^{1\times(h/2)}$. The model is trained end-to-end using binary cross-entropy loss:

$$\mathcal{L}(\hat{y}, y) = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log \hat{y}_i + (1 - y_i)\log(1 - \hat{y}_i)\right]. \tag{43}$$

IV. Experimental Setups

A. Datasets and External Knowledge Graph

MIMIC Benchmarks. We evaluate PIKE on two widely-used critical care datasets: MIMIC-III [19] and MIMIC-IV [20] from Beth Israel Deaconess Medical Center. Following the established EHR benchmark preprocessing pipeline [2, 3], we extract 61 clinical variables comprising vital signs, laboratory values, and one-hot encoded Glasgow Coma Scale subscales. Then, we use a 48-hour observation window and discretize time into 12-hour bins, resulting in $T = 4$ timesteps for the time-series modality. All methods are evaluated on the same discretized inputs to ensure fair comparison. To prevent temporal data leakage, we consolidate observations into 12-hour bins, focusing on the first four bins (48 hours total) of each ICU stay, consistent with early prediction scenarios in critical care settings. Clinical notes are preprocessed following Clinical-Longformer [7]: removing de-identification placeholders, normalizing whitespace, and converting to lowercase.

A key design choice is preserving missing value indicators rather than performing imputation, reflecting the clinical observation that missingness carries diagnostic information [4]. Datasets are split 70:10:20 for training, validation, and test. Table I shows statistics.

TABLE I
DATASET STATISTICS AFTER PREPROCESSING. NUMBERS IN PARENTHESES SHOW PERCENTAGES. OUT. DENOTES IN-HOSPITAL MORTALITY, RE. DENOTES 30-DAY READMISSION

Dataset	Split	Samples	Label$_{Out.}$	Label$_{Re.}$
MIMIC-III	Train	13,482 (70%)	1,523 (11.3%)	2,502 (18.6%)
	Val	1,927 (10%)	218 (11.3%)	357 (18.5%)
	Test	3,853 (20%)	435 (11.3%)	715 (18.6%)
MIMIC-IV	Train	17,339 (70%)	1,995 (11.5%)	2,755 (15.9%)
	Val	2,478 (10%)	285 (11.5%)	394 (15.9%)
	Test	4,955 (20%)	570 (11.5%)	787 (15.9%)

External Knowledge Graph. We use PrimeKG [17], a precision medicine knowledge graph with 17,080 diseases and 4,050,249 relationships across ten biological scales. PrimeKG provides comprehensive ICU-relevant disease coverage, rich textual descriptions (symptoms, risk factors, treatments), and clean multi-relational structure for multi-hop retrieval. Each disease node is pre-encoded using BGE-M3 [32] for semantic grounding.

TABLE II
IN-HOSPITAL MORTALITY AND 30-DAY READMISSION PREDICTION RESULTS ON MIMIC-III AND MIMIC-IV DATASETS. **BOLD** INDICATES THE BEST PERFORMANCE. ALL METRICS ARE MULTIPLIED BY 100 FOR READABILITY.

Methods	Task 1: Mortality Prediction						Task 2: Readmission Prediction					
	MIMIC-III			MIMIC-IV			MIMIC-III			MIMIC-IV		
	AUROC (↑)	AUPRC (↑)	min(+P,Se) (↑)	AUROC (↑)	AUPRC (↑)	min(+P,Se) (↑)	AUROC (↑)	AUPRC (↑)	min(+P,Se) (↑)	AUROC (↑)	AUPRC (↑)	min(+P,Se) (↑)
GRAM [9]	84.11±0.78	50.14±1.43	47.94±1.66	85.87±0.47	48.58±1.26	48.65±1.34	77.13±1.13	50.97±1.58	47.77±1.05	79.20±0.59	45.39±1.32	45.64±1.06
VecoCare [25]	83.43±1.49	47.28±2.68	47.92±2.22	76.93±1.82	46.18±2.76	47.22±2.63	55.37±2.00	55.35±1.72	53.71±1.72	79.17±1.21	50.58±1.48	51.01±1.36
M3Care [8]	83.33±1.24	47.86±2.33	49.96±1.99	76.80±1.55	46.29±2.62	45.38±2.32	88.14±0.78	54.06±2.04	54.30±1.31	83.31±1.79	78.87±1.31	51.95±1.05
KAME [10]	84.00±0.87	50.41±1.68	47.82±1.69	85.88±0.57	50.13±1.57	48.80±1.53	77.33±1.06	51.25±2.02	48.24±1.30	78.59±0.83	43.11±1.24	44.59±0.91
CGL [27]	83.61±0.79	48.54±1.77	46.04±1.79	86.52±0.58	50.94±1.39	48.81±1.48	76.76±0.99	49.60±1.87	48.57±1.28	78.63±0.70	43.22±1.04	44.16±0.92
KerPrint [16]	84.70±0.89	52.19±2.14	49.14±1.81	86.39±0.43	52.07±1.52	49.10±1.66	78.09±0.89	50.92±1.52	48.15±1.33	79.51±0.67	48.48±1.46	47.02±1.19
MedPath [15]	85.01±0.83	49.82±1.57	47.19±1.65	86.94±0.49	51.10±1.45	51.63±1.43	77.21±0.92	48.52±1.45	46.52±1.10	78.56±0.68	43.08±1.25	44.69±1.04
MedRetriever [26]	85.02±0.72	50.93±1.78	47.35±1.82	87.10±0.51	51.94±1.67	51.81±1.53	77.06±1.21	49.74±1.75	47.71±1.60	78.83±0.75	43.69±1.27	44.46±0.99
GraphCare [11]	85.25±0.83	51.10±1.84	47.47±1.78	87.22±0.47	54.73±1.67	52.70±1.78	77.98±0.98	50.14±1.52	47.64±1.35	78.86±0.75	43.96±1.05	44.59±0.82
EMERGE [2]	85.65±0.77	53.06±1.93	49.66±1.97	87.58±0.51	56.76±1.52	53.40±1.58	78.34±1.09	51.63±1.64	48.70±1.37	80.28±0.67	51.86±1.43	48.96±1.13
PIKE (Ours)	**90.85±0.68**	**64.09±2.18**	**58.72±2.31**	**93.86±0.47**	**74.21±1.71**	**66.67±2.06**	**83.37±0.83**	**59.58±1.92**	**55.23±1.83**	**86.86±0.71**	**65.50±1.63**	**58.83±1.61**

Evaluation Metrics. We adopt three evaluation metrics for clinical binary classification: AUROC as primary metric for handling class imbalance [33]; AUPRC for minority class performance and min(+P, Se) as balanced precision-sensitivity composite [34].

B. Baseline Models

We evaluate PIKE through three sets of experiments corresponding to our main contributions:

- **EHR Prediction Models.** We include multimodal EHR baselines (M3Care [8], VecoCare [25]), knowledge-enhanced methods (GRAM [9], KAME [10], CGL [27], KerPrint [16], MedPath [15], MedRetriever [26]), and LLM-facilitated models (GraphCare [11], EMERGE [2]).
- **Hybrid Set-Sequence Encoder Replacement.** To validate the Hybrid Set-Sequence Encoder, we replace it entirely with alternative temporal encoders: RNN [5], LSTM [12], GRU-D [4], Transformer [31], STraTS [6], and Raindrop [24].
- **Multimodal Fusion Methods.** To evaluate BLRF, we replace it with: Concat [35], Bidirectional [31] (bidirectional cross-attention), Cross Attention [31] (unidirectional cross-attention), Gated [18], and MAG [36].

C. Implementation Details

All experiments are conducted on 2 NVIDIA RTX 4090 GPUs (24GB VRAM) with CUDA 12.1. We implement PIKE in Python 3.10 with PyTorch 2.2.0 [37]. We use AdamW optimizer [38] with weight decay 0.01, batch size 64, and train for a maximum of 50 epochs with early stopping based on validation AUPRC. For text encoding, we use BGE-M3 [32] ($d = 1024$) to encode KG entity names and relation descriptions, and Clinical-Longformer [7] ($d = 768$) to encode clinical notes. The final knowledge representation concatenates entity, relation, and note embeddings, yielding dimension $2 \times 1024 + 768 = 2816$. Based on sensitivity analysis (Section V-C1 and Section V-C2), we set hidden dimension $h = 16$, learning rate $\eta = 10^{-4}$, dropout rate 0.1, tensor rank $r = 32$, 4 attention heads, graph expansion hops $H = 2$, BGE similarity threshold $\theta = 0.5$, and maximum subgraph size $N_{max} = 50$ nodes. We report test performance as mean $\pm$ std via bootstrapping with 1000 resamples [39].

V. EXPERIMENTAL RESULTS AND ANALYSIS

A. Main Results

Table II reports mortality and readmission prediction performance. PIKE achieves 90.85% AUROC (MIMIC-III) and 93.86% (MIMIC-IV) for mortality, securing substantial AUPRC absolute improvements of 11.03% and 17.45% over the best baseline (EMERGE). Similarly, for readmission, PIKE outperforms baselines with 83.37% and 86.86% AUROC, yielding AUPRC improvements of 7.95% and 13.64%.

This performance differential may be attributed to the distinct handling of data sparsity in the critical 48-hour window. Generative baselines such as EMERGE depend on RAG retrieval from clinical notes; however, early-stage notes are brief and sparse, which may cause retrieval of noise rather than signal. In contrast, PIKE employs two deterministic mechanisms: (1) structured knowledge expansion that grounds patient states in PrimeKG regardless of note length, and (2) the Sequence Encoder that initializes decay rates with physiological half-lives. Ablation studies (Section V-B1 and Section V-B2) confirm that removing either mechanism results in substantial performance drops, demonstrating their effectiveness for robust representation learning when explicit observations are minimal.

B. Ablation Studies

1) Modality Contribution Analysis: As defined in Section III, PIKE processes three input modalities: Time-Series ($\mathbf{X}_{TS}$), Clinical Notes ($\mathcal{X}_{Note}$), and Knowledge Graph ($\mathcal{X}_G$). Table III evaluates the contribution of each modality. Among single modalities, KG-Only (which uses diagnosis codes with PrimeKG expansion) achieves the highest AUROC (88.02% MIMIC-III, 90.34% MIMIC-IV for mortality), followed by TS-Only (85.12%, 88.43%), while Notes-Only yields the lowest performance (73.36%, 77.63%). The combination TS + KG achieves the strongest two-modality performance (90.65% MIMIC-III, 93.50% MIMIC-IV), and adding Clinical Notes (Full PIKE) yields only marginal gain (<0.4% AUROC). These results suggest that the Hybrid Set-Sequence Encoder and Graph Embedding module effectively capture complementary temporal and semantic signals, while the current note encoder uses a simple pooling strategy that limits its contribution.

TABLE III
ABLATION STUDY RESULTS ON MODALITY CONTRIBUTION, TIME-SERIES ENCODER COMPARISON, AND FUSION ARCHITECTURE. **BOLD** INDICATES THE BEST PERFORMANCE, UNDERLINE INDICATES SECOND-BEST. ALL METRICS ARE MULTIPLIED BY 100 FOR READABILITY.

Methods	Task 1: Mortality Prediction						Task 2: Readmission Prediction					
	MIMIC-III			MIMIC-IV			MIMIC-III			MIMIC-IV		
	AUROC (↑)	AUPRC (↑)	min(+P,Se) (↑)	AUROC (↑)	AUPRC (↑)	min(+P,Se) (↑)	AUROC (↑)	AUPRC (↑)	min(+P,Se) (↑)	AUROC (↑)	AUPRC (↑)	min(+P,Se) (↑)
Modality Contribution												
TS-Only	85.12±1.01	49.60±2.48	47.82±2.12	88.43±0.75	59.05±2.06	55.94±1.99	77.46±1.04	50.86±1.89	47.77±1.71	79.47±0.95	53.50±1.79	51.21±1.66
KG-Only	88.02±0.82	51.12±2.62	53.33±2.44	90.34±0.62	61.72±2.04	57.79±2.15	79.88±0.93	51.40±2.02	49.93±1.81	84.36±0.75	57.73±1.86	53.76±1.70
Notes-Only	73.36±1.24	25.66±1.76	31.49±1.74	77.63±0.97	32.55±1.98	36.08±1.18	69.71±1.08	32.18±1.63	36.22±0.72	72.75±0.96	34.22±1.69	36.49±1.05
TS+KG	90.65±0.75	63.80±2.30	58.50±2.25	93.50±0.49	73.50±1.75	66.40±2.05	83.10±0.85	59.20±1.90	55.00±1.80	86.20±0.75	65.10±1.65	58.00±1.70
KG+Notes	88.02±0.79	51.65±2.66	52.42±2.32	90.71±0.60	63.64±2.00	58.60±2.09	79.23±0.94	51.14±2.03	50.21±1.90	84.41±0.75	58.08±1.79	54.00±1.80
TS+Notes	84.74±0.92	47.98±2.41	47.48±2.13	88.34±0.78	59.06±2.05	53.68±1.83	78.14±0.98	49.57±1.93	48.54±1.80	80.99±0.89	54.63±1.77	50.70±1.52
Hybrid Set-Sequence Encoder Replacement												
RNN [5]	87.50±0.67	58.20±2.30	52.71±2.21	90.00±0.55	64.79±1.82	61.05±2.04	79.50±0.86	53.99±1.94	51.15±1.84	82.50±0.75	56.98±1.73	52.89±1.73
LSTM [12]	88.00±0.71	59.32±2.38	54.27±2.34	90.50±0.52	66.31±1.84	62.04±2.11	80.00±0.85	54.19±1.96	52.69±1.84	83.00±0.74	57.73±1.76	53.04±1.80
GRU-D [4]	88.50±0.70	59.41±2.35	55.85±2.34	91.00±0.55	68.47±1.94	64.26±2.02	80.50±0.87	55.39±1.93	53.83±1.88	83.00±0.75	58.93±1.72	54.53±1.73
Transformer [31]	89.00±0.64	61.57±2.26	56.44±2.33	90.50±0.49	68.84±1.79	63.73±2.07	81.50±0.85	58.21±1.85	54.24±1.82	83.50±0.74	60.12±1.65	54.95±1.83
STraTS [6]	87.80±0.75	51.04±2.60	51.26±2.31	89.50±0.80	54.42±2.62	50.80±2.27	81.00±0.87	55.48±1.97	51.68±1.90	81.50±0.94	51.72±1.97	49.37±1.68
Raindrop [24]	89.50±0.65	61.83±2.34	56.42±2.33	92.00±0.50	70.23±1.90	64.91±2.01	82.00±0.89	57.24±1.94	53.01±1.87	85.10±0.74	62.44±1.73	57.81±1.75
Fusion Architecture Comparison												
Concat [35]	90.80±0.60	64.03±2.12	58.61±2.34	92.50±0.54	71.92±1.75	66.19±2.08	82.50±0.84	59.10±1.81	54.22±1.87	85.00±0.71	62.21±1.79	56.69±1.74
Bidirectional [31]	90.60±0.61	63.26±2.19	58.27±2.07	93.00±0.50	72.72±1.80	66.55±2.05	82.80±0.83	59.10±1.82	54.90±1.70	85.50±0.71	63.69±1.67	56.81±1.80
Cross Attention [31]	90.20±0.63	63.75±2.22	57.41±2.15	92.80±0.54	72.92±1.75	66.19±2.08	82.60±0.84	58.10±1.81	54.22±1.87	84.80±0.71	62.31±1.67	56.61±1.74
Gated [18]	90.00±0.65	62.70±2.21	57.45±2.32	92.60±0.47	71.79±1.81	65.84±2.06	81.00±0.86	56.19±1.94	51.19±1.81	85.20±0.73	63.32±1.65	56.91±1.72
MAG [36]	90.25±0.61	63.03±2.12	58.10±2.07	92.50±0.50	71.72±1.80	65.55±2.05	82.50±0.83	58.08±1.82	54.89±1.70	84.50±0.71	62.59±1.67	55.61±1.80
PIKE (Ours)	**90.85±0.68**	**64.09±2.18**	**58.72±2.31**	**93.86±0.47**	**74.21±1.71**	**66.67±2.06**	**83.37±0.83**	**59.58±1.92**	**55.23±1.83**	**86.86±0.71**	**65.50±1.63**	**58.83±1.61**

2) Hybrid Set-Sequence Encoder Replacement: We adapt Transformer [31] to multivariate time series using learned positional encodings and feature-wise embeddings. All replacement encoders use the same hidden dimension to ensure fair comparison. Table III reports the results. PIKE achieves 90.85% AUROC (MIMIC-III) and 93.86% (MIMIC-IV) for mortality, outperforming the second-best alternative encoder Raindrop by 1.35% and 1.86%. Attention-based models (Transformer, STraTS) may be less effective when $T = 4$ timesteps provide limited context to learn inter-variable dependencies. Moreover, discretization to 12-hour bins collapses fine-grained timing information, reducing the advantage of STraTS [6] which is designed for irregular sampling. In contrast, our Sequence Encoder initializes decay rates with physiological half-lives, enabling variable-specific temporal dynamics even with limited observations. This suggests that the Hybrid Set-Sequence Encodesr is well-suited for short-sequence, sparse ICU data.

3) Fusion Architecture Comparison: PIKE achieves 90.85% and 93.86% AUROC for mortality, outperforming Bidirectional by 0.25% (MIMIC-III) and 0.86% (MIMIC-IV). For readmission, PIKE achieves 86.86% AUROC, surpassing Bidirectional by 1.36%. Notably, simple Concat fusion achieves competitive performance, which supports our claim that high-capacity fusion mechanisms may be brittle under short-sequence, sparse constraints. BLRF improves upon Concat by adding controlled cross-modal interactions through the low-rank bilinear pooling pathway while preserving stable gradient flow via the base concatenation pathway. These results suggest that BLRF effectively balances expressiveness and robustness for multimodal fusion under data scarcity.

C. Further Analysis

1) Hyperparameter Sensitivity: We analyze sensitivity to two key hyperparameters on MIMIC-IV (Figure 3). For hidden dimension $h \in \{16, \ldots, 512\}$, the best performance is achieved at $h = 128$. For learning rate $\eta \in \{10^{-1}, \ldots, 10^{-6}\}$, optimal performance occurs at $\eta = 10^{-3}$. We select $h = 16$, $\eta = 10^{-4}$ as the final configuration.

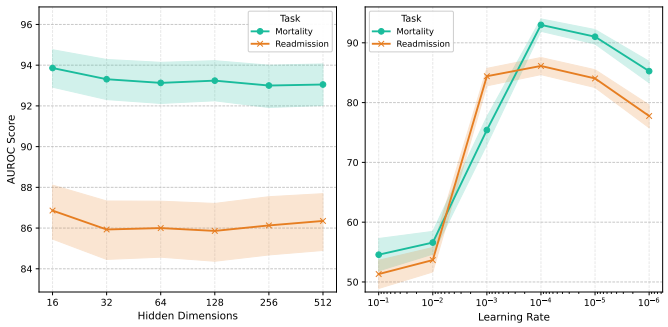

Fig. 3. Hyperparameter sensitivity analysis on MIMIC-IV

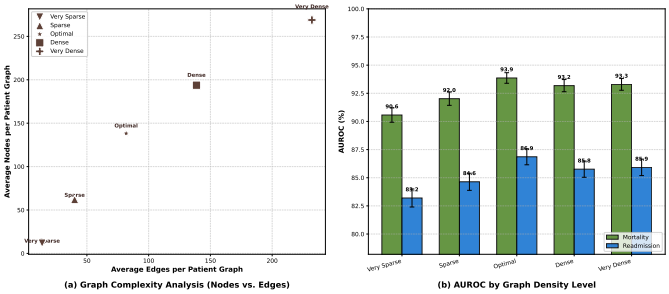

Fig. 4. Graph density analysis on MIMIC-IV.

2) Graph Density Analysis: Figure 4 investigates the impact of subgraph density controlled by maximum nodes N_{max} and expansion hops H. We define five density levels: Very Sparse ($N_{max} = 10, H = 1, \theta = 0.7$), Sparse ($N_{max} = 30, H = 1, \theta = 0.6$), Optimal ($N_{max} = 50, H = 2, \theta = 0.5$), Dense ($N_{max} = 70, H = 3, \theta = 0.4$), and Very Dense ($N_{max} = 100, H = 5, \theta = 0.3$). Increasing density beyond this point introduces irrelevant nodes (noise) and dilutes the semantic signal, degrading AUROC. This confirms that our

node selection module ($N_{max} = 50$) effectively balances semantic coverage and noise reduction, refuting the assumption that denser graphs yield better representations.

VI. Conclusion

We proposed PIKE, a framework for clinical outcome prediction under *short-sequence, sparse* early-stage ICU conditions. PIKE addresses three core challenges through: (1) a physiology-informed Hybrid Set-Sequence Encoder incorporating variable-name embeddings and half-life-initialized decay, (2) deterministic patient knowledge graph construction via controlled multi-hop expansion in PrimeKG with evidence-based relation scoring, and (3) Bidirectional Low-Rank Fusion (BLRF), a parameter-efficient dual-pathway architecture for robust cross-modal integration. Experiments on MIMIC-III and MIMIC-IV demonstrate that PIKE achieves absolute AUROC improvements of 11.03% and 17.45% respectively for mortality prediction compared to state-of-the-art baselines. These results demonstrate that incorporating physiological priors, curated biomedical knowledge, and parameter-efficient fusion is effective for short-sequence, sparse clinical prediction. Future work may explore extending this approach to other early-stage clinical tasks (e.g., decompensation, length-of-stay prediction) and evaluating robustness under cross-hospital distribution shift.

References

[1] J. Wang, J. Luo, M. Ye, X. Wang, Y. Zhong, A. Chang, G. Huang, Z. Yin, C. Xiao, J. Sun *et al.*, "Recent advances in predictive modeling with electronic health records," in *IJCAI: proceedings of the conference*, vol. 2024, 2024, p. 8272.

[2] Y. Zhu, C. Ren, Z. Wang, X. Zheng, S. Xie, J. Feng, X. Zhu, Z. Li, L. Ma, and C. Pan, "Emerge: Enhancing multimodal electronic health records predictive modeling with retrieval-augmented generation," in *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. New York, NY, USA: Association for Computing Machinery, 2024, pp. 3549–3559.

[3] H. Harutyunyan, H. Khachatrian, D. C. Kale, G. Ver Steeg, and A. Galstyan, "Multitask learning and benchmarking with clinical time series data," *Scientific data*, vol. 6, no. 1, p. 96, 2019.

[4] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, "Recurrent neural networks for multivariate time series with missing values," *Scientific reports*, vol. 8, no. 1, p. 6085, 2018.

[5] A. Sherstinsky, "Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network," *Physica D: Nonlinear Phenomena*, vol. 404, p. 132306, 2020.

[6] S. N. Shukla and B. Marlin, "Multi-time attention networks for irregularly sampled time series," in *International Conference on Learning Representations*, 2021.

[7] Y. Li, R. M. Wehbe, F. S. Ahmad, H. Wang, and Y. Luo, "A comparative study of pretrained language models for long clinical text," *Journal of the American Medical Informatics Association*, vol. 30, no. 2, pp. 340–347, 2023.

[8] C. Zhang, X. Chu, L. Ma, Y. Zhu, Y. Wang, J. Wang, and J. Zhao, "M3care: Learning with missing modalities in multimodal healthcare data," in *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, 2022, pp. 2418–2428.

[9] E. Choi, M. T. Bahadori, L. Song, W. F. Stewart, and J. Sun, "Gram: graph-based attention model for healthcare representation learning," in *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 2017, pp. 787–795.

[10] F. Ma, Q. You, H. Xiao, R. Chitta, J. Zhou, and J. Gao, "Kame: Knowledge-based attention model for diagnosis prediction in healthcare," in *Proceedings of the 27th ACM international conference on information and knowledge management*, 2018, pp. 743–752.

[11] P. Jiang, C. Xiao, A. Cross, and J. Sun, "Graphcare: Enhancing healthcare predictions with personalized knowledge graphs," in *The Twelfth International Conference on Learning Representations*, 2024.

[12] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[13] M. Horn, M. Moor, C. Bock, B. Rieck, and K. Borgwardt, "Set functions for time series," in *International Conference on Machine Learning*. PMLR, 2020, pp. 4353–4363.

[14] M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Poczos, R. R. Salakhutdinov, and A. J. Smola, "Deep sets," *Advances in neural information processing systems*, vol. 30, 2017.

[15] M. Ye, S. Cui, Y. Wang, J. Luo, C. Xiao, and F. Ma, "Medpath: Augmenting health risk prediction via medical knowledge paths," in *Proceedings of the Web Conference 2021*, 2021, pp. 1397–1409.

[16] K. Yang, Y. Xu, P. Zou, H. Ding, J. Zhao, Y. Wang, and B. Xie, "Kerprint: local-global knowledge graph enhanced diagnosis prediction for retrospective and prospective interpretations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 4, 2023, pp. 5357–5365.

[17] P. Chandak, K. Huang, and M. Zitnik, "Building a knowledge graph to enable precision medicine," *Scientific Data*, vol. 10, no. 1, p. 67, 2023.

[18] J. Arevalo, T. Solorio, M. Montes-y Gómez, and F. A. González, "Gated multimodal units for information fusion," in *International Conference on Learning Representations Workshop*, 2017.

[19] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.

[20] A. E. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T. J. Pollard, S. Hao, B. Moody, B. Gow *et al.*, "Mimic-iv, a freely accessible electronic health record dataset," *Scientific data*, vol. 10, no. 1, p. 1, 2023.

[21] A. C. Guyton and J. E. Hall, *Guyton and Hall textbook of medical physiology*. Elsevier, 2011.

[22] G. H. Guyatt, A. D. Oxman, G. E. Vist, R. Kunz, Y. Falck-Ytter, P. Alonso-Coello, and H. J. Schünemann, "Grade: an emerging consensus on rating quality of evidence and strength of recommendations," *Bmj*, vol. 336, no. 7650, pp. 924–926, 2008.

[23] Z. Yu, J. Yu, J. Fan, and D. Tao, "Multi-modal factorized bilinear pooling with co-attention learning for visual question answering," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1821–1830.

[24] X. Zhang, M. Zeman, T. Tsiligkaridis, and M. Zitnik, "Graph-guided network for irregularly sampled multivariate time series," in *International Conference on Learning Representations*, 2022.

[25] Y. Xu, K. Yang, C. Zhang, P. Zou, Z. Wang, H. Ding, J. Zhao, Y. Wang, and B. Xie, "Vecocare: visit sequences-clinical notes joint learning for diagnosis prediction in healthcare data," in *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, ser. IJCAI '23, 2023.

[26] M. Ye, S. Cui, Y. Wang, J. Luo, C. Xiao, and F. Ma, "Medretriever: Target-driven interpretable health risk prediction via retrieving unstructured medical text," in *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 2021, pp. 2414–2423.

[27] C. Lu, C. K. Reddy, P. Chakraborty, S. Kleinberg, and Y. Ning, "Collaborative graph learning with auxiliary text for temporal event prediction in healthcare," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, 2021, pp. 3529–3535.

[28] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[29] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *International Conference on Learning Representations*, 2018.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[32] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, and Z. Liu, "M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation," in *Findings of the Association for Computational Linguistics ACL 2024*. Bangkok, Thailand: Association for Computational Linguistics, 2024, pp. 2318–2335.

[33] M. McDermott, H. Zhang, L. Hansen, G. Angelotti, and J. Gallifant, "A closer look at auroc and auprc under class imbalance," *Advances in Neural Information Processing Systems*, vol. 37, pp. 44 102–44 163, 2024.

[34] X. Ma, Y. Wang, X. Chu, L. Ma, W. Tang, J. Zhao, Y. Yuan, and G. Wang, "Patient health representation learning via correlational sparse prior of medical features," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 11, pp. 11 769–11 783, 2022.

[35] D. Kiela, E. Grave, A. Joulin, and T. Mikolov, "Efficient large-scale multi-modal classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[36] W. Rahman, M. K. Hasan, S. Lee, A. B. Zadeh, C. Mao, L.-P. Morency, and E. Hoque, "Integrating multimodal information in large pretrained transformers," in *Proceedings of the 58th annual meeting of the association for computational linguistics*, 2020, pp. 2359–2369.

[37] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[38] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019.

[39] F. Ma, R. Chitta, J. Zhou, Q. You, T. Sun, and J. Gao, "Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks," in *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 2017, pp. 1903–1911.

TABLE V
PHYSIOLOGICAL HALF-LIFE CATEGORIES AND CLINICAL RATIONALE

Category	τ_{prior}	**Clinical Rationale**
Vital Signs	0.5h	Rapid physiological changes reflecting immediate clinical state
Blood Gas	1.5h	Respiratory compensation and acid-base dynamics
Metabolic	2.0h	Glucose homeostasis and metabolic regulation
Kidney Markers	4.0h	Reflects slower renal clearance kinetics (e.g., creatinine changes more gradually)
Electrolytes	6.0h	Renal regulation and cellular buffering mechanisms
Hematology	24h	Slow production/destruction cycles of blood cells
Cardiac Markers	48h	Troponin release kinetics reflecting myocardial injury
Liver Enzymes	48h	Long biological half-life reflecting hepatocyte turnover

TABLE VI
REPRESENTATIVE VARIABLE MAPPING TO HALF-LIFE CATEGORIES

Category	**Count**	**Example Variables**
Vital Signs	11	Heart Rate, Systolic/Diastolic/Mean BP, Respiratory Rate, SpO2, Temperature, GCS components
Blood Gas	5	pH, pCO2, pO2, Base Excess, FiO2
Metabolic	3	Glucose, Lactate, Ammonia
Kidney Markers	4	Creatinine, BUN, GFR, Uric Acid
Electrolytes	10	Sodium, Potassium, Chloride, Bicarbonate, Calcium, Magnesium, Phosphorus
Hematology	14	WBC, RBC, Hemoglobin, Hematocrit, Platelets, MCV, MCH, PT, PTT, INR
Cardiac Markers	1	Troponin
Liver Enzymes	13	AST, ALT, ALP, GGT, Bilirubin, Albumin, Total Protein

APPENDIX

The scoring function $\omega : \mathcal{R}_{KG} \rightarrow \{1, 2, 3\}$ defined in Section III-B assigns weights to prioritize relations with direct clinical actionability. We adopt a heuristic mapping inspired by the GRADE evidence hierarchy [22], which ranks clinical evidence quality. While GRADE is designed for evaluating clinical recommendations, we adapt it as an inductive bias to prioritize interventional relations (indication, contraindication) over mechanistic pathways and associative co-occurrences. This is not a claim that PrimeKG relations correspond to GRADE evidence levels, but rather a principled prior for relation importance in clinical prediction tasks.

We categorize PrimeKG relation types into three groups based on their clinical utility: (1) *Interventional* relations directly inform therapeutic decisions and safety warnings, (2) *Mechanistic* relations describe molecular targets and phenotypic effects, and (3) *Associative* relations provide biological context and co-occurrence patterns. Table IV provides the mapping used in our experiments; relation types not listed are merged into the associative category.

These scores are used in the graph expansion phase (Section III-B) to bias neighbor sampling toward high-priority relations when the subgraph size constraint N_{max} is reached, ensuring that interventional knowledge is preferentially retained.

TABLE IV
RELATION SCORING FUNCTION $\omega(r)$ FOR PRIMEKG RELATION TYPES USED IN OUR EXPERIMENTS. SCORING IS INSPIRED BY THE GRADE EVIDENCE HIERARCHY [22], PRIORITIZING INTERVENTIONAL EVIDENCE OVER MECHANISTIC AND ASSOCIATIVE RELATIONSHIPS.

Category	$\omega(r)$	**Relations** $r \in \mathcal{R}_{KG}$
$\mathcal{R}_{interv}$ *(Interventional)*	3	`indication`, `contraindication`, `off-label use` Direct therapeutic decisions and safety warnings
$\mathcal{R}_{mech}$ *(Mechanistic)*	2	`drug_targets`, `disease_mechanism`, `drug_effect`, `exposure_disease` Molecular pathways and phenotypic outcomes
$\mathcal{R}_{assoc}$ *(Associative)*	1	`phenotype`, `protein_interaction`, `disease_association`, `pathway` Biological context and co-occurrences

The Sequence Encoder (Section III-A) initializes decay rates τ_f for each clinical variable f using physiologically motivated priors. We define τ_{prior} as an *effective decay time constant* that represents the characteristic persistence of clinical measurements, not necessarily the strict pharmacokinetic half-life. This initialization enables the model to capture variable-specific temporal dynamics from the first training step, rather than learning these dynamics from scratch. For example, vital signs (heart rate, blood pressure) change rapidly and require short time constants (0.5h), while slow biomarkers reflect gradual physiological processes and require longer time constants (48h).

Table V details the physiological half-life categories derived from clinical textbooks [21] and clinical reasoning. Table VI shows the mapping of clinical variables to categories; counts correspond to the 61-variable feature set used in our preprocessing (Section IV). The decay function is defined as $\gamma_f(\Delta t) = \exp(-\Delta t/\tau_f)$, where τ_f is initialized with τ_{prior} and then refined via learnable residuals δ_f (initialized to zero) during training, as described in Equation (6) of Section III-A. Ablation studies (Section V-B2) demonstrate that this physiological initialization outperforms random initialization, particularly under short-sequence constraints.