GUIDELINE-ANCHORED MULTIMODAL REPRESENTATION LEARNING WITH PATIENT-SPECIFIC PROTOTYPE MINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Clinical outcome prediction from electronic health records (EHRs) requires relating physiological changes to observations documented in clinical notes. Noise, missingness, and differences in how these modalities are recorded complicate this task. Although multimodal models learn associations useful for prediction, those associations need not correspond to clinically meaningful patterns. Clinical guidelines organize diagnostic criteria and monitoring indicators, yet their use as semantic anchors for patient representation learning remains limited. We propose **G**uideline-**A**nchored **M**ultimodal Fusion with **P**rototype Mining (GAMP), which uses guideline statements to guide the learning of patient-specific prototypes. Shared queries initialized from guideline knowledge extract patterns from each patient's physiological time series and notes. Attention over individual guideline statements calibrates these patterns, and a guideline alignment objective supplies a common semantic reference. Residual connections and reconstruction retain information from the patient's record. A dual-branch decoder then combines guideline and patient queries for prediction. Experiments on MIMIC-III show improved performance over the evaluated baselines for 48-hour in-hospital mortality prediction and 24-hour phenotype classification. Ablations examine the contributions of guideline initialization and calibration, while clinical case analysis illustrates the concepts emphasized by the learned representations.

1 INTRODUCTION

EHRs have become a cornerstone of modern healthcare systems, capturing rich, longitudinal patient information across multiple data modalities Ye et al. (2024); Niu et al. (2025). The structured component of EHRs consists of multivariate physiological time series—vital signs, laboratory results, and medication logs—which encode dynamic, fine-grained physiological changes over a patient's stay. Complementing these quantitative measurements, unstructured clinical notes authored by physicians and nurses provide rich contextual observations, diagnostic reasoning, and narrative descriptions not easily reducible to numerical values. Integrating these heterogeneous modalities can improve clinical prediction by combining complementary physiological and contextual information. Chen & Liu (2025); Wu et al. (2025).

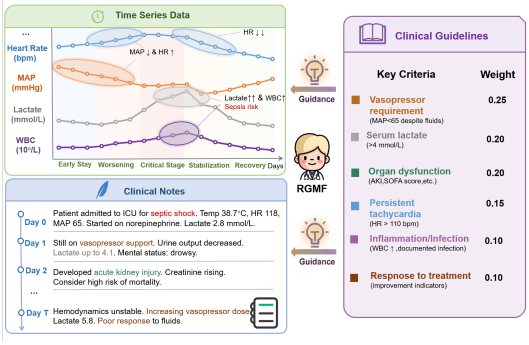


Figure 1: Clinical guidelines (right) serve as a global knowledge anchor to guide the extraction and alignment of distinct temporal patterns, or prototypes, from both continuous time-series physiological data (top left) and unstructured clinical notes (bottom left).

Deep learning has significantly advanced multimodal EHR analysis. Early approaches relied on simple fusion strategies, such as feature concatenation or prediction ensembles, which provided limited modeling of cross-modal interactions

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.

Yang et al. (2021). Recent transformer-based architectures have enabled more sophisticated cross-attention and adapter-based fusion, achieving strong performance on several EHR benchmarksYe et al. (2024); Nash et al. (2024). Nevertheless, two core challenges still remain. Current approaches mostly learn superficial statistical associations instead of clinically relevant relationships, and fail to capture the pathophysiological mechanisms that underpin clinical reasoningWu et al. (2025); Niu et al. (2025). Moreover, the majority of these models operate as black boxes, offering limited interpretability or evidence traceability, critically impeding their adoption in safety-critical clinical settings Rodis et al. (2024).

Prototype-based learning has emerged as a promising paradigm for enhancing interpretability Snell et al. (2017). By identifying representative prototype instances and constraining their distances from input representations, prototype-based methods enable human inspection of learned decision patterns and have demonstrated effectiveness across healthcare tasks Cai et al. (2025); Wang et al. (2024). Recent efforts have extended this paradigm to multimodal EHR analysis by jointly extracting cross-modal prototypes from time-series and clinical notes Wang et al. (2025). However, a critical gap persists: without a clinically grounded semantic reference anchoring the prototype space, learned representations tend to drift toward statistically salient but clinically irrelevant patterns, rendering predictions unreliable for evidence-based decision support Pellicer et al. (2025). Furthermore, existing methods predominantly focus on input-side augmentation and do not systematically impose knowledge-level constraints during representation learning, failing to explicitly align both temporal and textual EHR features with evidence-based clinical guidelines Lu et al. (2024); Zhao et al. (2025). To address these gaps, we anchor prototype learning and final prediction to clinical guidelines, which serve as a global knowledge reference, as illustrated in Figure 1: shared temporal patterns, or prototypes, are extracted from physiological time-series and clinical notes under the guidance of guideline statements.

In this paper, we propose the **G**uideline-**A**nchored **M**ultimodal **F**usion with **P**rototype Mining (GAMP) framework. In summary, our contributions are three-fold:

- We propose a guideline-anchored prototype alignment module that initializes shared prototypes from clinical guideline embeddings and aligns them with guideline knowledge through a dedicated latent-space loss, improving noise filtering and interpretability.
- We design a dual-branch decoder that separately queries the patient representation using guideline and patient-specific information.The guideline branch enables traceable adherence to standard clinical criteria, while the patient-specific branch captures subtle off-norm physiological variations.
- We evaluate GAMP on MIMIC-III for 48-hour in-hospital mortality prediction and 24-hour phenotype classification, where it consistently outperforms the evaluated baselines.

2 Related Work

2.1 Multimodal Learning in Healthcare

Clinical time-series models address missing values, irregular sampling, and variation in the duration of patient records. GRU-D, IP-Net, SeFT, mTAND, and DGM-O approach these problems through mechanisms for handling missingness, interpolation, set-based representations, or temporal modeling Che et al. (2018); Shukla & Marlin (2019); Horn et al. (2020); Shukla & Marlin (2021); Wu et al. (2021). These methods model physiological trajectories but do not use the contextual information available in clinical notes. Multimodal approaches combine these trajectories with text and, in some settings, tabular or imaging data. Temporal-clinical note networks, UTDE, DAFT, MedFuse, and DrFuse explore temporal fusion, feature modulation, and disentangled representations to combine information across modalities Yang et al. (2021); Zhang et al. (2023); Pölsterl et al. (2021); Hayat et al. (2022); Yao et al. (2024). More recent Transformer and prompt-based methods improve interactions between time series and text Ye et al. (2024); Niu et al. (2025). This line of work establishes how complementary observations can be combined. GAMP addresses the clinical reference used during that combination by introducing guideline statements into the learning of intermediate patient representations.

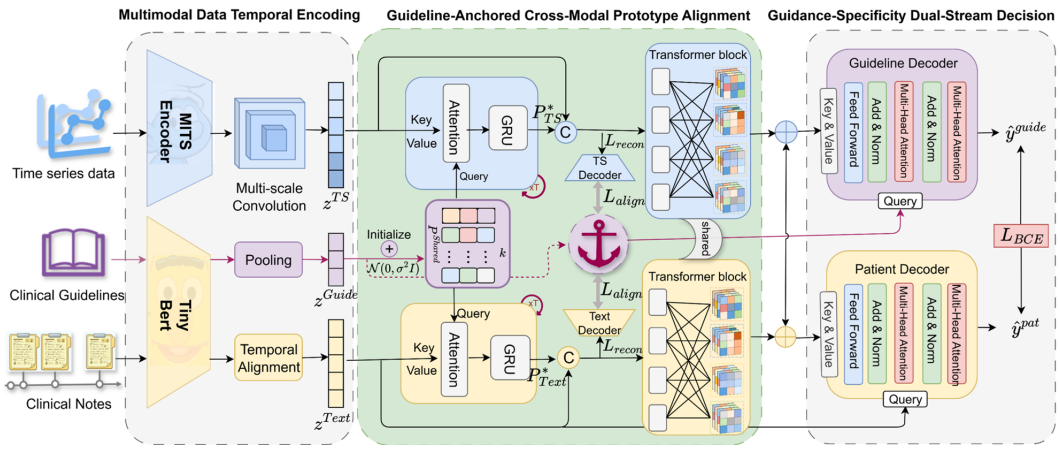


Figure 2: GAMP initializes shared queries from clinical guidelines and mines prototypes from each patient's time series and notes. Statement-level attention calibrates the prototypes, while reconstruction retains information from the record. A dual-branch decoder queries the resulting representation with guidelines and patient notes.

2.2 Prototype-based Pattern Learning

Prototype-based methods represent data through a set of characteristic patterns in a learned space. They have been applied to few-shot learning, unsupervised representation learning, and EHR prediction Snell et al. (2017); Chen et al. (2023); Gao et al. (2019); Li et al. (2020); Cai et al. (2025). In healthcare, ProtoEHR models similarities at code, visit, and patient levels, while ProtoTS uses hierarchical temporal prototypes for forecasting Cai et al. (2025); Peng et al. (2025). For multimodal EHRs, CTPD uses slot attention to extract shared prototypes from irregular time series and clinical notes Wang et al. (2025). These approaches make recurring patterns accessible through a compact representation, but the clinical meaning of an observation-driven prototype still depends on what the model learns from the data. GAMP introduces a guideline reference at initialization and calibration. Its prototypes remain functions of individual patient records, with guideline statements supplying prior knowledge about the clinical concepts those records may express.

2.3 Medical Knowledge Injection

Medical knowledge can enter a learning system through pretrained parameters, explicit concept relations, or retrieved documents. BioBERT and PubMedBERT encode biomedical corpora into language representations Lee et al. (2020); Gu et al. (2021). Graph-based approaches introduce relations among medical concepts through graph-enhanced pretraining or reasoning Yasunaga et al. (2022); Zhao et al. (2025). Retrieval-based methods select medical documents to provide context for downstream tasks Lu et al. (2024); Zhao et al. (2025). These approaches differ in where knowledge enters the model and how it is used. Our focus is on its role during discriminative patient representation learning. Guideline embeddings initialize prototype queries, provide statement-level attention keys and values, and define the alignment target. Clinical knowledge thus participates in forming the features used for prediction, alongside the information extracted from a patient's measurements and notes.

3 Method

As illustrated in Figure 2, GAMP consists of three cascaded modules: **A).** a cross-modal temporal encoding module that projects heterogeneous EHR data into a unified temporal-semantic space, **B).** a guideline-anchored prototype alignment module that injects medical knowledge into the feature space, and **C).** an evidence-specificity dual-stream decision module that produces both guideline-conformant and patient-specific predictions.

3.1 PROBLEM FORMULATION

Let $\mathcal{D} = \{(\mathbf{X}_i^{TS}, \mathbf{X}_i^{Txt}, \mathbf{y}_i)\}_{i=1}^N$ denote a training set of N patients. Each record contains a multivariate time series $\mathbf{X}^{TS} = \{\mathbf{x}_1^{TS}, \ldots, \mathbf{x}_T^{TS}\} \in \mathbb{R}^{T \times D_{TS}}$, where $\mathbf{x}_t^{Ts} \in \mathbb{R}^{D_{TS}}$ contains physiological measurements such as vital signs and laboratory results. Measurements are sampled irregularly and may contain missing values. Clinical notes are represented by $\mathbf{X}^{Txt} = \{x_1^{Txt}, \ldots, x_T^{Txt}\}$, with each note associated with a position on the patient timeline. The labels $\mathbf{y} \in \{0,1\}^C$ specify the C prediction targets. We also use a guideline collection $\mathcal{G} = \{g_1, \ldots, g_S\}$, where each statement g_s describes clinical criteria or a stage of care, such as evaluation, monitoring, or intervention. This collection supplies a shared reference across patients; the time-series and text observations provide the individual evidence from which prototypes are mined.We learn a prediction function $f : (\mathbf{X}^{TS}, \mathbf{X}^{Txt}, \mathcal{G}) \to \hat{\mathbf{y}}$ from patient observations and shared guideline knowledge.

3.2 TEMPORAL ENCODING OF MULTIMODAL EHRS

Physiological measurements and clinical notes differ in sampling frequency, temporal granularity, and representation. Measurements describe numerical trajectories, whereas notes record observations at discrete times in clinical language. Separate encoders accommodate these differences and map both modalities to a shared feature dimension D. Time interpolation aligns the numerical observations to a regular grid, and positional encodings retain the location of notes on the patient timeline. This gives prototype queries access to both the content and temporal context of the record.

Time Series Encoding. For a multivariate time series , a learned gate combines observations with time-dependent statistics:

$$\tilde{\mathbf{x}}_t^{TS} = \mathbf{g}_t \odot \mathbf{x}_t^{TS} + (1 - \mathbf{g}_t) \odot \mu_t, \tag{1}$$

where $\mathbf{g}_t = \sigma(\mathbf{W}_g \mathbf{x}_t + \mathbf{b}_g)$ controls the contribution of each input and μ_t is the moving average at time step t.The imputed sequence is then aligned to a regular time grid via mTAND interpolation Shukla & Marlin (2021), which assigns higher weights to observations closer in time. For each query time t, attention is restricted to past observations ($t_s \leq t$, causal masking):

$$\begin{aligned} \mathbf{z}_t^{\text{interp}} &= \sum_{s:\, t_s \leq t} \alpha_{s,t} \cdot \tilde{\mathbf{x}}_{t_s}^{TS}, \\ \alpha_{s,t} &= \frac{\exp(\text{sim}(\mathbf{q}(t), \mathbf{k}(t_s))\,/\tau)}{\sum_{s':\, t_{s'} \leq t} \exp(\text{sim}(\mathbf{q}(t), \mathbf{k}(t_{s'}))\,/\tau)}, \end{aligned} \tag{2}$$

where $\mathbf{q}(t) = \mathbf{W}_q\, \boldsymbol{\phi}(t)$ and $\mathbf{k}(t_s) = \mathbf{W}_k\, \boldsymbol{\phi}(t_s)$ are learnable time embeddings, sim denotes cosine similarity, and τ is the temperature coefficient.The interpolated sequence passes through cascaded 1D convolutions with increasing dilation rates. Their different receptive fields capture local changes and patterns extending over longer intervals. We fuse and project these outputs to obtain $\mathbf{Z}_{TS}^{\text{MS}} \in \mathbb{R}^{T \times D}$, which supplies the time-series features for prototype mining.

Text Encoding. Clinical notes are encoded using Bert-tiny Turc et al. (2019); Bhargava et al. (2021), a domain-adapted BERT model pretrained on clinical text from MIMIC. We extract the [CLS] token as the document-level representation at each time step:

$$h_t^{Txt} = E_{\text{BERT}}(x_t^{Txt}) \tag{3}$$

We project each note embedding to dimension D and add a learned encoding of its position on the patient timeline:

$$\mathbf{z}_t^{Txt} = \mathbf{W}_p \mathbf{h}_t^{Txt} + \text{PE}(t), \tag{4}$$

where $\text{PE}(t)$ is a learned positional encoding and $\mathbf{W}_p$ maps the BERT output to dimension D. The positional term preserves when a note occurs in the record, allowing its content to be interpreted relative to the physiological trajectory. The resulting sequence is $\mathbf{Z}_{Txt} = \{\mathbf{z}_1^{Txt}, \ldots, \mathbf{z}_T^{Txt}\} \in \mathbb{R}^{T \times D}$.

Clinical guidelines are encoded statement by statement using the same Bert-tiny backbone and projection $\mathbf{W}_p$. We retain a bank of $S > 1$ statement embeddings and a separate pooled summary:

$$\begin{aligned} \mathbf{z}_{G,s} &= \mathbf{W}_p \, E_{\text{BERT}}(g_s), \quad \bar{\mathbf{z}}_G = \frac{1}{S}\sum_{s=1}^{S} \mathbf{z}_{G,s}, \\ \mathbf{Z}_G &= [\mathbf{z}_{G,1}; \ldots; \mathbf{z}_{G,S}] \in \mathbb{R}^{S\times D}. \end{aligned} \tag{5}$$

The statement bank and pooled summary serve different purposes. The bank keeps individual clinical criteria available as separate attention keys and values, so a prototype can assign different weights to different statements. The summary provides a global prior for query initialization, the alignment target, and the guideline prediction query. Pooling is therefore used where a single shared reference is needed, while calibration retains the finer statement-level representation.

3.3 Guideline-Anchored Cross-Modal Prototype Alignment Module

Existing prototype methods may drift toward statistically salient but clinically irrelevant patterns, limiting both noise filtering and interpretability. To mitigate this limitation, we initialize shared queries from guideline priors, calibrate modality-specific prototypes through guideline cross-attention, and explicitly align them with the guideline embedding in latent space. The alignment loss suppresses spurious patterns and preserves clinical traceability, while reconstruction retains patient-specific information.

Knowledge-Constrained Prototype Mining. To inject clinical prior knowledge into prototype extraction, we first generate K initial prototype slots by perturbing the pooled guideline representation $\bar{z}_G \in \mathbb{R}^D$. Specifically, each slot is initialized by adding an independent Gaussian perturbation to $\bar{z}_G$:

$$\mathbf{P}_{\text{share}} = \left[\, \bar{\mathbf{Z}}_G + \epsilon_1;\ \bar{\mathbf{Z}}_G + \epsilon_2;\ \ldots;\ \bar{\mathbf{Z}}_G + \epsilon_K \,\right], \tag{6}$$

where $P_{\text{share}} \in \mathbb{R}^{K\times D}$ and $\epsilon_k \sim \mathcal{N}(0, \sigma^2 I_D)$ for $k = 1, \ldots, K$. Each slot $P_{\text{share},k} = \bar{z}_G + \epsilon_k$ shares the same global guideline prior while introducing controlled diversity for subsequent prototype mining. These shared queries are then iteratively refined through Slot Attention over modality-specific features to extract patient-specific prototypes. At each iteration, the current prototype slots attend to the modality features, and a GRU updates each slot using the aggregated evidence from the patient record. Taking the time-series modality as an example, we initialize the slots as $P_{TS}^{(0)} = P_{\text{share}}$ and perform R iterative updates:

$$\mathbf{Q}_{TS}^{(r)} = \mathbf{P}_{\text{TS}}^{(r-1)}\mathbf{W}_Q, \quad \mathbf{K}_{TS} = \mathbf{Z}_{TS}^{\text{MS}}\mathbf{W}_K, \quad \mathbf{V}_{TS} = \mathbf{Z}_{TS}^{\text{MS}}\mathbf{W}_V. \tag{7}$$

$$\begin{aligned} \mathbf{\Pi}_{TS}^{(r)} &= \text{Softmax}\left(\frac{\mathbf{Q}_{TS}^{(r)}\mathbf{K}_{TS}^{\top}}{\sqrt{D}}\right)\mathbf{V}_{TS}, \\ \mathbf{P}_{TS}^{(r)} &= f(\text{GRU}(\mathbf{\Pi}_{TS}^{(r)}, \mathbf{P}_{TS}^{(r-1)})), \end{aligned} \tag{8}$$

where r indexes the Slot Attention iteration, and the GRU recurrently updates each prototype using the evidence aggregated from the time-series features and its representation from the previous iteration. $f(\cdot)$ denotes an MLP refinement layer. After R iterations, $P_{TS} = P_{TS}^{(R)} \in \mathbb{R}^{K\times D}$ denotes the raw time-series prototypes. The text prototypes P_{Txt} are obtained using the same iterative procedure over Z_{Txt}.Despite guideline-informed initialization, raw prototypes $\mathbf{P}_{TS}$ and $\mathbf{P}_{Txt}$ may still deviate from clinical standards due to data noise and artifacts. To correct this deviation, we use $\mathbf{Z}_G$ (the gold standard encoding full medical guidelines) to calibrate both sets of raw prototypes via cross-attention:

$$\begin{aligned} \mathbf{P}_{TS}^{*} &= \mathbf{P}_{TS} + \text{Attention}(\mathbf{P}_{TS}, \mathbf{Z}_G, \mathbf{Z}_G), \\ \mathbf{P}_{Txt}^{*} &= \mathbf{P}_{Txt} + \text{Attention}(\mathbf{P}_{Txt}, \mathbf{Z}_G, \mathbf{Z}_G). \end{aligned} \tag{9}$$

This step pulls raw prototypes toward clinically meaningful regions in the embedding space, ensuring learned representations conform to evidence-based medical standards.

Learning Objectives. To enforce consistent semantics across modalities, we introduce an align loss inspired by InfoNCE Oord et al. (2018), which aligns time series and text prototypes with the guideline embedding $\mathbf{Z}_G$ (the sole gold standard). Concretely, each prototype is pulled closer to $\mathbf{Z}_G$

while being pushed away from other prototypes:

$$\mathcal{L}_{ali} = \mathcal{L}_{ali}^{TS} + \mathcal{L}_{ali}^{Txt},$$
$$\mathcal{L}_{ali}^{m} = -\sum_{k=1}^{K} \log \frac{\exp(a_{m,k})}{\exp(a_{m,k}) + \sum_{j=1}^{K} \exp(b_{m,k,j})}. \tag{10}$$

where K denotes the number of prototypes, $\mathcal{L}_{ali}^{TS}$ and $\mathcal{L}_{ali}^{Txt}$ represent the alignment losses for time-series and text modalities, respectively, while $a_{m,k}$ and $b_{m,k,j}$ correspond to positive and negative prototype matching scores. Using $\mathbf{Z}_G$ as the unified positive target suppresses prototypes dominated by clinically irrelevant correlations and ties the retained patterns to a shared guideline reference, thereby improving noise filtering and interpretability.

To prevent prototype representations from losing inherent information of original multimodal data, we set reconstruction constraints. Specifically, we employ a temporal decoder to reconstruct the guideline-aligned temporal prototype $\mathbf{P}_{TS}^{*}$ and obtain the standardized regular temporal sequence $\hat{\mathbf{Z}}_{TS}^{MS}$. Similarly, we utilize a text embedding decoder to implement text feature reconstruction $\hat{\mathbf{Z}}^{Txt}$ using the guideline-aligned text prototype $\mathbf{P}_{Txt}^{*}$. Both decoders adopt the Transformer decoder structureVaswani et al. (2017), and mean square error loss is adopted as the constraint loss function.

$$\mathcal{L}_{recon} = \text{MSE}(\mathbf{Z}_{TS}^{MS}, \hat{\mathbf{Z}}_{TS}^{MS}) + \text{MSE}(\mathbf{Z}_{Txt}, \hat{\mathbf{Z}}_{Txt}). \tag{11}$$

Intra-Modality Interaction. To capture global cross-modal, cross-temporal, and cross-prototype dependencies, we concatenate four component sequences into a sequence and feed it into an L-layer Transformer encoder: guideline-aligned time series prototypes $\mathbf{P}_{TS}^{*}$, time series multi-scale embeddings $\mathbf{Z}_{TS}^{MS}$, guideline-aligned text prototypes $\mathbf{P}_{Txt}^{*}$, and temporally-aligned text embeddings $\mathbf{Z}_{Txt}$. Self-attention enables full interaction across all elements, and the final output forms the comprehensive patient representation:

$$\mathbf{H}_{Pat} = \text{Transformer}\big([\mathbf{Z}_{TS}, \mathbf{P}_{TS}^{*}, \mathbf{Z}_{Txt}, \mathbf{P}_{Txt}^{*}]\big). \tag{12}$$

3.4 Dual-Branch Prediction

A single prediction stream may overemphasize either standardized evidence or individual variation. We therefore use complementary queries over the shared patient representation $\mathbf{H}_{Pat}$: the guideline embedding $\mathbf{Z}_G$ retrieves evidence consistent with clinical criteria, whereas the patient text embedding $\mathbf{Z}_{Txt}$ retrieves atypical symptoms and comorbidities not fully covered by guidelines. Their fusion mirrors clinicians' joint consideration of standard criteria and patient-specific deviations.

Branch 1: Guideline-Evidence Branch. The pooled summary queries patient features; h_{guide} pools the attention output and projects it to C task logits:

$$\hat{\mathbf{y}}_{guide} = \sigma\left(h_{guide}\left(\text{Attention}(\bar{\mathbf{z}}_G, \mathbf{H}_{Pat}, \mathbf{H}_{Pat})\right)\right). \tag{13}$$

Branch 2: Individual-Specificity Branch. Clinical-note embeddings query patient features, with h_{pat} pooling and projecting the output to C logits:

$$\hat{\mathbf{y}}_{pat} = \sigma\left(h_{pat}\left(\text{Attention}(\mathbf{Z}_{Txt}, \mathbf{H}_{Pat}, \mathbf{H}_{Pat})\right)\right). \tag{14}$$

Both branches attend to the same $\mathbf{H}_{Pat}$, but their query sources differ. The guideline summary is shared, whereas the note queries vary with the patient. Their outputs have the same C-dimensional prediction space and are fused through a weighted sum:

$$\hat{\mathbf{y}}_{final} = w \cdot \hat{\mathbf{y}}_{guide} + (1 - w) \cdot \hat{\mathbf{y}}_{pat}, \tag{15}$$

where $w \in [0, 1]$ is a validation-selected fusion weight. The model is trained end-to-end with the multi-objective loss:

$$\mathcal{L}_{total} = \mathcal{L}_{BCE}(\hat{\mathbf{y}}_{final}, \mathbf{y}) + \lambda_1 \mathcal{L}_{ali} + \lambda_2 \mathcal{L}_{recon}, \tag{16}$$

where $\mathcal{L}_{BCE}$ is binary/multi-label cross-entropy loss and $\lambda_1, \lambda_2 \geq 0$ weight the regularizers.

Table 1: Comparison of our method with baselines on 48-IHM and 24-PHE tasks using the MIMIC-III dataset. Values are reported as mean ± std over five repeated runs. The best results are highlighted in **Bold**. † denotes the method proposed in this work.

Method		48-IHM			24-PHE		
		AUROC↑	AUPRC↑	F1↑	AUROC↑	AUPRC↑	F1↑
Methods on Single-modal	CNN	79.029±0.32	35.368±0.67	38.489±4.21	69.853±0.28	31.111±0.26	36.111±0.37
	RNN	77.290±0.74	34.471±0.28	35.287±2.86	63.996±0.11	33.245±0.27	34.226±0.44
	UTDE	78.950±0.85	37.182±0.28	37.870±0.62	69.108±1.57	31.194±0.87	30.194±0.64
	GRU-D	52.681±3.26	12.740±2.43	20.963±0.71	66.722±0.22	34.241±0.55	32.411±0.23
	SeFT	47.386±0.74	10.719±1.37	23.941±9.93	60.985±4.09	23.434±0.14	30.515±0.78
	mTAND	77.370±0.20	33.292±0.10	37.792±0.12	70.178±0.18	31.451±0.10	36.704±0.14
	DGM-O	69.036±6.34	26.592±5.14	30.107±7.74	47.552±10.02	16.419±0.82	26.948±0.77
	LSTM	77.454±0.87	34.752±1.48	37.781±1.21	71.622±1.23	33.690±0.28	37.890±0.49
	ProtoTS	79.421±1.06	38.522±0.26	39.465±0.26	71.423±1.42	34.450±0.33	35.849±0.08
	IP-Net	72.651±0.48	27.154±0.73	32.534±1.47	70.653±0.13	32.171±0.29	36.516±0.15
Methods on Multi-modal	DAFT	81.530±0.32	40.401±0.25	39.951±0.64	76.188±0.48	42.790±0.51	44.172±0.39
	MedFuse	80.020±0.20	39.002±0.71	37.290±0.22	73.561±0.45	38.667±0.64	39.015±0.21
	DrFuse	79.971±1.46	38.941±0.71	38.751±1.52	70.290±0.18	34.442±0.40	37.771±0.31
	CTPD	83.483±0.72	43.614±2.04	41.939±0.12	79.323±0.69	48.933±0.34	50.125±0.40
	GAMP†	**85.567**±0.22	**49.515**±0.12	**49.943**±0.09	**87.579**±0.04	**53.555**±0.11	**52.707**±0.28

Table 2: Ablation study on GAMP. Each row shows the performance of removing one key component.

	48-IHM			24-PHE		
	AUROC↑	AUPRC↑	F1↑	AUROC↑	AUPRC↑	F1↑
GAMP (Full)	85.567±0.22	49.515±0.12	49.943±0.09	87.579±0.04	53.555±0.11	52.707±0.28
w/o Timestamp Emb.	83.921±0.12	46.783±0.77	47.321±0.19	80.134±0.05	50.412±0.24	49.876±0.18
w/o Multi-scale Emb.	82.653±0.22	44.915±1.01	45.788±0.78	78.776±0.03	48.639±0.01	47.954±0.21
w/o G-Init	81.024±0.29	42.138±0.17	42.917±0.03	77.083±0.11	46.271±1.83	45.338±0.04
w/o G-Calib	79.438±0.32	40.522±1.28	41.264±0.28	75.619±0.04	44.887±0.28	43.521±0.29
w/o Dual-Branch	80.752±0.44	42.987±1.37	43.691±0.72	76.890±0.19	46.054±0.20	44.679±0.49

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Dataset. We use the MIMIC-III v1.4 dataset Johnson et al. (2016) following the Multi-modal MIMIC framework Zhang et al. (2023). Adult ICU patients with both valid time-series and clinical notes are retained; stays with incomplete or insufficient records are excluded. Key vital signs and laboratory results are imputed, normalized, and resampled to a unified temporal granularity; clinical notes are de-identified, cleaned, and tokenized. The dataset is split 7:1:2 at the patient level into training, validation, and test sets.

Clinical Guidelines. Source materials include disease definitions from *Wikipedia*, research articles from *PubMed*, 45,000 clinical practice guidelines, and medical textbooks. MedCPT retrieves documents relevant to diseases in MIMIC. Qwen2.5-72B Yang et al. (2024) then condenses these documents into disease-specific statements covering diagnostic criteria and key indicators. Detailed clinical guideline processing pipeline is provided in Appendix F.

Tasks. We evaluate two clinical prediction tasks. The 48-hour *In-Hospital Mortality Prediction (48-IHM)* task is binary ($C = 1$) and concerns the patient's risk of death during hospitalization. The 24-hour *Phenotype Classification (24-PHE)* task is multilabel ($C = 25$) and identifies coexist-

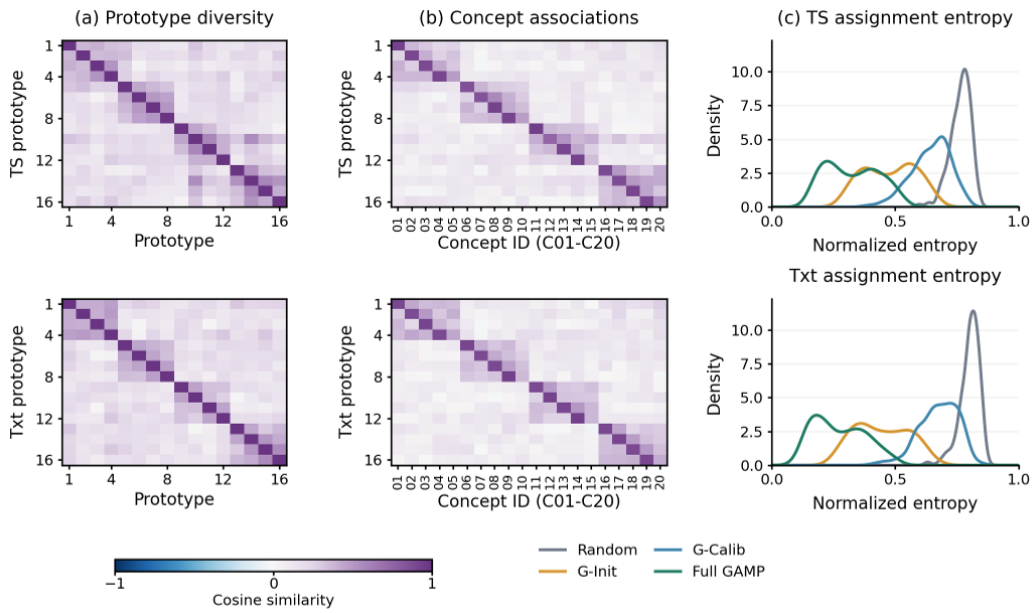


Figure 3: Analysis of Learned Clinical Prototypes at $K = 16$: Cross-Prototype Diversity, Guideline-Concept Alignment, and Assignment Entropy.

ing conditions, including congestive heart failure, renal failure, and respiratory failure. We report AUROC, AUPRC, and F1 for both tasks, with higher values indicating better performance.

4.2 Baseline Comparison

We compare GAMP with single-modal and multimodal models to examine performance relative to both physiological modeling and cross-modal fusion. The single-modal baselines include CNN LeCun et al. (1998), RNN Elman (1990), UTDE Zhang et al. (2023), GRU-D Che et al. (2018), SeFT Horn et al. (2020), mTAND Shukla & Marlin (2021), DGM-O Wu et al. (2021), LSTM Schmidhuber & Hochreiter (1997), ProtoTS Peng et al. (2025), and IP-Net Shukla & Marlin (2019). Multimodal baselines include DAFT Pölsterl et al. (2021), MedFuse Hayat et al. (2022), DrFuse Yao et al. (2024), and CTPD Wang et al. (2025).

Table 1 shows that GAMP achieves the highest AUROC, AUPRC, and F1 on both tasks. On 48-IHM, it obtains an AUROC of 85.567% and an AUPRC of 49.515%, improving over CTPD by 2.08 and 5.90 points. On 24-PHE, AUROC reaches 87.579% and AUPRC reaches 53.555%, with gains of 8.26 and 6.62 points. F1 also improves, from 41.939% to 49.943% on 48-IHM and from 50.125% to 52.707% on 24-PHE. The comparison with CTPD is particularly relevant because it also extracts cross-modal prototypes from time series and notes. These results establish a predictive advantage for the full GAMP model on the evaluated tasks.

4.3 Ablation Study

To validate the effectiveness of each core component, we ablate five variants of GAMP. Table 2 presents the results. Removing timestamp embeddings and multi-scale convolutions reduces the AUROC by 1.65% and 2.91% on 48-IHM, and 7.45% and 8.80% on 24-PHE. Performance declines most significantly without guideline-based mechanisms. The absence of guideline-guided initialization leads to 4.54% and 10.50% AUROC decreases, while cross-attention calibration removal causes the maximum performance loss of 6.13% and 11.96%. Eliminating the dual-branch decoder brings an AUROC decline above 4.8% on both tasks. These results confirm that temporal modeling provides basic feature extraction capability, and embedded guideline knowledge dominates model prediction performance. Details are in Appendix E.

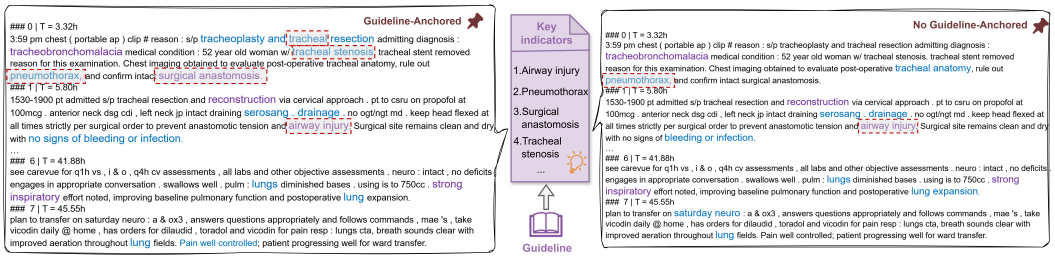


Figure 4: Qualitative case visualization illustrating differences in attention distributions: the guideline-anchored GAMP model (left) versus the baseline model without guideline constraints (right), generated on a randomly sampled patient clinical record.

4.4 Clinical Prototype Analysis

We organize the analysis around three complementary questions: **i).** how prototype count affects prediction, **ii).** whether the prototypes are diverse and capture clinically meaningful semantics, and **iii).** how confidently observations are assigned to them. Firstly, We compare $K \in \{4, 8, 16, 32, 64\}$ on 48-IHM and 24-PHE. The guideline-anchored model achieves its highest AUROC and AUPRC at $K = 16$ among the tested settings. We therefore use 16 prototypes for the following analyses.

The pairwise cosine-similarity heatmap among the 16 prototypes exhibits clear heterogeneity rather than uniformly high similarity, suggesting that the prototypes do not collapse into redundant representations. To further examine whether these diverse prototypes capture clinically meaningful semantics, we compute their cosine similarity with the top 20 guideline-derived concepts that occur frequently and receive high attention weights. The resulting prototype–concept similarity heatmap shows selective associations, with different prototypes preferentially aligned with different clinical concepts, supporting the semantic relevance of the learned prototype space. Related results for $K = 16$ are shown in Figure 3, with corresponding results for $K \in \{4, 8, 32, 64\}$ in the Appendix B.

Finally, we compare prototype-assignment entropy under Random, G-Init, G-Calib, and Full GAMP. As shown in subfigure (c) of Figure 3, the progressive change in the entropy distribution indicates that guideline initialization and calibration reshape prototype utilization, while the full model shifts the assignment distribution toward lower entropy, yielding more concentrated prototype assignments.

4.5 Case Study

We perform qualitative case analysis on a randomly sampled test-set patient record, comparing attention weight distributions between the vanilla baseline model and GAMP with the guideline-anchored alignment module. As shown in Figure 4 (purple: top 20% attention; blue: top 40%), GAMP focuses on clinically decisive terms such as airway injury, pneumothorax, surgical anastomosis and tracheal stenosis, aligned with critical care diagnostic benchmarks.The model's prioritized regions strongly overlap with core terminology from official tracheoplasty/tracheal resection guidelines (red boxes), confirming GAMP avoids spurious EHR associations and its predictions are bounded by authoritative clinical guidelines.

5 Conclusion

In this paper, we proposed the **G**uideline-**A**nchored **M**ultimodal Fusion with **P**rototype Mining (GAMP) framework for clinical outcome prediction from multimodal EHR data. GAMP leverages clinical practice guidelines as knowledge anchors to guide prototype initialization and calibration,ensuring that the learned representations remain clinically meaningful and the model output is trustworthy for real-world deployment. A dual-branch decision module further integrates standardized guideline evidence with patient-specific features, making predictions both allowing the prediction module to use both guideline-derived and patient-specific information. Our experiments on the MIMIC-III dataset demonstrate the effectiveness of GAMP in multi-modal EHR analysis.

AI USE STATEMENT

Generative AI tools assisted with language editing, narrative organization, and mathematical consistency checks.

ETHICS STATEMENT

This study uses de-identified MIMIC-III records for retrospective outcome prediction. The reported experiments assess predictive performance on these records; they do not evaluate use in patient care.

REPRODUCIBILITY STATEMENT

The Method section defines the encoders, prototype learning procedure, and training objectives. The Experimental Setup describes the cohort selection, data split, preprocessing, and training configuration. Results are reported across five random seeds.We will release our source code and experiment configs on GitHub upon acceptance.

REFERENCES

Prajjwal Bhargava, Aleksandr Drozd, and Anna Rogers. Generalization in NLI: Ways (not) to go beyond simple heuristics. 2021.

Zi Cai, Yu Liu, Zhiyao Luo, and Tingting Zhu. Protoehr: Hierarchical prototype learning for ehr-based healthcare predictions. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pp. 169–178, 2025.

Zhengpat 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.

Binyue Chen and Guohua Liu. D4care: A deep dynamic memory-driven cross-modal feature representation network for clinical outcome prediction. *Applied Sciences*, 15(11):6054, 2025.

Xiudi Chen, Hui Wu, and Xiaodong Shi. Consistent prototype learning for few-shot continual relation extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7409–7422, 2023.

Jeffrey L Elman. Finding structure in time. *Cognitive Science*, 14(2):179–211, 1990.

Tianyu Gao, Xu Han, Zhiyuan Liu, and Maosong Sun. Hybrid attention-based prototypical networks for noisy few-shot relation classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 6407–6414, 2019.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific Language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23, 2021.

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.

Alistair Johnson, Tom Pollard, and Roger Mark. MIMIC-III Clinical Database. *PhysioNet*, September 2016. doi: 10.13026/C2XW26. URL `https://doi.org/10.13026/C2XW26`. Version 1.4.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

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, Pan Zhou, Caiming Xiong, and Steven CH Hoi. Prototypical contrastive learning of unsupervised representations. *arXiv preprint arXiv:2005.04966*, 2020.

Yuxing Lu, Xukai Zhao, and Jinzhuo Wang. Clinicalrag: Enhancing clinical decision support through heterogeneous knowledge retrieval. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pp. 64–68, 2024.

Christian Nash, Rajesh Nair, and Syed Mohsen Naqvi. Cross-modal attention for multimodal information fusion: A novel approach to attention deficit hyperactivity disorder detection. In *2024 27th International Conference on Information Fusion (FUSION)*, pp. 1–6. IEEE, 2024.

Shuai Niu, Jing Ma, Hongzhan Lin, Liang Bai, Zhihua Wang, Richard Yi Da Xu, Guo Li, Xian Yang, et al. Promedts: A self-supervised, prompt-guided multimodal approach for integrating medical text and time series. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 5915–5928, 2025.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Alvaro Lopez Pellicer, Andre Mariucci, Plamen Angelov, Marwan Bukhari, and Jemma G Kerns. Protomedx: Towards explainable multi-modal prototype learning for bone health classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7357–7366, 2025.

Ziheng Peng, Shijie Ren, Xinyue Gu, Linxiao Yang, Xiting Wang, and Liang Sun. Protots: Learning hierarchical prototypes for explainable time series forecasting. *arXiv preprint arXiv:2509.23159*, 2025.

Sebastian Pölsterl, Tom Nuno Wolf, and Christian Wachinger. Combining 3d image and tabular data via the dynamic affine feature map transform. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 688–698. Springer, 2021.

Nikolaos Rodis, Christos Sardianos, Panagiotis Radoglou-Grammatikis, Panagiotis Sarigiannidis, Iraklis Varlamis, and Georgios Th Papadopoulos. Multimodal explainable artificial intelligence: A comprehensive review of methodological advances and future research directions. *IEEE Access*, 12:159794–159820, 2024.

Jürgen Schmidhuber and Sepp Hochreiter. Long short-term memory. *Neural Computation*, 9(8): 1735–1780, 1997.

Satya Narayan Shukla and Benjamin M Marlin. Interpolation-prediction networks for irregularly sampled time series. *arXiv preprint arXiv:1909.07782*, 2019.

Satya Narayan Shukla and Benjamin M Marlin. Multi-time attention networks for irregularly sampled time series. *arXiv preprint arXiv:2101.10318*, 2021.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in Neural Information Processing Systems*, 30, 2017.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Well-read students learn better: The impact of student initialization on knowledge distillation. *arXiv preprint arXiv:1908.08962*, 13:3, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

Fuying Wang, Feng Wu, Yihan Tang, and Lequan Yu. CTPD: Cross-modal temporal pattern discovery for enhanced multimodal electronic health records analysis. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 6783–6799, 2025.

Yunyun Wang, Jian Mao, Chenguang Guo, and Songcan Chen. Contrastive prototype-guided generation for generalized zero-shot learning. *Neural Networks*, 176:106324, 2024.

Feng Wu, Tsai Hor Chan, Fuying Wang, Guosheng Yin, and Lequan Yu. Cross-modal alignment via variational copula modelling. *arXiv preprint arXiv:2511.03196*, 2025.

Yinjun Wu, Jingchao Ni, Wei Cheng, Bo Zong, Dongjin Song, Zhengzhang Chen, Yanchi Liu, Xuchao Zhang, Haifeng Chen, and Susan B Davidson. Dynamic gaussian mixture based deep generative model for robust forecasting on sparse multivariate time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 651–659, 2021.

An Yang, Baosong Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

Haiyang Yang, Li Kuang, and FengQiang Xia. Multimodal temporal-clinical note network for mortality prediction. *Journal of Biomedical Semantics*, 12(1):3, 2021.

Wenfang Yao, Kejing Yin, William K Cheung, Jia Liu, and Jing Qin. Drfuse: Learning disentangled representation for clinical multi-modal fusion with missing modality and modal inconsistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 16416–16424, 2024.

Michihiro Yasunaga, Antoine Bosselut, Hongyu Ren, Xikun Zhang, Christopher D Manning, Percy S Liang, and Jure Leskovec. Deep bidirectional language-knowledge graph pretraining. *Advances in Neural Information Processing Systems*, 35:37309–37323, 2022.

Jiexia Ye, Weiqi Zhang, Ziyue Li, Jia Li, Meng Zhao, and Fugee Tsung. Medualtime: A dual-adapter language model for medical time series-text multimodal learning. *arXiv preprint arXiv:2406.06620*, 2024.

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.

Xuejiao Zhao, Siyan Liu, Su-Yin Yang, and Chunyan Miao. Medrag: Enhancing retrieval-augmented generation with knowledge graph-elicited reasoning for healthcare copilot. In *Proceedings of the ACM Web Conference 2025*, pp. 4442–4457, 2025.

A MIMIC-III TASK AND DATASET STATISTICS

We summarize the label distributions for in-hospital mortality prediction (IHM) and phenotype classification (Pheno) on MIMIC-III in Table 3. The phenotype classification task includes 25 disease labels. Each phenotype is treated as a binary label, and the table reports the total number of examples and the numbers of positive and negative examples for each label. The phenotype labels exhibit class imbalance, with negative examples outnumbering positive examples for every listed label.

Table 3: Label distributions for the MIMIC-III tasks. IHM denotes in-hospital mortality prediction within 48 hours, and Pheno denotes phenotype classification within 24 hours. Counts are reproduced from the supplied dataset summary.

Task	Label	Positive	Negative	Total
IHM	In-hospital mortality	2528	16788	19,316
PHE	Acute and unspecified renal failure	8,247	30,398	38,645
	Acute cerebrovascular disease	2,835	35,810	38,645
	Acute myocardial infarction	3,985	34,660	38,645
	Cardiac dysrhythmias	12,412	24,233	38,645
	Chronic kidney disease	5,164	33,481	38,645
	Chronic obstructive pulmonary disease and bronchiectasis	5,005	33,640	38,645
	Complications of surgical procedures or medical care	7,993	30,652	38,645
	Conduction disorders	2,781	35,864	38,645
	Congestive heart failure; nonhypertensive	10,362	28,283	38,645
	Coronary atherosclerosis and other heart disease	12,475	26,170	38,645
	Diabetes mellitus with complications	3,683	34,962	38,645
	Diabetes mellitus without complication	7,475	31,170	38,645
	Disorders of lipid metabolism	11,201	27,444	38,645
	Essential hypertension	16,207	22,438	38,645
	Fluid and electrolyte disorders	10,344	28,301	38,645
	Gastrointestinal hemorrhage	2,838	35,807	38,645
	Hypertension with complications and secondary hypertension	5,127	33,518	38,645
	Other liver diseases	3,439	35,206	38,645
	Other lower respiratory disease	2,014	36,631	38,645
	Other upper respiratory disease	1,553	37,092	38,645
	Pleurisy; pneumothorax; pulmonary collapse	3,355	35,290	38,645
	Pneumonia (except that caused by tuberculosis or sexually transmitted disease)	5,336	33,309	38,645
	Respiratory failure; insufficiency; arrest (adult)	6,946	31,699	38,645
	Septicemia (except in labor)	5,483	33,162	38,645
	Shock	3,015	35,630	38,645

B PROTOTYPE ANALYSIS

How prototype count affects prediction? We analyze the sensitivity of our framework to the number of prototypes K on both the 48-IHM and 24-PHE tasks. The model achieves its highest AUROC and AUPRC at $K = 16$ among the tested settings, and performance gradually degrades as K moves further away from this value in either direction, indicating that the model benefits from a moderate number of diverse clinical prototypes but suffers from under-representation (too few prototypes) or noisy aggregation (too many). Furthermore, we evaluate the impact of guideline guidance (i.e., G-Init and G-Calib) across all K settings. The results consistently show substantial performance gains when guideline-based prototype initialization and calibration are enabled, confirming the effectiveness of the guideline anchoring mechanism regardless of the prototype count. The visualization results are provided in Figure 5.

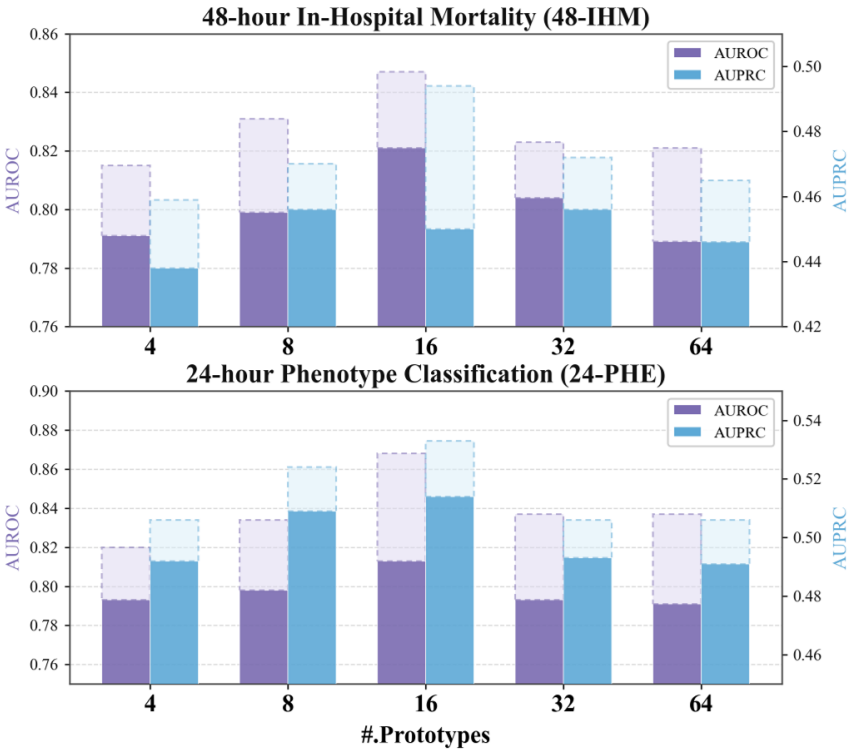


Figure 5: Sensitivity analysis of the number of prototypes K on 48-IHM and 24-PHE. The lighter, dashed-bordered bars on top indicate the performance gain from guideline anchoring.

whether the prototypes are diverse and capture clinically meaningful semantics?To examine whether the observations in Figure 3 are specific to $K = 16$, we further visualize prototype diversity and prototype–concept associations for $K \in \{4, 8, 32, 64\}$. As shown in Figure 6, 7,, 8 and 9, the learned prototypes consistently exhibit non-uniform pairwise similarities and selective associations with guideline-derived concepts across different K.The top 20 guideline-derived concepts with high frequency and attention weights are shown in Table 4 .These results suggest that the prototype structure observed at $K = 16$ is qualitatively consistent across different prototype configurations.

Table 4: The top 20 guideline-derived concepts with high frequency and attention weights

ID	concept	ID	concept	ID	concept
01	sinus rhythm	08	mitral valve	15	hemodynamic monitor
02	valve leaflets	09	respiratory failure	16	chest pain
03	portable ap	10	renal failure	17	glycemic control
04	lung sounds	11	fluid balance	18	stress ulcer
05	pleural effusion	12	blood products	19	bowel sounds
06	aortic valve	13	full code	20	iv contrast
07	vital signs	14	heart rhythm		

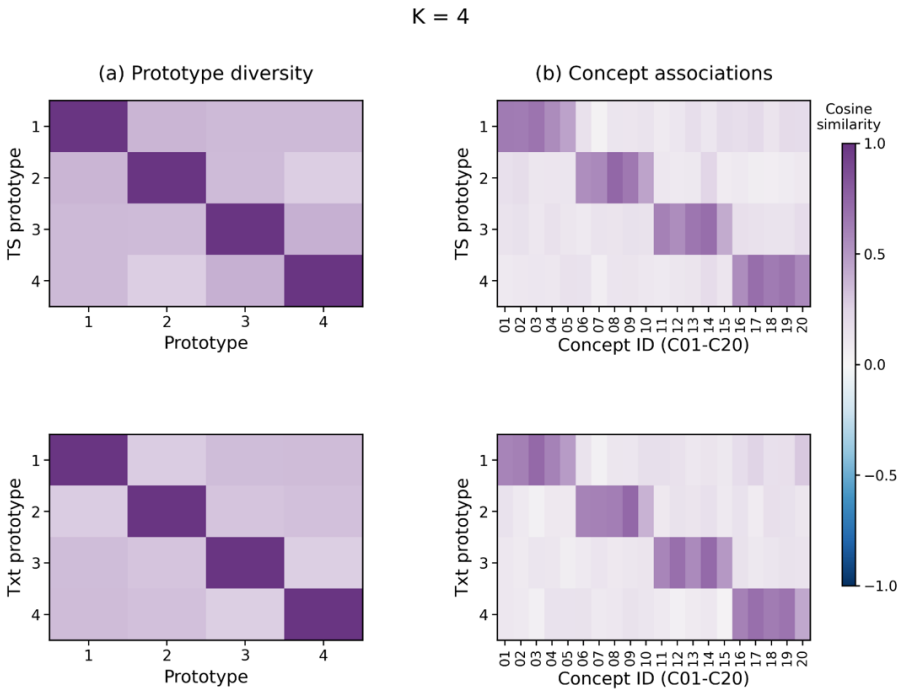


Figure 6: Analysis of Learned Clinical Prototypes at $K = 4$: Cross-Prototype Diversity and Guideline-Concept Alignment.Top and bottom rows correspond to time-series and text prototypes, respectively.

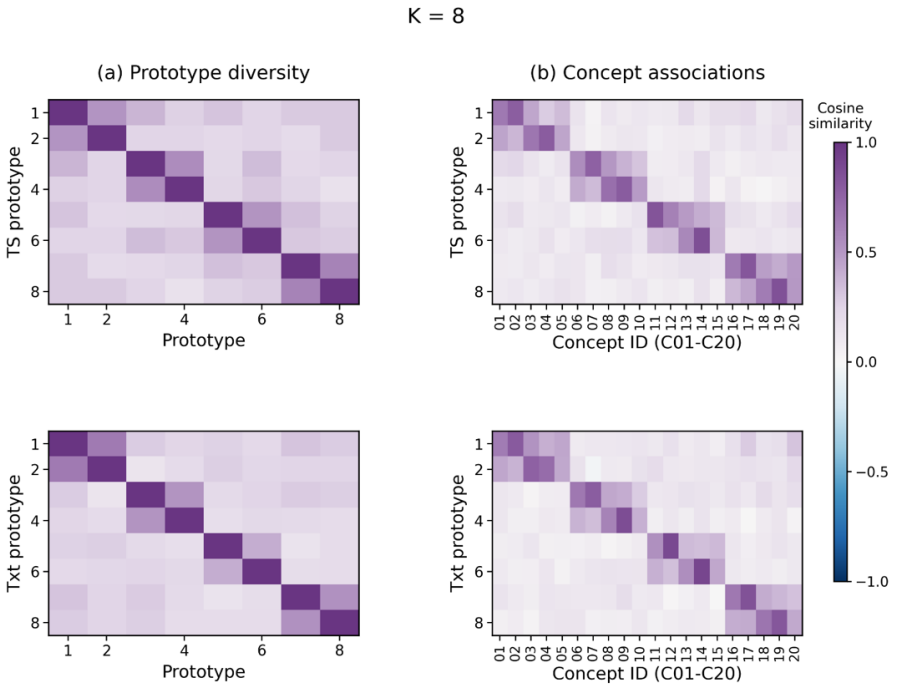


Figure 7: Analysis of Learned Clinical Prototypes at $K = 8$: Cross-Prototype Diversity and Guideline-Concept Alignment.Top and bottom rows correspond to time-series and text prototypes, respectively.

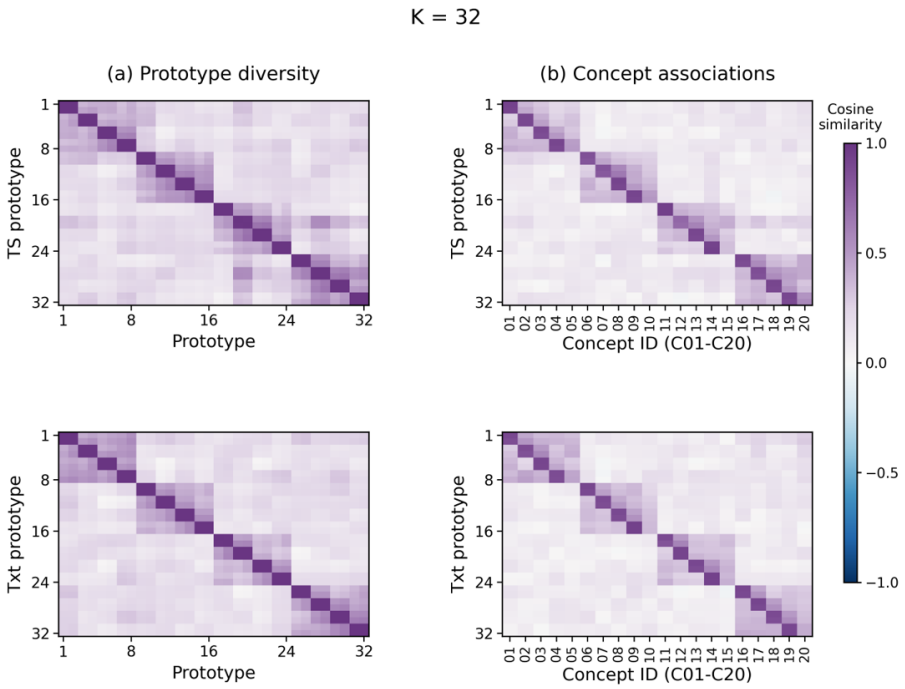


Figure 8: Analysis of Learned Clinical Prototypes at $K = 32$: Cross-Prototype Diversity and Guideline-Concept Alignment.Top and bottom rows correspond to time-series and text prototypes, respectively.

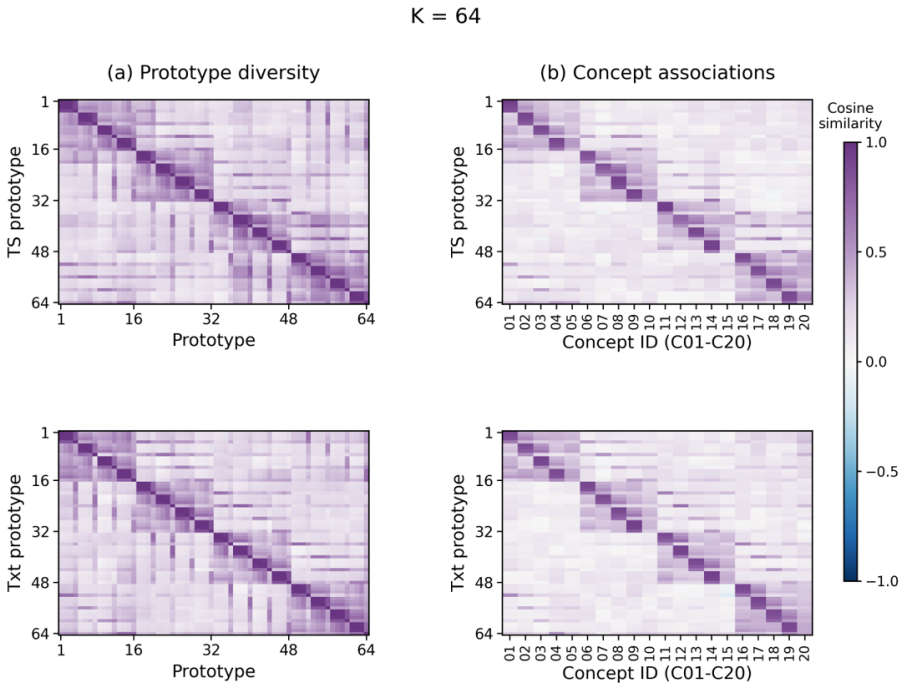


Figure 9: Analysis of Learned Clinical Prototypes at $K = 64$: Cross-Prototype Diversity and Guideline-Concept Alignment.Top and bottom rows correspond to time-series and text prototypes, respectively.

C HYPER-PARAMETER ANALYSIS

To evaluate the impact of hyperparameters in GAMP, we analyze three key hyperparameters: λ_1 and λ_2 in Eq. 16, as well as w in Eq. 15. λ_1 determines the weight for aligning temporal prototypes, text prototypes and clinical guidelines; λ_2 controls the weight of the prototype reconstruction loss, while w regulates the proportion of the Guideline-Evidence Branch and the Individual-Specificity Branch within the two-branch predictor. Figure 10 illustrates the performance trends under varying values of these hyperparameters. GAMP achieves the best performance when $\lambda_1 = 0.1$, $\lambda_2 = 0.5$, and $w = 0.05$. We observe that model performance first increases and then decreases as λ_1 rises. This pattern arises because an excessively large weight forces prototypes to over-align with clinical guidelines, which undermines the optimization of the primary classification loss. On the contrary, an overly small weight also leads to performance degradation, highlighting the role of clinical guidelines in refining multi-modal feature representations. The parameter λ_2 exhibits a similar pattern.

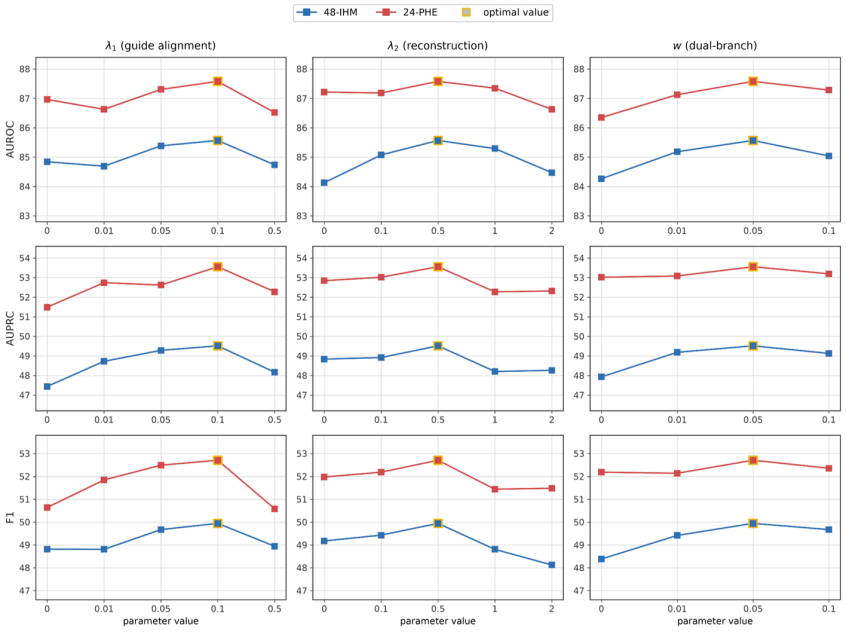


Figure 10: Sensitivity of AUROC, AUPRC, and F1 to λ_1, λ_2, and w on 48-IHM and 24-PHE. Columns vary λ_1, λ_2, and w; yellow markers denote the selected values.

D BASELINE MODELS

Single-modal Baselines.

- **CNN** (LeCun et al., 1998): A convolutional neural network extracts local patterns through shared convolutional filters. Applied along the temporal dimension, these filters capture short-range patterns in time-series inputs.
- **RNN** (Elman, 1990): A recurrent neural network processes observations sequentially and updates a hidden state, allowing information from previous time steps to inform the representation of the current observation.
- **UTDE** (Zhang et al., 2023): Unified Time Discretization Embedding combines representations obtained from hand-crafted imputation and learned interpolation through a gating mechanism, enabling adaptive representation learning for irregularly sampled time series.
- **GRU-D** (Che et al., 2018): GRU-D extends gated recurrent units with trainable decay mechanisms for input values and hidden states. It incorporates observation masks and time intervals to capture temporal dependencies and informative missingness.

- **SeFT** (Horn et al., 2020): Set Functions for Time Series treats an irregularly sampled time series as a set of observations and learns a permutation-invariant aggregation of their representations for classification.
- **mTAND** (Shukla & Marlin, 2021): Multi-Time Attention Networks learn continuous-time embeddings and use time-based attention to map irregular observations to a fixed-length representation for downstream prediction.
- **DGM-O** (Wu et al., 2021): This dynamic Gaussian mixture-based generative approach models sparse multivariate time series by tracking transitions in latent clustering structure, thereby capturing evolving distributions and temporal dependencies.
- **LSTM** (Schmidhuber & Hochreiter, 1997): Long Short-Term Memory is a recurrent architecture that uses memory cells and gating mechanisms to retain information over long temporal spans and mitigate difficulties in learning long-range dependencies.
- **ProtoTS** (Peng et al., 2025): ProtoTS is an interpretable forecasting framework that learns a hierarchy of temporal prototypes. Coarse prototypes represent global patterns, while finer prototypes describe local variations, supporting explanations at multiple levels of granularity.
- **IP-Net** (Shukla & Marlin, 2019): Interpolation-Prediction Networks combine a semi-parametric interpolation network with a prediction network. The interpolation stage shares information across variables to construct representations of sparse, irregularly sampled time series.

Multimodal Baselines.

- **DAFT** (Pölsterl et al., 2021): Dynamic Affine Feature Map Transform is a general-purpose CNN module that conditions the scaling and shifting of convolutional feature maps on a patient's tabular clinical information, enabling image–tabular fusion.
- **MedFuse** (Hayat et al., 2022): MedFuse uses an LSTM-based fusion module to combine clinical time-series representations with chest X-ray image representations. Its design accommodates both unimodal and multimodal inputs, including cases with missing imaging data.
- **DrFuse** (Yao et al., 2024): DrFuse separates modality-shared and modality-specific representations to address missing modalities and uses disease-wise attention to account for differences in the relevance of each modality across patients and prediction targets.
- **CTPD** (Wang et al., 2025): Cross-Modal Temporal Pattern Discovery learns corresponding temporal patterns from clinical time series and notes using shared initial representations and Slot Attention. Contrastive alignment and modality-specific reconstruction objectives encourage the learned patterns to retain complementary cross-modal information.

E Experimental Details

Implementation Details. Each run takes approximately 0.5 hours on a single NVIDIA GeForce RTX 4090 GPU. Metrics are averaged over five runs with different random seeds. We use Adam Kingma & Ba (2015) with a batch size of 128 and an initial learning rate of 4×10^{-5}. Training lasts for at most 100 epochs, with early stopping based on validation AUROC. BERT-tiny Turc et al. (2019); Bhargava et al. (2021) uses a maximum context length of 512 tokens. We perform three Slot Attention iterations $R = 3$. We set $\sigma = 0.01$ in Eq. 6 and use two Transformer encoder layers for cross-modal interaction.

Ablation Details. We consider the following ablation variants:

- **GAMP (w/o Timestamp Embeddings):** We remove the learnable temporal positional encodings from the text encoding pathway, i.e., we set $\mathbf{z}_t^{txt} = \mathbf{W}_p \mathbf{h}_t^{txt}$ without $\mathrm{PE}(t)$. This evaluates the contribution of temporal alignment between clinical notes and the time-series timeline.

- **GAMP (w/o Multi-scale Embeddings):** We replace the cascaded dilated convolutions (dilation rates $\{1, 2, 4\}$) in the time-series encoder with a single-layer 1D convolution (kernel size 3, dilation rate 1), removing the multi-resolution temporal feature extraction capability.
- **GAMP (w/o G-Init):** We replace the guideline-informed prototype initialization $P_{share} = \left[\bar{z}_G + \epsilon_1; \ldots; \bar{z}_G + \epsilon_K\right]$. with a random Gaussian initialization $\mathbf{P}_{Shared} \sim \mathcal{N}(0, \mathbf{I})$, eliminating the clinical prior injected at the prototype seed stage.
- **GAMP (w/o G-Calib):** We remove the cross-attention calibration step $\mathbf{P}^*_{TS} = P_{TS} + \text{Attention}(\mathbf{P}_{TS}, \mathbf{Z}_G, \mathbf{Z}_G)$ (and its text counterpart), directly using the raw prototypes $\mathbf{P}_{TS}$ and $\mathbf{P}_{txt}$ for subsequent interactions. This isolates the contribution of the guideline anchoring mechanism.
- **GAMP (w/o Dual-Branch):** We replace the dual-branch decision module with a single-branch predictor that directly attends over $\mathbf{H}_{Pat}$ using a learned query, removing both the guideline-evidence branch and the individual-specificity branch.

F Detailed Clinical Guideline Processing Pipeline

We collect source materials from four categories: **Wikipedia**[1] providing standardized disease definitions, **PubMed**[2] offering frontier research achievements, 45,000 verified clinical practice **guidelines**, and authoritative medical **textbooks** involving surgery, imaging and pharmacology. Targeting diseases contained in the MIMIC dataset, we first adopt the medical-oriented retriever MedCPT to extract relevant documents as initial evidence. The large language model Qwen2.5-72B Yang et al. (2024) is then utilized to eliminate redundant irrelevant content and satisfy length limits. Driven by structured prompts, the LLM integrates valid information to form standardized clinical guidelines. Core diagnostic standards and vital indicators are retained while noisy and duplicate contents are discarded. Eventually, concise disease-specific guidelines adaptive to MIMIC are produced. Examples of guidelines and elaborately designed prompts are given in Figure 11 and Figure 12.

[1]Wikipedia `https://huggingface.co/datasets/wikimedia/wikipedia`
[2]PubMed `https://ftp.ncbi.nlm.nih.gov/pubmed/baseline`

```
% ============================================================
% Source: 2019 ESC Guidelines (Knuuti et al., Eur Heart J 2020;41:407–477)
% ============================================================
DISEASE: Chronic Coronary Syndromes (CCS)
--- DIAGNOSIS ---
1. Symptom Assessment: Evaluate chest pain characteristics (typical/atypical/non-anginal).
   Rule out unstable angina or acute coronary syndrome (ACS) first.
2. Basic Work-up: Resting ECG, echocardiography (assess LVEF), fasting glucose,
   lipid profile, serum creatinine. Chest X-ray in selected patients.
3. Pre-test Probability (PTP): Estimate PTP of obstructive CAD based on age, sex,
   and symptom type. Low PTP (<5%): no further testing; Intermediate (5–65%):
   non-invasive functional imaging or coronary CTA; High PTP (>65%): consider
   invasive coronary angiography.
4. Non-invasive Testing: Functional imaging (stress echocardiography, SPECT, PET)
   or coronary CTA recommended as initial test for symptomatic patients where
   obstructive CAD cannot be excluded clinically.
5. Event Risk Stratification: After CAD is confirmed, assess risk of death or
   myocardial infarction to guide treatment decisions (medical therapy vs.
   revascularization).
--- TREATMENT ---
1. Lifestyle Modification: Smoking cessation; Mediterranean diet; regular aerobic
   exercise (≥150 min/week); weight control; cardiac rehabilitation.
2. Anti-anginal Drugs (symptom relief):
   - First-line: Beta-blockers and/or calcium channel blockers.
   - Second-line (if symptoms persist): Long-acting nitrates, ivabradine,
     ranolazine, nicorandil, or trimetazidine.
3. Event Prevention:
   - Aspirin 75–100 mg/day for secondary prevention.
   - High-intensity statins (atorvastatin/rosuvastatin) regardless of LDL-C level.
   - ACE inhibitors in high-risk patients (diabetes, reduced LVEF, CKD).
4. Revascularization: PCI or CABG recommended for patients with high event risk
   and significant stenosis (FFR ≤0.80 or visual stenosis ≥90%).
--- KEY RISK MARKERS ---
- LVEF <50%, Diabetes, CKD (eGFR <60), Multi-vessel CAD, Elevated troponin/NT-proBNP.
```

Figure 11: An example of the synthesized clinical guideline for chronic coronary syndromes (CCS).

Prompt for Clinical Guideline Synthesis
System Prompt (Role)
You are a clinical guideline expert. Your task is to synthesize retrieved medical documents into concise, standardized, and evidence-based clinical guidelines for MIMIC dataset diseases.
Task Prompt
Input: Multiple retrieved documents about a specific disease from the MIMIC dataset.
Task: Synthesize these documents into a single concise clinical guideline by:
1.**Preserve:*
-Diagnostic criteria
- Key clinical indicators (lab values, vital signs thresholds)
-Treatment recommendations- Risk stratification factors
2. **Remove:*
- Redundant information (merge overlapping content)
- Irrelevant content not applicable to ICU/critical care
- Duplicate descriptions from multiple sources
Noise and non-clinical details
3. **Format:**
Factors
- Structured sections: Overview, Diagnostic Criteria, Key Indicators, Treatment, Risk
- Maximum 500 words per disease- Use bullet points for clarity
- Clinical precision, no speculation
**Output:* A clean, standardized clinical guideline ready for use as a knowledge anchor in machine learning.

Concise Version (Single Prompt)
> You are a clinical guideline expert. Given retrieved medical documents about a specific disease from MIMIC dataset, synthesize them into a concise clinical guideline (S500 words). Preserve: diagnostic criteria, key clinical indicators, treatment recommendations, risk factors Remove: redundancy, irrelevant ICU information, duplicates, noise. Output structured sections: Overview, Diagnostic Criteria, Key Indicators, Treatment, Risk Factors. Use bullet Point.Be clinically precise.

Figure 12: Our designed prompt for Qwen2.5-72B to synthesize clinical guidelines.