Retrieval-Augmented Multimodal Language Model with Heterogeneous Evidence Grounding for Clinical Prognosis

Anonymous authors
Paper under double-blind review

Abstract

Clinical prognosis requires reasoning over heterogeneous evidence: continuous waveforms, dense laboratory measurements, longitudinal patient history, and the empirical outcome patterns of similar patients. Multimodal language models offer a principled substrate for this integration, yet current systems operate over an impoverished version of the clinical record. Structured measurements lose their numerical meaning when serialized as text. Each encounter is treated in isolation, discarding the prognostic signal in serial waveform change. And all predictions are grounded exclusively in knowledge memorised during pre-training, with no mechanism to consult the outcomes of clinically similar patients. We introduce PRISM (Population-Retrieved Inference for Structured Multimodal Prognosis), which recovers each of these missing evidence streams through a matched representational strategy: per-feature numerical tokenization that preserves measurement fidelity, temporal aggregation via a learnable recency decay that distills longitudinal ECG context, and a dual-modality population memory bank that retrieves outcome-concordant historical patients as non-parametric evidence for language model reasoning. On the 1,443-task MDS-ED benchmark (MIMIC-IV), PRISM achieves state-of-the-art performance across every task group: 86.2% Diagnosis, 93.1% Deterioration, 91.8% ICU admission, and 92.8% Mortality (AUROC). Our ablations reveal that the three evidence streams serve complementary roles: numerical encoding drives laboratory-dependent diagnoses, population evidence sharpens diagnosis and ICU admission prediction, and temporal context resolves trajectory-sensitive outcomes, suggesting broader design principles for grounding clinical language models in structured, longitudinal, and population-level evidence.

1 Introduction

Clinical prognosis is fundamentally a multimodal reasoning problem. Predicting whether a patient will deteriorate, require intensive care, or die in hospital demands the simultaneous interpretation of continuous physiological waveforms, dense panels of laboratory analytes and vital signs, longitudinal acquisition histories, and the implicit epidemiological priors that experienced clinicians carry from thousands of prior cases (Smith et al., 2013; Raghunath et al., 2020). Large language models have emerged as a compelling backbone for this integration: their capacity for cross-modal attention over heterogeneous token sequences offers, in principle, a unified substrate for fusing structured and unstructured clinical data into coherent prognostic judgements (Singhal et al., 2023). In practice, however, current multimodal clinical LLMs leave substantial predictive signal unrecovered due to three structural limitations in how they represent, contextualise, and reason over clinical evidence.

The most immediate of these concerns how structured clinical data enters the model. Systems that fuse waveform encoders with language model backbones (Liu et al., 2023; Li et al., 2023; Tang et al., 2026; Pham et al., 2025) typically rely on *textualization*: laboratory values, vital signs, and demographics are serialized into natural-language strings and consumed through the tokenizer. This conversion destroys the ordinal and ratio-scale structure of numerical measurements at the input boundary. A troponin of 0.11 ng/mL and a troponin of 0.01 ng/mL become sequences of subword tokens whose clinical distance is opaque to the embedding layer (Gorishniy et al., 2021). Analytes such as troponin, lactate, and creatinine, among the

strongest independent predictors of acute deterioration and in-hospital mortality (Ahmed et al., 2014; Kruse et al., 2011), are either collapsed into coarse text summaries or discarded entirely. No amount of downstream model capacity can recover information eliminated before the forward pass.

A related limitation concerns the temporal dimension. Serial changes in a patient's ECG carry independent prognostic information beyond any single acquisition (Lehmacher et al., 2020): a new ST deviation against a previously normal baseline has different clinical meaning than the same morphology in a patient with chronic abnormalities. Yet current multimodal systems encode only the index-visit waveform (Tang et al., 2026; Zhao et al., 2025; Pham et al., 2025), treating each encounter as if it were the patient's first and forfeiting the prognostic signal embedded in longitudinal change. This is a particularly costly omission given that ECG foundation models have shown that waveform representations encode risk information well beyond traditional diagnostic categories (Raghunath et al., 2020; Li et al., 2025; Tsai et al., 2023).

Finally, even the largest language models reason exclusively from knowledge memorised during pre-training (Singhal et al., 2023). They have no mechanism to consult the empirical outcome distribution of clinically similar patients in the target population, the very evidence that anchors a clinician's bedside risk estimate. Retrieval-augmented generation has proven effective for grounding language models in external text corpora (Lewis et al., 2020) and has been applied to clinical NLP with knowledge graphs (Zhu et al., 2024), but this paradigm has not been extended to structured patient populations, where the retrievable evidence is not a document but a real historical patient with a documented outcome.

We introduce **PRISM** (**P**opulation-**R**etrieved **I**nference for **S**tructured **M**ultimodal Prognosis)[1], a multimodal architecture that grounds language model prognostic reasoning in population-level evidence while preserving the full fidelity of structured clinical data. Rather than treating the language model as a self-contained reasoner that must absorb all clinical context through its tokenizer, PRISM equips it with three external capacities: a dense numerical encoder that represents every clinical feature as an individual learned token, a temporal aggregation mechanism that distills longitudinal waveform history into a compact representation through learnable recency decay, and a population retrieval module that supplies outcome-concordant evidence from historically similar patients as non-parametric context. Each capacity is realised through a dedicated component: an FT-Transformer (Gorishniy et al., 2021) for structured EHR encoding, a learnable exponential recency-decay module for ECG history aggregation, and a dual-modality population memory bank over ECG and EHR embeddings that retrieves outcome-concordant historical patients as non-parametric context. Their outputs are projected as soft tokens into a shared sequence consumed by a medical language model (i.e., MedGemma (Sellergren et al., 2025)) fine-tuned with low-rank adaptation (Hu et al., 2022), and the unified architecture produces predictions across all target tasks in a single forward pass.

We evaluate PRISM on the MDS-ED benchmark (Lopez Alcaraz et al., 2025), the most comprehensive multimodal clinical prognosis benchmark available, comprising 1,443 binary prediction tasks over raw 12-lead ECG waveforms and 470 structured clinical variables from MIMIC-IV (Johnson et al., 2023). PRISM establishes a new state of the art across all four task groups: 86.24% on Diagnosis (1,428 ICD-10 tasks), 93.12% on Deterioration (6 tasks), 91.79% on ICU admission (2 tasks), and 92.82% on Mortality (7 tasks). The proportion of tasks exceeding AUROC 80% rises from 60.2% to 79.3%, a gain of 277 tasks now classified at high discriminative performance. Specifically, PRISM achieves high discriminative performance on 522 more tasks than the MDS-ED (Lopez Alcaraz et al., 2025) baseline and 277 more than the UniPACT (Tang et al., 2026). A systematic ablation study confirms that each architectural component contributes distinctly and that their benefits are complementary rather than redundant. Our principal contributions are:

- We propose PRISM, a multimodal architecture for clinical prognosis that unifies dense numerical encoding of structured clinical records, temporal aggregation of longitudinal waveform history, and population-level retrieval-augmented generation within a single language model reasoning framework.
- We demonstrate that per-feature numerical tokenization recovers clinically significant predictive signal lost by the textualization strategy used in prior multimodal clinical LLMs, with gains concentrated on outcomes whose discrimination depends on continuous-valued laboratory measurements.

[1] We will release the complete codebase upon acceptance.

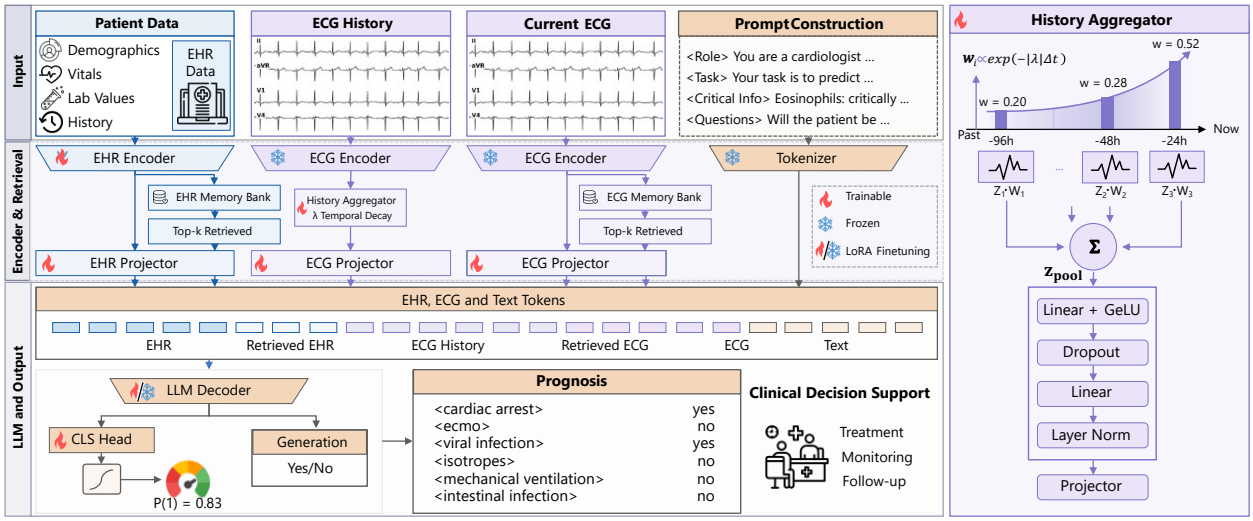

Figure 1: Overview of the PRISM architecture. Five modality streams (structured EHR, current ECG, longitudinal ECG history, retrieved EHR context, and retrieved ECG context) are independently encoded and projected as soft tokens into a shared sequence consumed by a LoRA-adapted MedGemma-4B decoder. The population memory bank (centre) retrieves outcome-concordant training patients from dual EHR and ECG indices constructed after Stage 1. The temporal history aggregator (right) pools prior ECG embeddings through learned exponential recency decay. Predictions across all target tasks are produced in a single forward pass via both generative and classification pathways.

- We introduce a temporal history aggregator with a single learnable recency-decay scalar that pools prior-visit ECG embeddings into a compact longitudinal representation at negligible parameter cost.
- We extend retrieval-augmented generation from text corpora to empirical patient populations by indexing per-patient ECG and EHR embeddings and injecting outcome-concordant neighbors as auditable, case-based soft-token context.

2 Related Work

Multimodal clinical LLMs and structured data encoding. The soft-token alignment paradigm, in which encoder outputs are projected into a language model's embedding space as learnable tokens (Liu et al., 2023; Li et al., 2023), has become the dominant approach for multimodal clinical prediction. Recent systems apply this paradigm to align ECG waveform encoders (Li et al., 2025; Tian et al., 2024; Hung et al., 2025) with LLM backbones for diagnostic and prognostic tasks (Tang et al., 2026; Pham et al., 2025; Zhao et al., 2025). A persistent design choice across these systems is that structured clinical data, including laboratory values, vital signs, and demographics, enters the language model as serialized natural-language strings rather than through encoders that preserve numerical fidelity. Transformer architectures designed for heterogeneous tabular data, such as the FT-Transformer (Gorishniy et al., 2021), map each feature to an individual token via learned per-feature projections and consistently outperform text-based representations on numerical benchmarks, yet have not been integrated into multimodal LLM pipelines for clinical prediction. Separately, pre-trained EHR Transformers such as Med-BERT (Rasmy et al., 2021) and BEHRT (Li et al., 2020) learn effective representations from longitudinal diagnostic code sequences but target discrete event vocabularies rather than the dense continuous-valued features that characterise acute-care data. We integrate an FT-Transformer based encoder directly into the multimodal LLM pipeline, encoding all 470 clinical features as individual numerical tokens and preserving the measurement fidelity that textualization discards.

Longitudinal modelling in clinical prediction. Physiological signals acquired serially carry prognostic information beyond any single measurement. In the ECG domain, serial waveform evolution provides independent predictive value for adverse cardiac events in acute presentations (Lehmacher et al., 2020), and large-scale studies have confirmed that ECG waveforms encode risk patterns invisible to expert interpretation (Raghunath et al., 2020; Tsai et al., 2023). Existing multimodal clinical LLMs, however, encode only

the index-visit acquisition, treating each encounter in isolation. Longitudinal EHR models (Rasmy et al., 2021; Li et al., 2020) capture temporal dependencies over months or years of discrete clinical events, but the problem of incorporating short-horizon, multi-acquisition waveform histories within a language model architecture remains unaddressed. We address this gap through a temporal history aggregator that pools prior-visit ECG embeddings via a single learnable recency-decay scalar, injecting longitudinal waveform context into the language model at negligible parameter cost.

Retrieval-augmented reasoning in healthcare. Retrieval-augmented generation (or RAG) improves factual consistency by grounding language model outputs in external evidence (Lewis et al., 2020). In clinical settings, this motivation is particularly strong: even large language models encoding substantial medical knowledge exhibit factual inconsistencies under expert scrutiny (Singhal et al., 2023). EMERGE (Zhu et al., 2024) applies retrieval over a clinical knowledge graph to generate task-relevant patient summaries for mortality and readmission prediction. All prior clinical RAG systems, however, retrieve from curated knowledge sources such as guidelines, ontologies, or literature. Retrieval from an empirical patient population, where the indexed evidence consists of per-patient embeddings with documented outcomes rather than text passages, remains unexplored. We close this gap by indexing learned patient representations from the training cohort into a dual-modality memory bank and surfacing clinically similar cases with known outcomes as additional context for the language model, replacing curated knowledge retrieval with direct population-grounded inference.

3 Method

3.1 Problem Formulation and Overview

Let each clinical encounter be characterized by a tuple $(\mathbf{X}_{\text{ehr}}, \mathbf{X}_{\text{ecg}}, \mathcal{H}, \mathbf{p})$, where $\mathbf{X}_{\text{ehr}} \in \mathbb{R}^D$ is a D-dimensional structured clinical feature vector (here D=470, comprising laboratory analytes, vital signs, demographics, and categorical indicators), $\mathbf{X}_{\text{ecg}} \in \mathbb{R}^{12\times T}$ is a raw 12-lead ECG waveform of length T, $\mathcal{H} = \{(\mathbf{X}_{\text{ecg}}^{(i)}, \delta_i)\}_{i=1}^{N}$ is a set of $N \geq 0$ prior ECG acquisitions with corresponding temporal offsets $\delta_i > 0$ (in hours before the index visit), and $\mathbf{p}$ is a natural-language clinical prompt specifying the prediction target. Given a set of $\mathcal{T}$ binary outcome tasks (here $|\mathcal{T}|$=1,443), the goal is to learn a function f that maps each encounter to a vector of probabilities $\hat{\mathbf{y}} = f(\mathbf{X}_{\text{ehr}}, \mathbf{X}_{\text{ecg}}, \mathcal{H}, \mathbf{p}) \in [0,1]^{|\mathcal{T}|}$, where each $\hat{y}_t$ estimates $P(y_t{=}1)$ for outcome t. PRISM decomposes this function into five encoding streams, a population retrieval mechanism, and a shared language model decoder, as illustrated in Figure 1. Each stream produces a fixed-length embedding that is projected as a sequence of soft tokens into the language model's embedding space. The following subsections detail each component; prompt design and data construction are deferred to Appendices I and J.

3.2 Structured EHR Encoding

We encode the complete D-dimensional clinical feature vector using a Feature Tokenizer Transformer (FT-Transformer) (Gorishniy et al., 2021), which preserves the numerical fidelity of each clinical measurement by mapping it to an individual learned token rather than serialising it as text.

Concretely, each continuous feature x_j is projected to a d_{model}-dimensional token via a learned per-feature affine transformation $\mathbf{t}_j = x_j\mathbf{w}_j + \mathbf{b}_j$, where $\mathbf{w}_j \in \mathbb{R}^{d_{\text{model}}}$ and $\mathbf{b}_j \in \mathbb{R}^{d_{\text{model}}}$ are feature-specific parameters. Categorical features are embedded through trainable lookup tables. A learnable `[CLS]` token $\mathbf{t}_{\text{cls}}$ is prepended to the resulting sequence $[\mathbf{t}_{\text{cls}}, \mathbf{t}_1, \ldots, \mathbf{t}_D]$, which is processed by an L-layer Transformer encoder with d_{model}=192, H=4 attention heads, and pre-layer normalization (Xiong et al., 2020). The `[CLS]` output is projected through a two-layer feed-forward network (FFN) to produce the EHR embedding:

$$\mathbf{z}_{\text{EHR}} = \text{FFN}\Big(\text{Transformer}^{L}([\mathbf{t}_{\text{cls}}, \mathbf{t}_1, \ldots, \mathbf{t}_D])_{\text{cls}}\Big) \in \mathbb{R}^{512}. \tag{1}$$

To capture short-horizon physiological trends beyond the index measurement, 96-hour prior vital sign histories are summarised per channel into seven aggregate statistics (minimum, maximum, mean, standard

deviation, first, last, and slope), yielding 42 supplementary features concatenated to the base vector before encoding. Missing values are imputed with training-fold medians to prevent information leakage across data splits. The EHR encoder is trained end-to-end during Stage 1 and frozen thereafter.

3.3 ECG Waveform Encoding

The index-visit ECG waveform $\mathbf{X}_{\text{ecg}} \in \mathbb{R}^{12\times T}$ is encoded by D-BETA (Hung et al., 2025), a waveform-language foundation model pre-trained through two complementary objectives: masked modality modelling, which reconstructs masked segments of both the waveform and paired clinical text, and a contrastive objective with a nearest-neighbor negative sampling strategy that aligns waveform and text representations in a shared embedding space. The encoder produces a fixed-dimensional waveform embedding $\mathbf{z}_{\text{ECG}} \in \mathbb{R}^{d_{\text{ecg}}}$.

D-BETA is selected as default ECG encoder for its joint waveform-text pre-training, which organises the embedding space by clinical co-occurrence. This geometric property benefits both patient-level encoding and population memory bank retrieval (see Section 3.5). The encoder remains frozen throughout both training stages, with embeddings pre-computed offline.

3.4 Temporal History Aggregation

Serial ECG changes carry prognostic information beyond any single acquisition (Lehmacher et al., 2020). To incorporate this longitudinal context, we introduce a temporal history aggregator that weights prior-visit ECG embeddings by their recency relative to the index encounter (Figure 1, right panel).

For a patient with N prior acquisitions obtained δ_i hours before the index visit, the aggregator computes normalised exponential recency-decay weights:

$$w_i = \frac{\exp\bigl(-|\lambda|\ \delta_i \,/\, \tau\bigr)}{\sum_{j=1}^{N} \exp\bigl(-|\lambda|\ \delta_j \,/\, \tau\bigr)}, \tag{2}$$

where $\lambda \in \mathbb{R}$ is a single learnable scalar initialised to 1.0 and $\tau = 96$ hours is a fixed normalization constant corresponding to the history window. The absolute value $|\lambda|$ enforces non-negative decay while permitting unconstrained gradient flow. The pooled history representation is then:

$$\mathbf{z}_{\text{pool}} = \sum_{i=1}^{N} w_i\ \mathbf{z}_{\text{ECG}}^{(i)} \ \in\ \mathbb{R}^{d_{\text{ecg}}}. \tag{3}$$

This weighted average is transformed by an adapter network comprising a linear projection, GELU activation, dropout (p=0.1), a second linear projection to 512 dimensions, and layer normalization:

$$\mathbf{z}_{\text{hist}} = \text{LayerNorm}(\mathbf{W}_2\ \text{Dropout}(\text{GELU}(\mathbf{W}_1\ \mathbf{z}_{\text{pool}} + \mathbf{b}_1)) + \mathbf{b}_2) \ \in\ \mathbb{R}^{512}. \tag{4}$$

For patients without prior acquisitions (N=0), a separate learnable no-history embedding initialised from $\mathcal{N}(0, 0.02^2)$ replaces $\mathbf{z}_{\text{hist}}$, allowing the model to distinguish the absence of history from a genuine temporal representation. The entire aggregation module, including λ, is trained during Stage 1 and frozen in Stage 2.

Upon convergence, $|\lambda|$ is directly interpretable as a rate constant whose reciprocal $\tau/|\lambda|$ defines the characteristic timescale over which prior ECG context remains prognostically relevant.

3.5 Population Memory Bank

To ground language model reasoning in empirical population evidence, PRISM constructs a dual-modality memory bank using FAISS (Douze et al., 2024) over per-patient embeddings from the training cohort. After Stage 1 training, the learned EHR and ECG encoders produce fixed-length representations for every training encounter. These embeddings are ℓ_2-normalised and stored in two independent nearest-neighbor indices: an EHR index operating over $\mathbf{z}_{\text{EHR}} \in \mathbb{R}^{512}$ that captures metabolic and organ-function similarity, and an ECG index operating over $\mathbf{z}_{\text{ECG}} \in \mathbb{R}^{d_{\text{ecg}}}$ that captures cardiac phenotypic similarity.

At inference, the query patient's EHR and ECG embeddings are used to retrieve the k nearest neighbors independently from each index using cosine similarity. The top-k retrieved embeddings from each modality are aggregated via mean pooling to produce context vectors:

$$\mathbf{c}_{\text{EHR}} = \frac{1}{k}\sum_{i=1}^{k} \mathbf{z}_{\text{EHR}}^{(\text{nn}_i)} \in \mathbb{R}^{512}, \qquad \mathbf{c}_{\text{ECG}} = \frac{1}{k}\sum_{i=1}^{k} \mathbf{z}_{\text{ECG}}^{(\text{nn}_i)} \in \mathbb{R}^{d_{\text{ecg}}}, \tag{5}$$

where nn_i denotes the i-th nearest neighbor. The neighborhood size k=3 is selected via held-out sensitivity analysis over $k \in \{1, 3, 5\}$ and held fixed across all task groups. To eliminate retrieval latency during training and inference, neighbor assignments are materialised as pre-computed row-aligned lookup tables after Stage 1, reducing each retrieval operation to a single array index. Appendix D provides a complete configuration of the retrieval index and its key details.

The dual-index design reflects a deliberate architectural choice. The EHR index groups patients by metabolic and organ-function profile, while the ECG index groups patients by cardiac phenotype. As the ablation study in Section 4.4 demonstrates, neither index alone provides meaningful benefit; only their combination consistently recovers outcome-concordant neighbors.

3.6 Multimodal Fusion and Language Model Adaptation

Soft-token projection. Each modality embedding is mapped to a sequence of n_m soft tokens in the language model's embedding space through a dedicated two-layer projector:

$$\mathbf{S}_m = \text{Reshape}_{n_m \times d_{\text{llm}}}\Big(\mathbf{W}_m^{(2)}\ \text{GELU}\Big(\mathbf{W}_m^{(1)}\ \text{LayerNorm}(\mathbf{z}_m) + \mathbf{b}_m^{(1)}\Big) + \mathbf{b}_m^{(2)}\Big), \tag{6}$$

where $\mathbf{W}_m^{(1)} \in \mathbb{R}^{1024 \times d_{\text{in}}}$, $\mathbf{W}_m^{(2)} \in \mathbb{R}^{n_m d_{\text{llm}} \times 1024}$, n_m=8 for all five modality streams, and d_{llm}=2560 for MedGemma-4B. The five modality streams are: current ECG ($\mathbf{z}_{\text{ECG}}$), structured EHR ($\mathbf{z}_{\text{EHR}}$), longitudinal ECG history ($\mathbf{z}_{\text{hist}}$), retrieved EHR context ($\mathbf{c}_{\text{EHR}}$), and retrieved ECG context ($\mathbf{c}_{\text{ECG}}$). Appendix C tabulates complete per-modality input and output dimensions for all five projectors.

Placeholder substitution. The tokenised clinical prompt $\mathbf{p}$ contains five designated placeholder tokens that serve as positional anchors within the text sequence. At forward time, each placeholder is replaced in-place by its corresponding $n_m \times d_{\text{llm}}$ soft-token matrix $\mathbf{S}_m$, while all remaining positions are embedded through the standard LLM vocabulary layer. The resulting sequence interleaves 40 modality tokens (5×8) with a variable number of text tokens into a unified embedding tensor $\mathbf{E} \in \mathbb{R}^{S \times d_{\text{llm}}}$ consumed directly by the decoder. This substitution-based alignment enables cross-modal reasoning within the LLM's self-attention layers without architectural modification.

Language model backbone and adaptation. We use MedGemma-4B (Sellergren et al., 2025), a decoder-based four-billion-parameter language model pre-trained on de-identified biomedical corpora. During Stage 1, all LLM parameters are frozen and only the encoders and projectors are optimized. During Stage 2, we apply Low-Rank Adaptation (LoRA) (Hu et al., 2022) to the query, key, value, output, gate, up-projection, and down-projection matrices of every self-attention layer with rank r=16, scaling factor α_{LoRA}=32, and dropout 0.05.

3.7 Training and Inference

PRISM is trained in two stages, each with a distinct set of trainable parameters and a shared composite objective.

Composite objective. Both stages minimise a weighted combination of causal language modelling loss on the answer tokens and a binary classification auxiliary loss:

$$\mathcal{L} = \mathcal{L}_{\text{CLM}} + \alpha\, \mathcal{L}_{\text{BCE}}, \tag{7}$$

where $\mathcal{L}_{\text{CLM}}$ is the standard next-token cross-entropy restricted to the answer span, $\mathcal{L}_{\text{BCE}}$ is binary cross-entropy applied through a linear classification head over the mean-pooled decoder hidden states at all prompt positions, and α=0.1 throughout.

Stage 1: encoder and projector training. The EHR FT-Transformer, the temporal history aggregator (including λ), and the three primary projectors (ECG, EHR, history) are jointly optimized while the D-BETA encoder and MedGemma-4B backbone remain frozen. Training uses AdamW optimizer with learning rate 2×10^{-4}, cosine decay with 100 warm-up steps, gradient clipping at $\|\mathbf{g}\|_2 \leq 1.0$, effective batch size 32 per device (micro-batch 4, gradient accumulation over 8 steps), and BF16 mixed precision on two NVIDIA H100 80 GB GPUs. Upon completion, the population memory bank indices are constructed from the trained encoder outputs.

Stage 2: retrieval-augmented language model adaptation. All Stage 1 parameters are frozen. The LoRA adapters, the two retrieval projectors (initialised from scratch), and the classification head (loaded from the Stage 1 checkpoint) are jointly optimized. The learning rate is reduced to 2×10^{-5}; the micro-batch size is reduced to two to accommodate the expanded five-modality token sequence within the 2,048-token budget. Checkpoints are selected by minimum validation loss on the held-out split. Full hyperparameter specifications are provided in Appendix B.

Inference pathways. At inference, PRISM supports two complementary output pathways evaluated on every task. The *generative pathway* decodes the LLM autoregressively to produce a Yes/No token, yielding a binary prediction reported as $\text{PRISM}_{\text{gen}}$. The *discriminative pathway* applies the linear head to the mean-pooled decoder hidden states, producing a continuous probability estimate $\hat{y} = P(y{=}1) \in [0, 1]$ without autoregressive decoding, reported as $\text{PRISM}_{\text{disc}}$. Both pathways are evaluated across all experiments and reported separately.

4 Results

4.1 Overall Benchmark Performance

All experiments are conducted on the standardized MDS-ED benchmark splits. Following Stage 1 and Stage 2 optimization on 4.5 million training samples, all models are evaluated on a held-out test set of 6,454 patient records. PRISM achieves state-of-the-art performance across all task groups. Table 1 reports AUROC across all four MDS-ED task groups for PRISM and 15 baseline systems spanning five model families. PRISM establishes a new state of the art on every task group: **86.24%** on Diagnosis (1,428 ICD-10 discharge tasks), **93.12%** on Deterioration (6 tasks), **91.79%** on ICU admission (2 tasks), and **92.82%** on Mortality (7 tasks), with the generative and discriminative pathways achieving the best score on complementary task groups. The performance hierarchy across model families mirrors clinical data requirements: single-modality ECG models and zero-shot frontier LLMs operate well below the supervised multimodal tier on Diagnosis tasks, confirming that neither waveform information nor parametric clinical knowledge alone can substitute for task-specific multimodal supervision (Table 1). Among the four task groups, Deterioration, ICU admission, and Mortality predict future adverse events from data available at ED arrival, whereas Diagnosis tasks predict discharge ICD-10 labels assigned at the end of the hospital stay, a temporally forward-looking classification problem grounded in a different clinical signal.

The two prior multimodal systems, MDS-ED (Lopez Alcaraz et al., 2025) and UniPACT (Tang et al., 2026), substantially narrow this gap by fusing ECG representations with clinical context (88.89% and 89.37% overall, respectively). PRISM improves over UniPACT across every task group, with the largest margin on Diagnosis (+2.26 absolute AUROC points), followed by Deterioration (+1.95 pp), ICU admission (+1.29 pp), and Mortality (+1.00 pp). This ordering is consistent with the architectural hypothesis motivating PRISM. The performance gap over prior methods widens precisely on outcomes where laboratory values carry the most discriminative signal and where the textualization bottleneck is most costly.

We estimated 95% confidence intervals via empirical bootstrap (1,000 resamples per task, macro-averaged across tasks within each group). The resulting intervals are 86.24 [85.92, 86.45] for Diagnosis, 92.90 [89.16,

Table 1: **Benchmark results across all 1,443 MDS-ED tasks.** AUROC (%) is macro-averaged within each task group: Diagnosis (1,428 ICD-10 discharge tasks), Deterioration (6 tasks), ICU admission (2 tasks), and Mortality (7 tasks). Single-modality ECG models and zero-shot frontier LLMs operate well below the multimodal tier, confirming that waveform or parametric knowledge alone cannot substitute for task-specific multimodal supervision. PRISM surpasses all prior methods across every task group under both inference pathways. **Bold with underline**: best. †Zero-shot, text-only input.

Method	Tasks · AUROC (%)			
	Diagnosis	Deterioration	ICU	Mortality
ECG Foundation Models · *waveform input*				
HeartBERT (Tahery et al., 2025)	56.35	58.24	57.28	61.11
ECGFM-KED (Tian et al., 2024)	58.75	64.76	61.32	65.76
ESI (Yu et al., 2024)	60.98	63.34	62.52	66.28
ECGFounder (Li et al., 2025)	61.52	69.35	64.13	69.42
Image-Based Models · *rendered ECG input*				
GEM (Lan et al., 2025)	61.52	55.80	60.40	65.70
PULSE (Liu et al., 2024)	58.47	61.10	64.70	68.70
TimeMaster (Zhang et al., 2025)	57.91	60.80	63.80	57.30
ECG-Instruct (M & M, 2025)	47.33	46.00	44.40	43.70
Signal-Based QA Models · *ECG waveform input*				
ECGQA-FSL (Tang et al., 2025)	50.56	50.46	51.06	52.15
Q-HEART (Pham et al., 2025)	50.15	55.42	54.89	56.32
ECG-Chat (Zhao et al., 2025)	49.70	57.54	56.40	55.78
Frontier LLMs† · *text-only input*				
GPT-5† (OpenAI, 2025)	55.68	78.12	70.04	62.27
Gemini-2.5 Pro† (Comanici et al., 2025)	57.82	78.72	70.70	72.38
Multimodal Baselines · *waveform + EHR*				
MDS-ED (Lopez Alcaraz et al., 2025)	82.56	90.70	90.63	91.68
UniPACT (Tang et al., 2026)	83.98	91.17	90.50	91.82
PRISM (Ours) · *waveform + EHR + history + retrieval*				
PRISM$_{gen}$	**86.24**	**92.90**	**91.79**	**92.82**
PRISM$_{disc}$	**86.06**	**93.12**	**91.55**	**92.79**

95.82] for Deterioration, 91.79 [90.92, 92.65] for ICU, and 92.78 [91.98, 93.50] for Mortality. A complete breakdown of per-task performance for Deterioration, ICU, and Mortality, with confidence intervals, is provided in Appendix E.

4.2 Task-Level Analysis

Task-level analysis reinforces the interpretation that PRISM's gains are concentrated where structured clinical data carries the most discriminative weight. Across the full benchmark, the MDS-ED baseline exceeds an AUROC of 80% on 623 of 1,443 tasks, while PRISM surpasses this threshold on 1,145 tasks (79.3%). This represents a significant gain of 522 additional tasks classified at high discriminative performance compared to the MDS-ED baseline, and 277 additional tasks compared to the UniPACT framework.

At the level of ICD chapters (Figure 2d), improvements are consistent across all clinical domains but graded by the degree to which each domain depends on laboratory context. Circulatory system diagnoses, where ECG waveforms already carry substantial discriminative information, show a gain of 1.39%. In contrast, musculoskeletal disorders (+6.33%) and blood and immune system conditions (+6.06%) show the largest improvements, consistent with the fact that these diagnostic categories rely primarily on laboratory values that prior systems discarded through textualization.

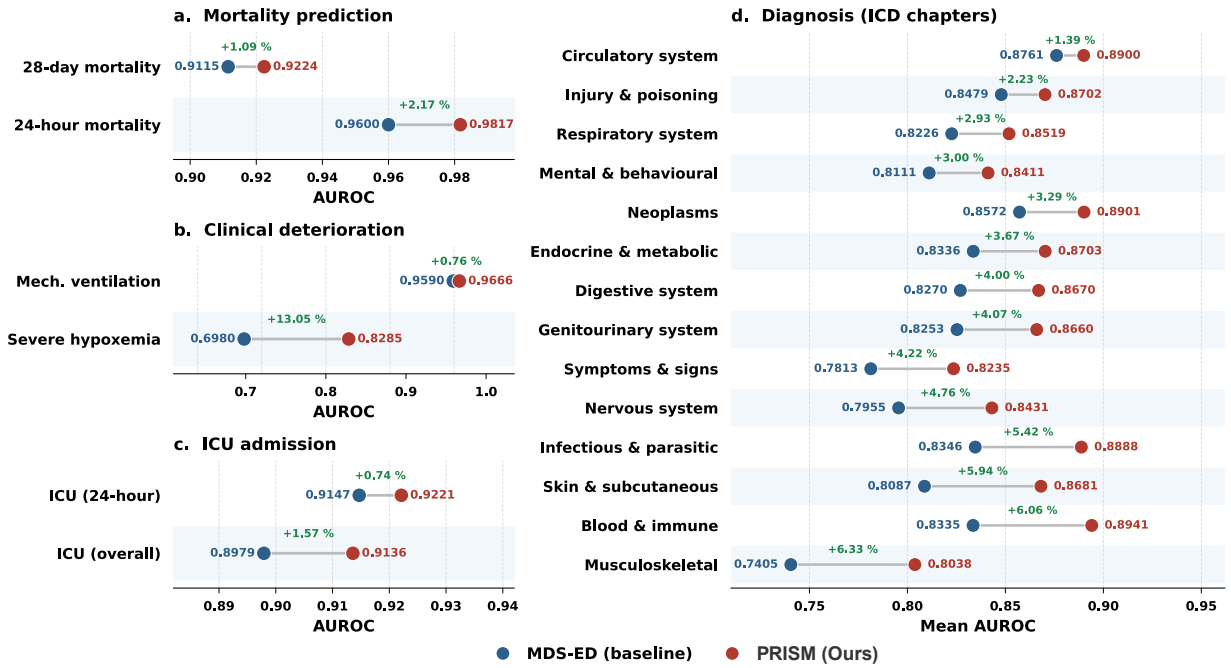

Figure 2: **PRISM performance vs. MDS-ED baseline across clinical prediction categories.** Comparison of AUROC for (a) mortality at multiple horizons, (b) clinical deterioration endpoints, (c) ICU admission, and (d) ICD-chapter-level diagnosis groups. Red and blue markers denote PRISM and MDS-ED, respectively. Green annotations indicate relative AUROC improvement. Gains are largest on laboratory-dependent outcomes (e.g., blood and immune disorders, musculoskeletal conditions) and on deterioration endpoints where metabolic context is mechanistically essential.

Within the Deterioration group, the most striking individual result is on severe hypoxemia, where PRISM achieves an AUROC of 0.8285 compared to 0.6980 for MDS-ED (+13.05%). This is a task for which arterial oxygen and respiratory function measurements provide mechanistically essential context that is unavailable to systems encoding only waveform and demographic features. Appendix G provides a comprehensive view of how these gains distribute across the 1,428 diagnosis tasks and ICD-10 chapters.

4.3 Population Memory Bank Analysis

Figure 3 presents a representative case study evaluating the clinical coherence of the population memory bank. For a test patient assigned ICD-10 label `Z79` (long-term drug therapy) with a positive outcome, the dual-modality index retrieves three training patients whose joint ECG and EHR embeddings are closest in the learned representation space. All three retrieved neighbors share both the discharge label and the positive outcome, with cosine similarities of 96.5%, 76.7%, and 99.5% to the query embedding. This demonstrates that the retrieval mechanism consistently identifies clinically concordant cases in the outcome-relevant embedding geometry, rather than surfacing records with merely superficial feature overlap.

Inspection of key clinical features reveals plausible physiological correspondences: the test patient and all retrieved neighbors share elevated glucose and elevated heart rate as flagged abnormalities, consistent with the metabolic context associated with long-term pharmacological management in this diagnostic category. This concordance provides evidence that Stage 1 training produces a representation geometry in which clinical similarity and prognostic outcome are jointly preserved.

The auditable nature of retrieved evidence, consisting of real patients with known outcomes, enables case-based interpretability absent from purely parametric reasoning. Extended retrieval examples across four clinical phenotypes—metabolic disorder, chronic vascular, endocrine management and vascular risk are provided in Appendix K, confirming consistent outcome-concordant neighborhood recovery and coherent diagnostic enrichment across heterogeneous presentation types.

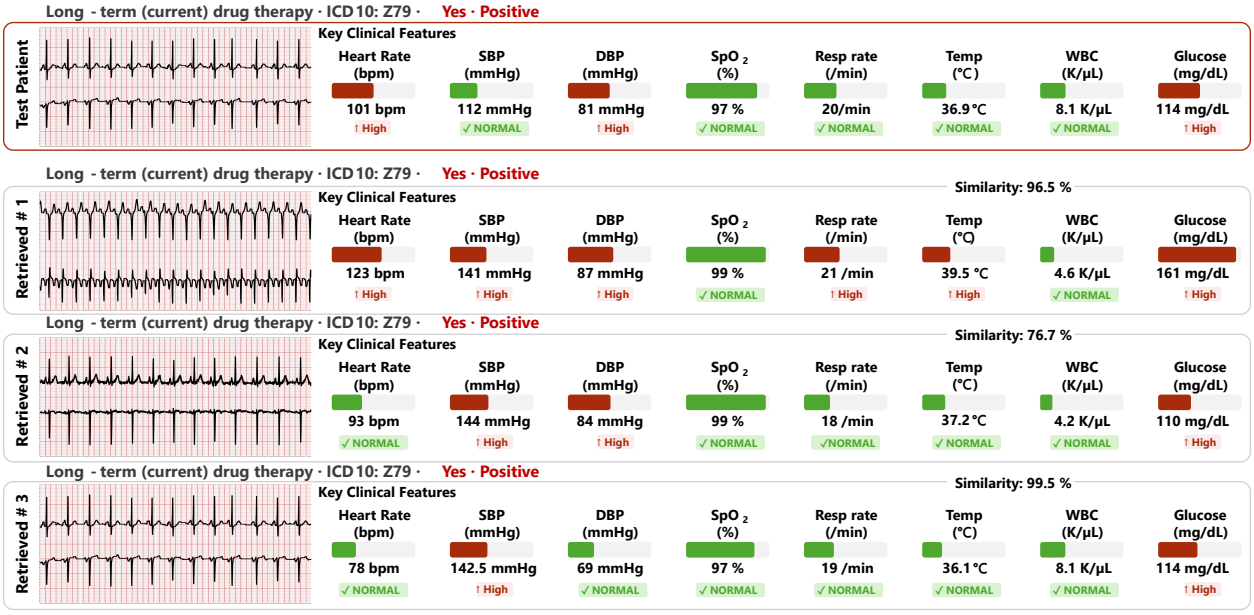

Figure 3: **Qualitative analysis of the population memory bank.** A test patient with ICD-10 label Z79 (long-term drug therapy, positive outcome) and three training-set neighbors retrieved from the dual-modality memory bank. All retrieved patients share the same discharge label and outcome. Cosine similarities to the query embedding (96.5%, 76.7%, 99.5%) and shared clinical abnormalities (elevated glucose, elevated heart rate) demonstrate that the learned embedding geometry jointly preserves clinical similarity and outcome concordance.

Table 2: **Ablation study on PRISM architecture and training configurations.** Each row removes or modifies a single component while holding all others fixed. Performance is reported as AUROC (%). Gen: generative pathway; Disc: discriminative pathway. Top-3 is the neighborhood size used in the full model; Top-1 and Top-5 are alternatives evaluated in sensitivity analysis. **Bold** indicates the best result per column. Disc. and Gen. refer to PRISM$_{disc}$ and PRISM$_{gen}$, respectively.

Ablation / Configuration	Diagnosis		Deterioration		ICU		Mortality	
	Disc	Gen	Disc	Gen	Disc	Gen	Disc	Gen
Encoder Ablations								
w/o EHR encoder	82.22	82.50	91.25	91.61	89.00	89.45	91.06	91.41
Text-serialized EHR (no tokenization)	79.14	80.07	90.19	91.18	86.69	87.86	89.85	90.18
w/o ECG encoder	80.37	80.61	91.39	91.44	87.65	88.34	90.47	90.93
w/o ECG history	82.72	82.92	92.45	91.95	89.41	89.69	91.13	91.56
Memory Bank Ablations								
w/o retrieval context	82.75	82.95	92.36	91.95	89.40	89.71	91.15	91.55
EHR retrieval only	82.75	82.95	92.47	92.25	89.42	89.69	91.15	91.53
ECG retrieval only	82.75	82.94	92.40	92.10	89.41	89.70	91.05	91.50
Neighborhood Size (k)								
$k = 1$	82.88	82.96	92.43	92.23	89.48	89.73	91.09	91.52
$k = 5$	83.76	83.96	92.53	92.13	89.44	89.70	91.11	91.49
Training Ablations								
Frozen retrieval projectors	81.53	81.67	88.88	89.08	89.42	89.66	87.42	87.76
LoRA off (backbone frozen)	68.97	73.26	82.30	84.79	82.26	84.55	86.74	86.17
Full PRISM (k=3)	**86.06**	**86.24**	**93.12**	**92.90**	**91.55**	**91.79**	**92.79**	**92.82**

4.4 Ablation Study

Table 2 presents systematic component removal experiments, revealing that PRISM's performance advantage is not attributable to any single component but emerges from three complementary contributions that are each necessary for the observed gains.

Table 3: **ECG encoder backbone ablation.** All other architectural components and hyperparameters are held fixed; only the waveform encoder varies. Best AUROC (%) across both inference pathways is reported per task group. Δ denotes the gap between D-BETA and the next-best encoder. **Bold**: best per column.

ECG Encoder	Diagnosis	Deterioration	ICU	Mortality
HeartBERT (Tahery et al., 2025)	83.65	91.07	90.79	91.61
ECGFM-KED (Tian et al., 2024)	82.71	90.57	90.27	90.90
ECGFounder (Li et al., 2025)	82.54	91.71	90.60	90.73
D-BETA (PRISM) (Hung et al., 2025)	**86.24**	**93.12**	**91.79**	**92.82**
Δ *vs. next-best*	+2.59	+1.41	+1.00	+1.21

Structured EHR encoding drives diagnostic performance. Removing the FT-Transformer produces the significant regression on Diagnosis AUROC (−3.84 pp Disc, −3.74 pp Gen), while losses on Deterioration and ICU tasks are comparatively modest. This gradient is clinically interpretable: ICD-10 discharge diagnosis is largely a laboratory-phenotyping problem where troponin, creatinine, and glucose are principal discriminators, whereas deterioration outcomes carry partial redundancy with ECG morphology. The result confirms that for a large fraction of clinical outcomes, the structured record is not merely an enrichment of the waveform signal but the primary source of discriminative information.

Numerical tokenization, isolated from the EHR pathway as a whole. Replacing the FT-Transformer soft tokens with text-serialized EHR features, holding the ECG encoder, prompt format, LoRA adapters, retrieval modules, and training setup fixed, underperforms full PRISM on every task group (Disc/Gen): Diagnosis 79.14/80.07 (−6.92/−6.17 pp), Deterioration 90.19/91.18 (−2.93/−1.72 pp), ICU 86.69/87.86 (−4.86/−3.93 pp), Mortality 89.85/90.18 (−2.94/−2.64 pp). It also underperforms *w/o EHR encoder* on every group and pathway, indicating that verbose numeric text is worse than omitting structured input entirely, likely because it crowds out context within a fixed token budget. It further required a 6,500-token context window versus PRISM's 2,048, raising inference time from roughly 4 to 24 hours. These results attribute PRISM's gains to numerical tokenization itself, not merely the presence of structured EHR data.

ECG modalities contribute dissociable and complementary signal. Removing the current-visit ECG encoder degrades Diagnosis by 5.69 pp (Disc) and 5.63 pp (Gen), confirming that waveform representations carry substantial signal complementary to the structured record. The effect is smaller but consistent on ICU (−3.90 pp) and Deterioration (−1.73 pp), reflecting partial redundancy between waveform morphology and laboratory features. Removing the longitudinal ECG history produces a more modest but interpretable degradation: Diagnosis falls by 3.34 pp, concentrated in tasks where disease trajectory rather than acute presentation determines the outcome.

Population retrieval requires dual-modality indexing. Activating either the EHR or ECG memory bank alone yields performance indistinguishable from no retrieval, while combining both produces substantial gains across all task groups. This superadditivity reflects complementary indexing structure: the EHR index groups patients by metabolic profile, the ECG index groups by cardiac phenotype, and only their intersection consistently recovers outcome-concordant neighbors. The neighborhood size exhibits a non-monotonic optimum at k=3, consistent across all task groups: k=1 underestimates the local population distribution, while k=5 dilutes the retrieval signal as boundary neighbors introduce inconsistent outcome labels.

Both training stages are essential. Freezing retrieval projectors degrades Mortality and Diagnosis most severely, revealing that these are the tasks where population evidence carries the most discriminative weight. Disabling LoRA adaptation entirely collapses performance across all task groups, most severely on Diagnosis (−17.09 pp Disc), confirming that a frozen backbone cannot learn to interpret the dense numerical features provided by the structured encoder. The two ablations expose complementary failure modes: retrieval misalignment and encoding rigidity.

ECG encoder selection. Table 3 evaluates three alternative ECG encoder backbones within the PRISM framework, with all other components held fixed. D-BETA outperforms HeartBERT, ECGFM-KED, and ECGFounder on every task group under both inference pathways, with its largest margin on Diagnosis (+2.59 pp), followed by Deterioration (+1.41 pp), precisely the tasks where the population memory bank contributes most. This advantage is not incidental to pre-training scale. D-BETA's joint waveform-text pre-training via masked modality modelling and nearest-neighbor contrastive objectives organises the embedding space by clinical co-occurrence rather than diagnostic label alone. Because PRISM constructs the population memory bank within this same space, a geometrically well-structured encoder improves both individual patient representations and the outcome-concordance of retrieved neighbors simultaneously.

5 Discussion and Conclusion

The results presented in this work challenge a foundational assumption underlying current multimodal clinical LLMs: that structured clinical data can be absorbed through a language model's tokenizer without meaningful information loss. PRISM's advantage over prior methods is largest on outcomes whose discriminative signal resides in continuous-valued laboratory measurements, suggesting that the textualization bottleneck is not a minor approximation but a structural limitation. Diagnosis tasks, which are fundamentally biochemical phenotyping problems, suffer most when numerical fidelity is discarded, and this cost is not recoverable through increased model capacity. The implication extends beyond this benchmark: the dominant paradigm of text-serialized clinical LLMs may systematically underperform on any outcome determined by continuous-valued biomarkers, a category encompassing the majority of acute-care prediction tasks.

The extension of retrieval-augmented generation from text corpora to empirical patient populations represents a conceptual shift in how language models can be grounded for clinical reasoning. The population memory bank surfaces real historical patients with documented outcomes, enabling auditable case-based reasoning that curated knowledge retrieval cannot provide. The ablation finding that dual-modality retrieval is strictly necessary reveals that outcome-concordant neighbor recovery requires physiological similarity across both metabolic and cardiac dimensions simultaneously, suggesting that population retrieval is most valuable in clinical settings where multiple complementary modalities jointly determine outcomes. Unlike parametric predictions that offer no provenance, each PRISM prediction can be accompanied by the specific historical patients that informed it, providing a concrete basis for clinical review aligned with emerging requirements for explainability in AI-assisted decision-making.

Limitations. All experiments use MIMIC-IV from a single institution; multi-site validation is needed to characterise robustness to variation in assay calibration, acquisition hardware, and patient demographics. The population memory bank requires site-specific construction from local historical data, though the two-stage architecture facilitates this without retraining the base encoders. Performance has not been stratified by demographic subgroups, an important direction given documented disparities in emergency department outcomes. Finally, while the soft-token projection architecture is backbone-agnostic, PRISM has been evaluated only with MedGemma-4B.

Conclusion. We introduced PRISM, a multimodal architecture that grounds language model prognostic reasoning in three sources of clinical evidence that prior systems discard: the full numerical fidelity of structured clinical records, the longitudinal context of serial ECG acquisitions, and population-level outcome statistics from historically similar patients. Evaluated across 1,443 binary prediction tasks on the MDS-ED benchmark, PRISM establishes a new state of the art on every task group, with systematic ablations confirming that each component contributes distinctly and that their benefits are complementary. Beyond the specific benchmark results, this work establishes that per-feature numerical encoding, lightweight temporal aggregation, and population-grounded retrieval are viable and effective design principles for multimodal clinical prediction with language models.

References

Amna N Ahmed, Ken Blonde, Daniel Hackam, Alla Iansavichene, and Marko Mrkobrada. Prognostic significance of elevated troponin in non-cardiac hospitalized patients: A systematic review and meta-analysis. *Annals of Medicine*, 46(8):653–663, 2014. doi: 10.3109/07853890.2014.959558.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, and Others. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. URL `https://arxiv.org/abs/2507.06261`.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. 2024.

Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Manh Pham Hung, Aaqib Saeed, and Dong Ma. Boosting masked ECG-text auto-encoders as discriminative learners. In *Forty-second International Conference on Machine Learning*, 2025.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023. doi: 10.1038/s41597-022-01899-x.

Ole Kruse, Niels Grunnet, and Charlotte Barfod. Blood lactate as a predictor for in-hospital mortality in patients admitted acutely to hospital: a systematic review. *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine*, 19:74, 2011. doi: 10.1186/1757-7241-19-74.

Xiang Lan, Feng Wu, Kai He, Qinghao Zhao, Shenda Hong, and Mengling Feng. GEM: Empowering MLLM for grounded ECG understanding with time series and images. *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Jonas Lehmacher, Johannes Tobias Neumann, Nils Arne Sørensen, Alina Goßling, Paul Michael Haller, Tau Sarra Hartikainen, Peter Clemmensen, Tanja Zeller, Stefan Blankenberg, and Dirk Westermann. Predictive value of serial ECGs in patients with suspected myocardial infarction. *Journal of Clinical Medicine*, 9(7):2303, 2020. doi: 10.3390/jcm9072303.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, volume 36, 2023.

Jun Li, Aaron D Aguirre, Valdery Moura Junior, Jiarui Jin, Che Liu, Lanhai Zhong, Chenxi Sun, Gari Clifford, M Brandon Westover, and Shenda Hong. An electrocardiogram foundation model built on over 10 million recordings. *NEJM AI*, 2(7):AIoa2401033, 2025.

Yikuan Li, Shishir Rao, Jose Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Thomas Lukasiewicz, Gholamreza Salimi-Khorshidi, and Kazem Rahimi. BEHRT: Transformer for electronic health records. *Scientific Reports*, 10(7155):1–12, 2020. doi: 10.1038/s41598-020-62922-y.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

Ruoqi Liu, Yuelin Bai, Xiang Yue, and Ping Zhang. Teach multimodal LLMs to comprehend electrocardiographic images. *arXiv preprint arXiv:2410.19008*, 2024.

Juan Miguel Lopez Alcaraz, Hjalmar Bouma, and Nils Strodthoff. Enhancing clinical decision support with physiological waveforms—a multimodal benchmark in emergency care. *Computers in Biology and Medicine*, 192:110196, 2025. doi: 10.1016/j.compbiomed.2025.110196.

Nandakishor M and Anjali M. High-accuracy ecg image interpretation using parameter-efficient lora fine-tuning with multimodal llama 3.2, 2025. URL `https://arxiv.org/abs/2501.18670`.

OpenAI. Openai GPT-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Hung Manh Pham, Jialu Tang, Aaqib Saeed, and Dong Ma. Q-HEART: ECG question answering via knowledge-informed multimodal LLMs. In *Proceedings of the European Conference on Artificial Intelligence (ECAI)*, volume 413 of *Frontiers in Artificial Intelligence and Applications*, pp. 4545–4552, 2025. doi: 10.3233/FAIA251356.

Sushravya Raghunath, Alvaro E Ulloa Cerna, Linyuan Jing, David P vanMaanen, Joshua V Stough, Dustin N Hartzel, Joseph B Leader, H Lester Kirchner, Martin C Stumpe, Ashraf Hafez, Arun Nemani, Tanner Carbonati, Kipp W Johnson, Katelyn Young, Christopher W Good, John M Pfeifer, Aalpen A Patel, Brian P Delisle, Amro Alsaid, Dominik Beer, Christopher M Haggerty, and Brandon K Fornwalt. Prediction of mortality from 12-lead electrocardiogram voltage data using a deep neural network. *Nature Medicine*, 26 (6):886–891, 2020. doi: 10.1038/s41591-020-0870-z.

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. Med-BERT: Pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *npj Digital Medicine*, 4(86):1–13, 2021. doi: 10.1038/s41746-021-00455-y.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cián Hughes, Charles Lau, and others. MedGemma technical report. *arXiv preprint arXiv:2507.05201*, 2025. URL `https://arxiv.org/abs/2507.05201`.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023. doi: 10.1038/s41586-023-06291-2.

Gary B Smith, David R Prytherch, Paul Meredith, Paul E Schmidt, and Peter I Featherstone. The ability of the National Early Warning Score (NEWS) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admission, and death. *Resuscitation*, 84(4):465–470, 2013. doi: 10.1016/j.resuscitation.2012.12.016.

Saedeh Tahery, Fatemeh Hamid Akhlaghi, and Termeh Amirsoleimani. Heartbert: A self-supervised ecg embedding model for efficient and effective medical signal analysis. 2025. URL `https://arxiv.org/abs/2411.11896`.

Jialu Tang, Tong Xia, Yuan Lu, Cecilia Mascolo, and Aaqib Saeed. Electrocardiogram–language model for few-shot question answering with meta learning. In *Proceedings of the Conference on Health, Inference, and Learning (CHIL)*, volume 287 of *Proceedings of Machine Learning Research*, pp. 89–104, 2025.

Jialu Tang, Tong Xia, Yuan Lu, and Aaqib Saeed. Unipact: A multimodal framework for prognostic question answering on raw ecg and structured ehr. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 22537–22541, 2026. doi: 10.1109/ICASSP55912.2026.11461333.

Yuanyuan Tian, Zhiyuan Li, Yanrui Jin, Mengxiao Wang, Xiaoyang Wei, Liqun Zhao, Yunqing Liu, Jinlei Liu, and Chengliang Liu. Foundation model of ECG diagnosis: Diagnostics and explanations of any form and rhythm on ECG. *Cell Reports Medicine*, 5:101875, 2024. doi: 10.1016/j.xcrm.2024.101875.

Dung-Jang Tsai, Yu-Sheng Lou, Chin-Sheng Lin, Wen-Hui Fang, Chia-Cheng Lee, Ching-Liang Ho, Chih-Hung Wang, and Chin Lin. Mortality risk prediction of the electrocardiogram as an informative indicator of cardiovascular diseases. *Digital Health*, 9:20552076231187247, 2023.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tie-Yan Liu. On layer normalization in the transformer architecture. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 10524–10533. PMLR, 2020. URL `https://proceedings.mlr.press/v119/xiong20b.html`.

Han Yu, Peikun Guo, and Akane Sano. ECG semantic integrator (ESI): A foundation ECG model pretrained with LLM-enhanced cardiological text. *arXiv preprint arXiv:2405.19366*, 2024.

Junru Zhang, Lang Feng, Xu Guo, Yuhan Wu, Yabo Dong, and Duanqing Xu. TimeMaster: Training time-series multimodal LLMs to reason via reinforcement learning. *arXiv preprint arXiv:2506.13705*, 2025.

Yubao Zhao, Jiaju Kang, Tian Zhang, Puyu Han, and Tong Chen. ECG-Chat: A large ECG-language model for cardiac disease diagnosis. In *IEEE International Conference on Multimedia and Expo (ICME)*, 2025. doi: 10.1109/ICME59968.2025.11209476.

Yinghao Zhu, Changyu Ren, Zixiang Wang, Xiaochen Zheng, Shiyun Xie, Junlan Feng, Xi Zhu, Zhoujun Li, Liantao Ma, and Chengwei 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 (CIKM)*, 2024. doi: 10.1145/3627673.3679582.

A Appendix

A. Training Hyperparameters 15

B. Soft-Token Projector Specifications 16

C. Population Memory Bank Configuration 16

D. Per-Task AUROC: Deterioration, ICU, and Mortality 17

E. Diagnosis AUROC by ICD-10 Chapter 17

F. AUROC Distribution Across Diagnosis Tasks 17

G. Cohort and Dataset Details 17

H. Clinical Prompt Design and Representative Examples 18

I. Training Data Construction and Prompt Generation 21

J. Population Memory Bank: Retrieval Across Clinical Phenotypes 23

B Training Hyperparameters

PRISM is trained in two stages with distinct sets of trainable parameters. Stage 1 jointly optimises the EHR FT-Transformer, the temporal history aggregator (including the learnable decay scalar λ), and the three primary soft-token projectors (ECG, EHR, history), while the MedGemma-4B backbone and D-BETA encoder remain frozen. Stage 2 freezes all Stage 1 parameters and jointly trains the LoRA adapters, the two retrieval projectors (initialised from scratch), and the classification head (loaded from the Stage 1 checkpoint). Table 4 reports all hyperparameters for both stages.

Table 4: Training hyperparameters for PRISM Stage 1 and Stage 2.

Hyperparameter	**Stage 1**	**Stage 2**
LLM backbone	MedGemma-4B (frozen)	MedGemma-4B + LoRA
LoRA rank / scaling	–	16 / 32
LoRA dropout	–	0.05
LoRA target modules	–	q, k, v, o, gate, up, down
EHR encoder d_{model}	192	frozen
EHR encoder layers / heads	3 / 4	frozen
EHR embedding d_{repr}	512	frozen
History decay λ (init)	1.0	frozen
Soft tokens per modality n_m	8	8
Projector hidden dim	1,024	1,024
Classification weight α	0.1	0.1
Memory bank k	–	3
Effective batch size	32 (micro 4 $\times$ accum 8)	2
Learning rate	2×10^{-4}	2×10^{-5}
LR schedule	Cosine (100 warm-up)	Cosine
Gradient clipping	$\|\mathbf{g}\|_2 \leq 1.0$	$\|\mathbf{g}\|_2 \leq 1.0$
Max sequence length	2,048 tokens	2,048 tokens
Mixed precision	BF16	BF16
Hardware	2× H100 80 GB	2× H100 80 GB

C Soft-Token Projector Specifications

Each modality embedding is mapped to a sequence of eight soft tokens in the language model's embedding space through a two-layer feed-forward projector with layer normalization, GELU activation, and a final linear expansion (Equation 6). All five projectors share an identical architecture and differ only in input dimensionality and training stage. The three primary projectors (current ECG, EHR, ECG history) are trained during Stage 1. The two retrieval projectors are initialised from scratch and trained during Stage 2 alongside the LoRA adapters. Table 5 summarises the specifications.

Table 5: Soft-token projector specifications. All projectors have a hidden dimension of 1,024 and produce an output shape of $8 \times 2{,}560$ (MedGemma-4B embedding size).

Modality stream	**Input dim.**	**Stage**
Current-visit ECG	d_{ecg}	1
Structured EHR	512	1
Longitudinal ECG history	512	1
Retrieved EHR context	512	2
Retrieved ECG context	d_{ecg}	2

D Population Memory Bank Configuration

The population memory bank comprises two independent nearest-neighbor indices constructed over per-patient embeddings from the 116,186 training-fold encounters after Stage 1 training. The EHR index op-

erates over 512-dimensional embeddings and retrieves patients with similar metabolic and organ-function profiles. The ECG index operates over d_{ecg}-dimensional embeddings and retrieves patients with similar cardiac phenotypes. Both indices use ℓ_2-normalised vectors with cosine similarity as the distance metric.

To eliminate retrieval latency during training and inference, neighbor assignments for all encounters are materialised as pre-computed row-aligned lookup tables immediately after index construction. At runtime, each retrieval operation reduces to a single array index per modality. The top-k retrieved embeddings from each index are independently mean-pooled to produce context vectors before projection into the language model's token space. Table 6 summarises the configuration.

Table 6: Population memory bank configuration for PRISM.

Property	**EHR Index**	**ECG Index**
Vector dimension	512	d_{ecg}
Distance metric	Cosine similarity	Cosine similarity
Indexed encounters	116,186	116,186
Neighbors retrieved (k)	3	3
Context aggregation	Mean pooling	Mean pooling
Index construction	After Stage 1	After Stage 1

E Per-Task AUROC: Deterioration, ICU, and Mortality

Table 7 reports per-task AUROC for all 15 non-diagnosis tasks on the held-out test set, together with 95% confidence intervals computed via empirical bootstrap (1,000 resamples per task). PRISM achieves AUROC above 0.95 on five tasks, including cardiac arrest, mechanical ventilation, and within-1-day mortality, and maintains AUROC above 0.90 on 13 of 15 tasks. Performance decreases as the mortality time horizon extends, consistent with the increasing contribution of chronic disease burden over longer intervals relative to the acute physiological state captured in the emergency department window.

F Diagnosis AUROC by ICD-10 Chapter

Table 8 aggregates AUROC across all 1,428 diagnosis tasks by ICD-10 clinical chapter under the generative pathway, sorted by mean AUROC. PRISM attains mean AUROC above 0.87 in 10 of 23 chapters and above 0.80 in all chapters. The highest-performing chapters are those in which specific laboratory abnormalities are directly pathognomonic: obstetric conditions (Chapter O, mean 0.98), infectious diseases (Chapter A, 0.92), and circulatory disorders (Chapter I, 0.89). Performance is comparatively lower in chapters dominated by symptom-based or musculoskeletal diagnoses, where the primary discriminating information resides in clinical presentation rather than the structured biochemical record.

G AUROC Distribution Across Diagnosis Tasks

Table 9 stratifies all 1,428 diagnosis tasks by AUROC range under the discriminative pathway. The majority of tasks (78.4%) exceed AUROC 0.80, and more than one third (35.4%) exceed 0.90. Only eight tasks (0.6%) fall below 0.60, all involving symptom-based or psychiatric diagnoses where the discriminating signal is largely absent from the structured laboratory and waveform record.

H Cohort and Dataset Details

PRISM is trained and evaluated on the MDS-ED benchmark (Lopez Alcaraz et al., 2025), derived from the MIMIC-IV and MIMIC-IV-ECG databases (Johnson et al., 2023). This dataset constitutes the largest publicly available source of linked ECG waveforms and structured EHR data for emergency department

Table 7: Per-task AUROC for all Deterioration, ICU, and Mortality tasks on the held-out test set, with 95% bootstrap confidence intervals (1,000 resamples). The stronger pathway per task is **bolded**.

Group	Endpoint	Test n	$\text{PRISM}_{\text{disc}}$ [95% CI]	$\text{PRISM}_{\text{gen}}$ [95% CI]
Deterioration	Cardiac arrest	68	0.9654 [0.9187, 0.9965]	**0.9857** [0.9627, 1.0000]
	Inotrope administration	27	**0.9560** [0.8765, 1.0000]	0.9313 [0.8076, 1.0000]
	Mechanical ventilation	437	0.9617 [0.9438, 0.9759]	**0.9670** [0.9517, 0.9789]
	ECMO cannulation	18	**0.9383** [0.8051, 1.0000]	**0.9383** [0.7916, 1.0000]
	Vasopressor initiation	125	**0.9265** [0.8717, 0.9701]	0.9238 [0.8742, 0.9672]
	Severe hypoxemia	54	**0.8278** [0.7069, 0.9293]	**0.8285** [0.7129, 0.9286]
ICU	Admission within 24h	1,588	0.9202 [0.9061, 0.9329]	**0.9226** [0.9100, 0.9349]
	Admission during stay	1,888	0.9117 [0.8987, 0.9239]	**0.9134** [0.9002, 0.9249]
Mortality	During hospitalisation	18	**1.0000** [1.0000, 1.0000]	**1.0000** [1.0000, 1.0000]
	Within 1 day	97	**0.9817** [0.9574, 0.9983]	0.9802 [0.9540, 0.9979]
	Within 7 days	245	0.9405 [0.9101, 0.9669]	**0.9428** [0.9086, 0.9680]
	Within 28 days	535	0.9191 [0.8939, 0.9414]	**0.9224** [0.8993, 0.9442]
	Within 90 days	922	0.8968 [0.8762, 0.9150]	**0.8971** [0.8772, 0.9158]
	Within 180 days	1,268	**0.8755** [0.8573, 0.8950]	0.8754 [0.8561, 0.8928]
	Within 365 days	1,813	0.8766 [0.8625, 0.8915]	**0.8775** [0.8629, 0.8926]
Deterioration macro-avg (6 tasks)			0.9288 [0.8958, 0.9541]	0.9290 [0.8916, 0.9582]
ICU macro-avg (2 tasks)			0.9159 [0.9070, 0.9247]	0.9179 [0.9092, 0.9265]
Mortality macro-avg (7 tasks)			0.9273 [0.9191, 0.9347]	0.9278 [0.9198, 0.9350]

encounters. We adopt the standardised 20-fold stratified splits defined by the MDS-ED protocol to ensure direct comparability with all existing baselines.

To represent the information available to a clinician at the point of disposition, all data is extracted from a 90-minute window beginning at ED arrival. Every encounter must have a raw 12-lead ECG, at least one structured laboratory measurement, and a complete discharge record with ICD-10-CM codes. The feature vector (D=470) includes demographics, biometrics, triage vitals, and laboratory panels; previous medications and free-text complaints are excluded.

Table 10 summarises the cohort statistics and experimental partitions. Table 11 describes the four clinical task groups predicted by PRISM.

I Clinical Prompt Design and Representative Examples

Each training and inference instance is presented to the language model as a structured clinical question-answering prompt. The prompt opens with modality placeholder tokens, each replaced at forward time by an $8 \times 2{,}560$ soft-token matrix from the corresponding projector. It then provides a natural-language role description, a patient summary comprising demographic information, a compact time-series of vital sign measurements recorded in the 96-hour peri-presentation window, a stratified abnormality summary of all 470 structured clinical features, and a prior ECG history listing. The prompt closes with a binary prediction question instructing the model to answer with Yes or No.

When a patient has no prior ECG acquisitions within the 96-hour window, the prompt contains four modality tokens. When prior recordings exist, a fifth token carrying the aggregated longitudinal ECG representation is inserted, resulting in a five-token prompt variant. The two boxes below illustrate representative prompts from the test set.

Table 8: Diagnosis AUROC by ICD-10 chapter ($\text{PRISM}_{\text{gen}}$), sorted by mean AUROC. N: number of ICD-10 codes per chapter.

ICD-10 Chapter	**Ch.**	***N***	**Mean**	**Range**
Pregnancy, childbirth, puerperium	O	3	0.9825	0.9475–1.0000
Certain infectious diseases	A	15	0.9189	0.8493–0.9835
Blood and immune disorders	D	50	0.8952	0.6600–1.0000
Poisoning and toxic effects	T	43	0.8888	0.6317–0.9889
Circulatory system	I	181	0.8900	0.7128–1.0000
Neoplasms	C	48	0.8886	0.2800–1.0000
External causes (V)	V	38	0.8838	0.7376–0.9698
Factors influencing health	Z	166	0.8788	0.5091–1.0000
Infectious and parasitic (B)	B	37	0.8766	0.7381–0.9835
External causes (W)	W	11	0.8732	0.7784–0.9705
Skin and subcutaneous	L	25	0.8681	0.6562–0.9835
Endocrine, nutritional, metabolic	E	107	0.8703	0.6875–1.0000
Genitourinary system	N	39	0.8660	0.6730–0.9952
Digestive system	K	119	0.8670	0.6942–1.0000
External causes (Y)	Y	27	0.8577	0.7427–0.9917
Injury and trauma	S	43	0.8517	0.6317–0.9877
Respiratory system	J	71	0.8519	0.5091–0.9905
Nervous system	G	67	0.8431	0.6159–0.9738
Mental and behavioral	F	71	0.8411	0.5102–0.9982
Congenital malformations	Q	3	0.8411	0.7842–0.9224
Eye and ear	H	24	0.8232	0.6311–0.9506
Symptoms, signs, abnormal findings	R	171	0.8235	0.5091–0.9961
Musculoskeletal system	M	68	0.8038	0.4201–1.0000
Overall (1,428 tasks)		**1,428**	**0.8624**	0.2800–1.0000

Table 9: Distribution of per-task AUROC across all 1,428 diagnosis tasks ($\text{PRISM}_{\text{disc}}$). Cumulative percentage is computed from the highest bin downward.

AUROC Range	**Tasks**	**Proportion**	**Cumulative**
≥ 0.90	506	35.4%	35.4%
0.80–0.89	613	42.9%	78.4%
0.70–0.79	260	18.2%	96.6%
0.60–0.69	41	2.9%	99.4%
< 0.60	8	0.6%	100%
Total	**1,428**	**100%**	

Table 10: MDS-ED cohort and partition statistics.

Metric	**Value**
Unique patients	71,098
ED visits	121,195
ECG-linked encounters	129,095
Training samples (task instances)	4.5M
Training set	116,186
Validation set	6,455
Test set	6,454

Table 11: Summary of the 1,443 binary clinical tasks predicted by PRISM.

Group	**Tasks**	**Examples / Endpoints**
Diagnosis	1,428	Full ICD-10 nosology (circulatory, musculoskeletal, etc.)
Deterioration	6	Intubation, cardiac arrest, vasopressors, severe hypoxemia
ICU Admission	2	Overall ICU admission, admission within 24 hours
Mortality	7	Time horizons: 24h, 1d, 7d, 28d, 90d, 180d, 365d

I.1 Box 1: Positive Case Without Prior ECG History

74-year-old male · 24-hour mortality · Ground truth: Yes

`[EHR] [Retrieved EHR] [ECG] [Retrieved ECG]`

You are a critical care physician with access to the patient's complete clinical record, current 12-lead ECG waveform, and Electronic Health Record (EHR). Your task is to predict whether a patient will experience mortality based on the ECG, the full 470-feature EHR summary, the vital signs time-series, and population evidence retrieved from similar patients.

Patient: 74-year-old, Black/African, male.
Biometrics: not recorded.

CRITICAL FINDINGS:
- Base excess: low at −20 mmol/L (normal −2.0–3.0 mmol/L).
- Glucose: critically elevated at 274 mg/dL (normal 70.0–100.0 mg/dL; 307→274 ↓, slope −0.107/hr).
- INR: elevated at 1.3 (normal <1.1).
- Lactate: critically elevated at 10 mmol/L (normal <2.0 mmol/L; 9.9→10 ↑, slope +0.0101/hr).
- pCO_2: critically elevated at 64 mmHg (normal 35.0–45.0 mmHg).
- PT: elevated at 15.1 sec (normal 11.0–13.5 sec).
- PTT: critically elevated at 60.5 sec (normal 25.0–35.0 sec).
- WBC: critically elevated at 15.6×10^3/µL (normal $4.5–11.0\times10^3$/µL).

WITHIN NORMAL LIMITS:
Creatinine; Fibrinogen; Ionized calcium; Hematocrit; Hemoglobin; Platelets; RDW; RBC; BUN.

Prior ECG History (within 96h):
No prior ECG recordings identified in the 96h window.

Will the patient die within 24 hours? Answer strictly with Yes or No.

Box 2: Positive Case With Prior ECG History

80-year-old female · ESI Level 3 · Two prior ECGs · 24-hour mortality · Ground truth: Yes

`[EHR] [Retrieved EHR] [ECG History] [ECG] [Retrieved ECG]`

You are a critical care physician with access to the patient's complete clinical record, current 12-lead ECG waveform, and Electronic Health Record (EHR). Your task is to predict whether a patient will experience mortality based on the ECG, the full 470-feature EHR summary, the vital signs time-series, and population evidence retrieved from similar patients.

Patient: 80-year-old, White, female. Acuity: ESI Level 3 (Urgent).
Biometrics: not recorded.

CRITICAL FINDINGS:
- AST: elevated at 44 U/L (normal 10.0–40.0 U/L).
- Creatinine: critically elevated at 2 mg/dL (normal 0.6–1.2 mg/dL).
- Eosinophils: critically low at 0% (normal 1.0–4.0%).
- Glucose: elevated at 125 mg/dL (normal 70.0–100.0 mg/dL).
- Lactate: critically elevated at 5.1 mmol/L (normal <2.0 mmol/L).
- Lymphocytes: low at 14% (normal 20.0–40.0%).
- Neutrophils: elevated at 85% (normal 50.0–70.0%).
- Phosphate: elevated at 5.5 mg/dL (normal 2.5–4.5 mg/dL).
- Potassium: elevated at 5.2 mEq/L (normal 3.5–5.0 mEq/L).
- Troponin T: critically elevated at 0.11 ng/mL (normal <0.01 ng/mL).
- BUN: critically elevated at 39 mg/dL (normal 7.0–20.0 mg/dL).

WITHIN NORMAL LIMITS:
ALT; Albumin; Alkaline phosphatase; Bands; Basophils; Bicarbonate; Total bilirubin; Total calcium; Chloride; Hematocrit; Hemoglobin; Magnesium; Platelets; RDW; RBC; Sodium; WBC.

Prior ECG History (within 96h):
• 0.6h before current acquisition
• 0.4h before current acquisition

Will the patient die within 24 hours? Answer strictly with Yes or No.

J Training Data Construction and Prompt Generation

Each training and evaluation instance presented to PRISM is a structured clinical question-answering record derived from the MDS-ED benchmark (Lopez Alcaraz et al., 2025). This appendix describes the end-to-end pipeline that transforms a raw MDS-ED task entry into the enriched prompt consumed by the MedGemma-4B backbone. The pipeline bridges the tabular MDS-ED data and the soft-token multimodal prompt format described in Sections 3.5 and 3.7; representative prompt instantiations are shown in Appendix I.

J.1 Source Data and Task Records

The 1,443 binary prediction tasks are sourced from the MDS-ED benchmark, which pairs each clinical encounter with an ordered set of question-and-answer records of the form *"Will the patient [outcome]? Answer: Yes / No"*. Each record contains a unique encounter identifier that maps to (i) a row in the MDS-ED feature CSV providing all 470 structured EHR features, (ii) a raw 12-lead ECG waveform path in MIMIC-IV-ECG, and (iii) a longitudinal patient entry in the 96-hour time-series store containing timestamped vital-sign measurements and a list of prior ECG acquisitions. No additional outcome annotation was performed; ground-truth binary labels are taken verbatim from the published benchmark to preserve strict comparability with all previously reported baselines on identical data partitions.

J.2 Population Statistics and Z-Score Normalization

Before any prompt is constructed, per-feature population statistics (mean and standard deviation) are computed over all numeric EHR columns from the 116,186 training-fold encounters (stratification folds 0–17 as defined by the MDS-ED protocol (Lopez Alcaraz et al., 2025)). These statistics serve two distinct roles. First, training-fold column medians are used for missing-value imputation in the FT-Transformer input vector (Section 3.2), preventing evaluation-fold leakage into the structured encoder. Second, they define the reference distribution for z-score-based abnormality flagging in the text prompt: a measurement is designated CRITICAL when it either falls outside the published clinical reference interval (Table 12) or yields a population z-score of magnitude greater than 2.0, whichever criterion is the more inclusive. Severity is

Table 12: Clinical reference intervals used for EHR abnormality flagging. Open bounds are indicated by a dash. Applied in conjunction with a population z-score threshold of $|z| > 2.0$ (Section J.2).

Cat.	**Analyte**	**Lo**	**Hi**	**Unit**	**Cat.**	**Analyte**	**Lo**	**Hi**	**Unit**
Vitals	Temperature	36.0	38.0	°C	Vitals	Systolic BP	90	140	mmHg
	Heart rate	60	100	bpm		Diastolic BP	60	90	mmHg
	Resp. rate	12	20	br/min		O_2 sat.	95	–	%
Haem.	WBC	4.5	11.0	10^3/µL	Haem.	Neutrophils	50	70	%
	hemoglobin	12.0	17.5	g/dL		Lymphocytes	20	40	%
	hematocrit	36	50	%		Eosinophils	1.0	4.0	%
	Platelets	150	400	10^3/µL		Basophils	0	1.0	%
	RBC	3.8	5.8	M/µL		Bands	–	5.0	%
	RDW	11.5	14.5	%					
Coag.	INR	–	1.1	–	Coag.	PTT	25	35	sec
	PT	11.0	13.5	sec		Fibrinogen	200	400	mg/dL
Chem.	Sodium	136	145	mEq/L	Chem.	Potassium	3.5	5.0	mEq/L
	Chloride	98	106	mEq/L		Bicarbonate	22	29	mEq/L
	BUN	7.0	20	mg/dL		Creatinine	0.6	1.2	mg/dL
	Glucose	70	100	mg/dL		Total calcium	8.5	10.5	mg/dL
	Phosphate	2.5	4.5	mg/dL		Magnesium	1.7	2.2	mg/dL
	Albumin	3.4	5.4	g/dL					
Hepatic	ALT	7.0	56	U/L	Hepatic	AST	10	40	U/L
	Alk. phos.	44	147	U/L		Total bili.	–	1.2	mg/dL
Cardiac	Troponin T	–	0.01	ng/mL	Cardiac	Total CK	–	200	U/L
	CK-MB	–	5.0	ng/mL					
Metabolic	Lactate	–	2.0	mmol/L	Metabolic	pCO_2	35	45	mmHg
	Base excess	−2	2.0	mmol/L		pH	7.35	7.45	–
	Ion. Ca	1.12	1.32	mmol/L					

further stratified as *elevated* or *low* when the fractional deviation from the nearest bound is at most 30%, and as *critically elevated* or *critically low* when the deviation exceeds 30%.

J.3 Language Model Prompt

The prompt for each instance is assembled from five ordered components.

Modality token prefix. Between four and five placeholder tokens are prepended to the prompt, one per active modality: the current-visit ECG waveform (`<ecg>`), the FT-Transformer EHR encoding (`<ehr>`), the retrieved EHR population context (`<retrieved_ehr>`), the aggregated longitudinal ECG history (`<ecg_history>`, inserted only when the patient has at least one prior acquisition within the 96-hour window), and the retrieved ECG population context (`<retrieved_ecg>`). At forward time, each placeholder is replaced in-place by an $8 \times 2{,}560$ soft-token matrix from the corresponding projector (Table 5). Encounters without prior ECG recordings receive the four-token variant; all others receive the five-token variant.

Role and task block. A fixed natural-language role description identifies the model as a critical care physician and specifies the prediction target for the current instance (e.g., *"Will the patient be diagnosed with 'Atrial fibrillation'?"*). This framing is derived verbatim from the binary question string in the MDS-ED task record and is standardised uniformly across all 1,443 tasks.

Patient summary block. The patient summary comprises four programmatically generated sub-sections: (i) a demographics and biometrics header containing age, sex, self-reported ethnicity, body-mass index, weight, height, and the Emergency Severity Index acuity level when recorded; (ii) a time-stamped vital-sign table spanning up to ten representative rows from the 96-hour window (when more than ten measurements are available, the earliest and nine most recent observations are retained); (iii) a stratified EHR abnormality summary classifying all 470 features under CRITICAL FINDINGS or WITHIN NORMAL LIMITS according to the criteria in Section J.2, including observed value, unit, reference interval, and trend direction; and (iv) a listing of the number and relative timestamps (in hours before the index ECG) of prior ECG acquisitions within the 96-hour window.

J.4 Memory Bank Lookup Integration

After Stage 1 training, the FT-Transformer and D-BETA encoder produce fixed-length embeddings for all 116,186 training-fold encounters. These embeddings populate two independent nearest-neighbor indices, one per modality, as described in Appendix D. During the prompt-construction pass that precedes Stage 2 training and evaluation, each encounter is assigned its top-k nearest neighbors from each index via a single lookup against pre-computed neighbor tables materialised immediately after Stage 1. The resulting neighbor row indices are embedded directly in the serialized data record so that the Stage 2 data loader can reconstruct the corresponding embedding vectors at batch time without incurring query latency during distributed GPU training.

J.5 Train, Validation, and Test Handling

Validation and test encounters (stratification folds 18 and 19, respectively) are processed through an identical pipeline, with one critical exception: their memory bank queries target the *training-fold* index exclusively, ensuring that no ground-truth outcome information from held-out partitions enters the retrieved population context.

K Population Memory Bank: Retrieval Across Clinical Phenotypes

Figure 4 extends the single-encounter retrieval analysis of Section 4.3 to four clinically distinct test-set encounters, each representative of a different phenotypic category in the MDS-ED cohort. For each encounter, the dual-modality memory bank retrieves k=50 nearest neighbors from the training population; ground-truth binary outcome labels and discharge diagnosis codes are aggregated by direct counting.

The four cases demonstrate systematic outcome enrichment aligned with clinical context. In the high-acuity metabolic case (Case A), where mechanical ventilation, ICU admission, 28-day mortality, and cardiac arrest are all ground-truth positive, retrieved neighbors exhibit prevalence rates of 30.0%, 40.0%, 26.0%, and 24.0% for these outcomes respectively, and their most frequent discharge diagnoses reflect the chronic disease burden associated with severe metabolic derangement. In the chronic vascular case (Case B), cardiac arrest prevalence among neighbors reaches 50.0% (25/50), the highest across all four cases, consistent with the haemodynamic vulnerability encoded in both the ECG and EHR embeddings. The endocrine management case (Case C) is notable for an ICU admission ground truth of absent despite a vasopressor-present label; the retrieved neighborhood captures this dissociation, with inflammatory rather than haemodynamic diagnoses predominating. The vascular risk case (Case D) exhibits the lowest adverse-outcome prevalence across all endpoints, with the neighborhood diagnosis distribution shifting toward sepsis-spectrum codes, consistent with the vascular-inflammatory overlap captured in this phenotype's embedding geometry.

These cases confirm that outcome-concordant neighborhood recovery and clinically coherent diagnostic enrichment are systematic properties of the dual-modality memory bank across phenotypically heterogeneous encounters, rather than an artefact of any single case type.

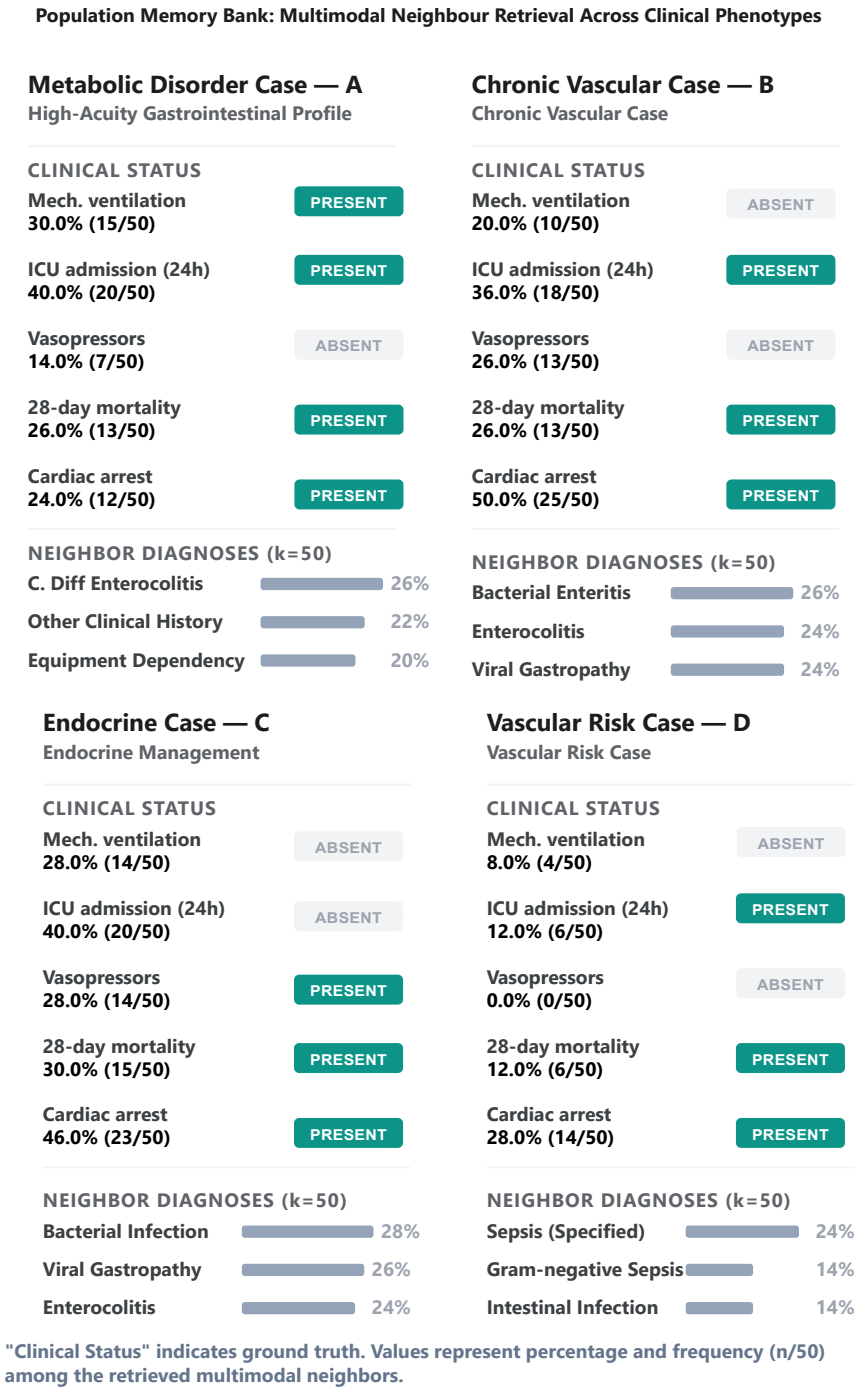

Figure 4: **Population memory bank retrieval across four clinical phenotypes (Cases A–D).** For each test encounter, k=50 nearest neighbors are retrieved from the dual-modality memory bank; clinical status (PRESENT/ABSENT) reflects the ground-truth outcome label of the index patient, and percentages with counts (n/50) report outcome prevalence among retrieved neighbors.