# SEPSISWEAVE: DYNAMIC ORGAN INTERACTION MODELING WITH ACTIVE EVIDENCE ACQUISITION FOR EARLY SEPSIS PREDICTION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Early sepsis often emerges through subtle and evolving interactions across organ systems. Most predictors, however, collapse these dynamics into a single latent representation and return an opaque risk score, obscuring the contributions of individual organs and their dependencies. We introduce **SepsisWeave**, an interactive framework that formulates early sepsis prediction as a sequential evidence-acquisition task. At its core, **DynaOrg** independently encodes the temporal states of six organ systems and learns patient-specific, time-varying interactions among them, exposing local and interaction-adjusted deterioration risks as structured, queryable evidence. We further develop a three-stage training strategy combining interaction-format warm-up, teacher-guided strategy imitation, and student-sampled preference optimization to enable valid evidence acquisition. On MIMIC-IV, SepsisWeave consistently outperforms direct single-turn prediction from standardized summaries across multiple LLM backbones, yielding a mean relative AUROC improvement of $10.12\%$. Preliminary external validation on eICU-CRD further suggests improved robustness under distribution shift. These results suggest that explicitly modeling and selectively acquiring cross-organ evidence provides an effective and inspectable alternative to fixed end-to-end inference.

## 1 INTRODUCTION

Sepsis, defined as life-threatening organ dysfunction caused by a dysregulated host response to infection, accounts for up to half of all hospital deaths and incurs over $24 billion in annual healthcare costs in the United States (Singer et al., 2016; Liu et al., 2014). Its early signs often do not manifest as pronounced abnormalities in a single organ, but rather as evolving patterns of coupled deterioration across multiple organ systems (Feng et al., 2025). However, most existing approaches to early sepsis prediction directly map the available clinical variables to a final risk score (Islam et al., 2023; Nemati et al., 2018). Such end-to-end black-box paradigms entangle local organ states and cross-organ information within a shared latent representation, leaving two important challenges unresolved: (i) *the complementary predictive value of dynamic cross-organ associations* is neither explicitly modeled nor readily isolated from local organ dynamics; and (ii) *the contribution and integration of different sources of evidence* remain insufficiently transparent to clinicians, making the decision process difficult to inspect.

Existing approaches to early sepsis prediction range from clinical scores such as SOFA and qSOFA to machine learning models trained on structured EHR variables (Singer et al., 2016; Desautels et al., 2016). More recent RNN- and Transformer-based methods capture longitudinal dependencies among these measurements (Lauritsen et al., 2020; Tang et al., 2024), while organ-structured and graph-based approaches further incorporate relationships among organ systems (Yin et al., 2025; Feng et al., 2025). In most cases, however, cross-organ information remains embedded within a single end-to-end prediction pathway rather than being exposed as a separable prediction component, limiting their ability to effectively inform organ deterioration. For example, renal function may appear relatively stable based on local state alone, while subtle but concordant changes across the cardiovascular and coagulation systems may indicate an elevated risk of subsequent renal deterioration (Figure 1). Motivated by this gap, we introduce **DynaOrg**, a structured model that ex-



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.

plicitly modeling patient-specific and time-varying cross-organ influence. DynaOrg disentangles organ-specific evidence from cross-organ evidence, producing both a *local deterioration risk* and an *interaction-adjusted risk*. By exposing this explicit cross-organ message-passing pathway, DynaOrg transforms otherwise entangled organ interactions into structured evidence that can be independently evaluated, clinically interpreted, and queried on demand by a downstream decision-making agent.

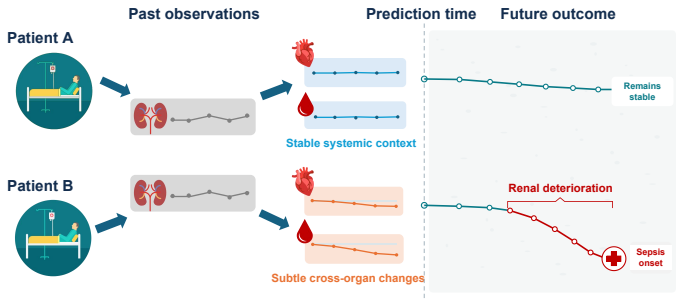



Figure 1: Motivating clinical scenario. Patients A and B exhibit similar local renal trajectories, but subtle cardiovascular and coagulation changes distinguish their systemic contexts. These differences precede divergent future trajectories.

Such on-demand access is important because cross-organ evidence is not equally informative for every patient. When local deterioration is already pronounced, it may provide limited additional value; when abnormalities are subtle or discordant across organs, it may reveal emerging systemic decline. More broadly, the evidence-acquisition process for early sepsis prediction should adapt to the patient's current state and the evidence already observed, rather than follow a fixed information pathway. However, most existing early sepsis predictors still rely on predefined inputs and fixed inference pathways (Islam et al., 2023; Nemati et al., 2018), while recent adaptive approaches mainly select additional measurements or trigger further diagnosis under predefined conditions (Yin et al., 2024; Shashikumar et al., 2025). This motivates a distinct question: can an agent autonomously decide *what* evidence to acquire, *when* to acquire it, and *when* enough evidence has been gathered to make a prediction?

To address this question, we introduce **SepsisWeave**, an interactive environment that formulates early sepsis prediction as a sequential evidence-acquisition process. This selective access mechanism encourages the agent to acquire clinically relevant evidence only when needed, reducing unnecessary information retrieval and improving inference efficiency. At the same time, each prediction is accompanied by a transparent acquisition trajectory that records which evidence sources were consulted and how newly acquired evidence influenced the decision, making the reasoning process more inspectable. By temporally grounding every query and exposing heterogeneous evidence sources through structured interfaces, SepsisWeave provides a controlled framework for studying the trade-off between predictive performance, evidence-acquisition cost, and decision interpretability.

Our main contributions are summarized as follows:

- We propose **DynaOrg** to explicitly capture dynamic cross-organ interactions, isolating interaction-adjusted deterioration risks as structured, queryable evidence.
- We introduce **SepsisWeave**, framing early sepsis prediction as active sequential evidence acquisition with transparent, inspectable decision trajectories.
- We demonstrate an average 10.12% AUROC gain on MIMIC-IV (Johnson et al., 2023) and strong out-of-distribution robustness on eICU-CRD (Pollard et al., 2018), while showing that our three-stage training effectively transfers acquisition strategies to smaller LLMs.

## 2 Related Work

**Organ-Aware Sepsis Prediction.** Sepsis is a systemic inflammatory response syndrome, and timely recognition and treatment are critical, with previous studies reporting a 4% reduction in survival odds for each hour of delay (Liu et al., 2017). Clinical screening tools such as qSOFA (Singer et al., 2016), MEWS (Subbe et al., 2001), NEWS (Smith et al., 2013), and SIRS (Bone et al., 1992) are widely used for sepsis recognition, but primarily capture existing physiological abnormalities rather than predict sepsis before its onset. With the growing availability of EHR data, deep learning methods have been widely investigated for early sepsis prediction (Islam et al., 2023; Shashikumar

et al., 2021), modeling clinical measurements as time series with RNNs, LSTMs, or Transformers to capture longitudinal dependencies (Bedoya et al., 2020; Tang et al., 2024). Recent approaches further incorporate unstructured clinical notes (Amrollahi et al., 2021; Goh et al., 2021; Shashikumar et al., 2025), yet typically encode all clinical variables into a unified risk representation, entangling organ-specific dynamics with cross-organ information. Recent graph-based time-series models have explored dynamic graphs to capture time-varying dependencies among clinical variables, such as RAINDROP (Zhang et al., 2022). In sepsis prediction, related studies have further incorporated organ-specific clinical calculators into dynamic temporal graphs for early prediction (Yin et al., 2025), while temporal cross-organ graphs have also been used to characterize sepsis subphenotypes (Feng et al., 2025). Distinct from these approaches, DynaOrg explicitly separates local organ deterioration from interaction-adjusted risk, exposing the incremental contribution of dynamic cross-organ information as structured evidence for selective querying in SepsisWeave.

**Active Clinical Evidence Acquisition.** Active information acquisition studies how a policy can sequentially acquire information and decide when to stop while balancing predictive utility and acquisition cost (Ji & Carin, 2007; Shim et al., 2018; Ma et al., 2018; Li & Oliva, 2021). In medicine, this framework has been applied to patient-specific symptom and test selection (Li et al., 2026), while clinical agents extend it to adaptive questioning, simulated consultations, test recommendation, and EHR retrieval (Li et al., 2024; Tu et al., 2025; Schmidgall et al., 2026; Ren et al., 2025; Shi et al., 2024; Wu et al., 2026). These studies also show that effective evidence gathering is distinct from diagnostic reasoning and should be evaluated separately (Long & Wei, 2026). For early sepsis prediction, SepsisLab actively selects raw clinical measurements to reduce uncertainty, whereas COMPOSER-LLM invokes note-based evidence within a predefined risk interval (Yin et al., 2024; Shashikumar et al., 2025). SepsisWeave instead formulates early sepsis prediction as a sequential evidence-acquisition task in which an agent adaptively selects among multiple complementary clinical tools, including independently queryable cross-organ evidence derived by DynaOrg, and decides when to stop and predict. This design enables controlled evaluation of both the predictive value of each evidence source and the agent's ability to acquire and use it.

# 3 METHOD

## 3.1 PROBLEM FORMULATION

We study early sepsis prediction from longitudinal ICU records. For patient $i$ at hourly prediction anchor $t$, let $\mathcal{O} = \{\text{cardiovascular}, \text{respiratory}, \text{renal}, \text{liver}, \text{coagulation}, \text{neurologic}\}$ denote the six organ systems, and let $\mathbf{X}_{i,t} = \{\mathbf{X}^o_{i,0:t}\}_{o\in\mathcal{O}}$ denote the organ-partitioned clinical history available up to $t$. Given the first Sepsis-3 onset time $\tau_i$, the prediction target is

$$y_{i,t} = \mathbb{I}(t < \tau_i \leq t + H),$$

indicating whether sepsis occurs within the next $H$ hours. We use $H = 4$ hours in the primary setting and exclude anchors at or after sepsis onset.

## 3.2 DYNAORG: DYNAMIC ORGAN INTERACTION MODELING

In this section, we propose DynaOrg, which models organ-specific status and cross-organ interactions to forecast near-term organ deterioration (Figure 2).

### 3.2.1 FROM LOCAL ORGAN STATES TO INTERACTION-ADJUSTED RISK

For each organ $o \in \mathcal{O}$, the clinical observation history $\mathbf{X}^o_{i,t}$ is mapped to an organ-specific representation $\mathbf{h}^o_{i,t} = \mathcal{E}_o\left(\mathbf{X}^o_{i,t}\right) + \mathbf{u}_o$, where $\mathcal{E}_o$ is a temporal encoder and $\mathbf{u}_o$ is a learned organ identity embedding. In parallel, shared clinical features $\mathbf{C}_{i,t}$ are encoded as $\mathbf{c}_{i,t} = \mathcal{E}_c(\mathbf{C}_{i,t})$. The local state of organ $o$ is then constructed as

$$\mathbf{l}^o_{i,t} = \mathrm{LN}\left(\mathbf{W}_v\mathbf{h}^o_{i,t} + \mathbf{W}_{cv}\mathbf{c}_{i,t}\right), \tag{1}$$

where $\mathrm{LN}(\cdot)$ denotes Layer Normalization.

To model dynamic dependencies across organs, DynaOrg parameterizes directed interaction weights $A^{o\leftarrow j}_{i,t}$ from source organ $j$ to target organ $o$ ($j \neq o$). We first compute an unnormalized interaction

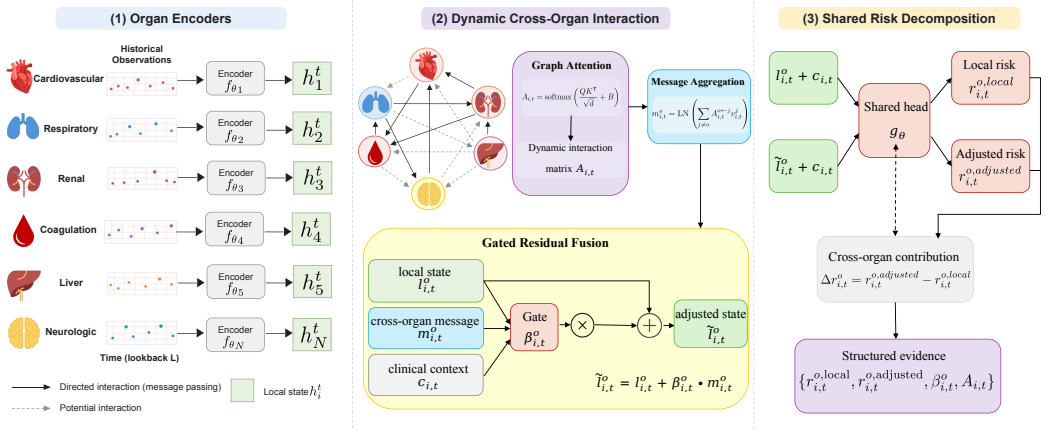



Figure 2: Overview of **DynaOrg**. Organ-specific encoders derive local representations from longitudinal histories, while a patient-specific dynamic graph models directed cross-organ interactions through message passing and gated fusion. A shared head $g_\theta$ outputs local and interaction-adjusted deterioration risks, whose difference captures cross-organ contributions as structured evidence for downstream agents.

score

$$s_{i,t}^{o\leftarrow j} = \frac{(\mathbf{q}_{i,t}^{o})^\top \mathbf{k}_{i,t}^{j}}{\sqrt{d_a}} + \phi_e\left(\mathbf{h}_{i,t}^{o} \| \mathbf{h}_{i,t}^{j}\right),$$

where $\mathbf{q}_{i,t}^{o}$ and $\mathbf{k}_{i,t}^{j}$ are linear projections of local representations, and $\phi_e$ is a pairwise compatibility network. Normalizing these scores via Softmax across all source organs $j \neq o$ yields the dynamic interaction matrix $\mathbf{A}_{i,t}$. The aggregated cross-organ message $\mathbf{m}_{i,t}^{o}$ for target organ $o$ is given by

$$\mathbf{m}_{i,t}^{o} = \mathrm{LN}\left(\sum_{j\neq o} A_{i,t}^{o\leftarrow j}\mathbf{v}_{i,t}^{j}\right),$$

where $\mathbf{v}_{i,t}^{j} = \mathbf{W}_v\mathbf{h}_{i,t}^{j} + \mathbf{W}_{cv}\mathbf{c}_{i,t}$.

To modulate the influence of cross-organ evidence, an organ-specific gate $\beta_{i,t}^{o} = \sigma\left(\phi_\beta\left(\mathbf{l}_{i,t}^{o} \| \mathbf{m}_{i,t}^{o} \| \mathbf{c}_{i,t}\right)\right)$ regulates the latent update, producing the adjusted representation

$$\widetilde{\mathbf{l}}_{i,t}^{o} = \mathrm{LN}\left(\mathbf{l}_{i,t}^{o} + \beta_{i,t}^{o}\phi_m\left(\mathbf{m}_{i,t}^{o} - \mathbf{l}_{i,t}^{o}\right)\right). \tag{2}$$

Finally, a shared prediction head $g_\theta$ computes both pre-interaction and interaction-adjusted risks:

$$r_{i,t}^{o,\mathrm{local}} = \sigma\left(g_\theta\left(\mathbf{l}_{i,t}^{o}, \mathbf{c}_{i,t}\right)\right), \qquad r_{i,t}^{o,\mathrm{adjusted}} = \sigma\left(g_\theta\left(\widetilde{\mathbf{l}}_{i,t}^{o}, \mathbf{c}_{i,t}\right)\right). \tag{3}$$

The shift in risk $\Delta r_{i,t}^{o} = r_{i,t}^{o,\mathrm{adjusted}} - r_{i,t}^{o,\mathrm{local}}$ isolates the predictive impact of cross-organ pathways. The tuple $(r_{i,t}^{o,\mathrm{local}}, r_{i,t}^{o,\mathrm{adjusted}}, \beta_{i,t}^{o}, \mathbf{A}_{i,t})$ forms the structured evidence exposed to downstream agents.

### 3.2.2 Training Objective

Let $z_{n,o} \in \{0,1\}$ be the binary deterioration label within the forecast horizon, with mask $M_{n,o}$ indicating observation validity (Appendix B.1). The overall loss $\mathcal{L}_{\mathrm{DynaOrg}}$ combines focal classification loss ($\mathcal{L}_{\mathrm{wors}}$), Huber regression loss for value trends ($\mathcal{L}_\Delta$), and interaction entropy regularization ($\mathcal{L}_{\mathrm{ent}}$):

$$\mathcal{L}_{\mathrm{DynaOrg}} = \mathcal{L}_{\mathrm{wors}}^{\mathrm{adjusted}} + \lambda_\Delta\mathcal{L}_\Delta^{\mathrm{adjusted}} + \lambda_{\mathrm{local}}\left(\mathcal{L}_{\mathrm{wors}}^{\mathrm{local}} + \lambda_\Delta\mathcal{L}_\Delta^{\mathrm{local}}\right) + \lambda_{\mathrm{ent}}\mathcal{L}_{\mathrm{ent}}. \tag{4}$$

The adjusted pathway is the primary objective, while auxiliary local supervision preserves an independently predictive pre-interaction reference. After joint training, we freeze the remaining DynaOrg components and refine only the latent message transformation $\phi_m$ with $\mathcal{L}_{\mathrm{wors}}^{\mathrm{adjusted}}$, constraining subsequent risk changes to the explicit cross-organ pathway.

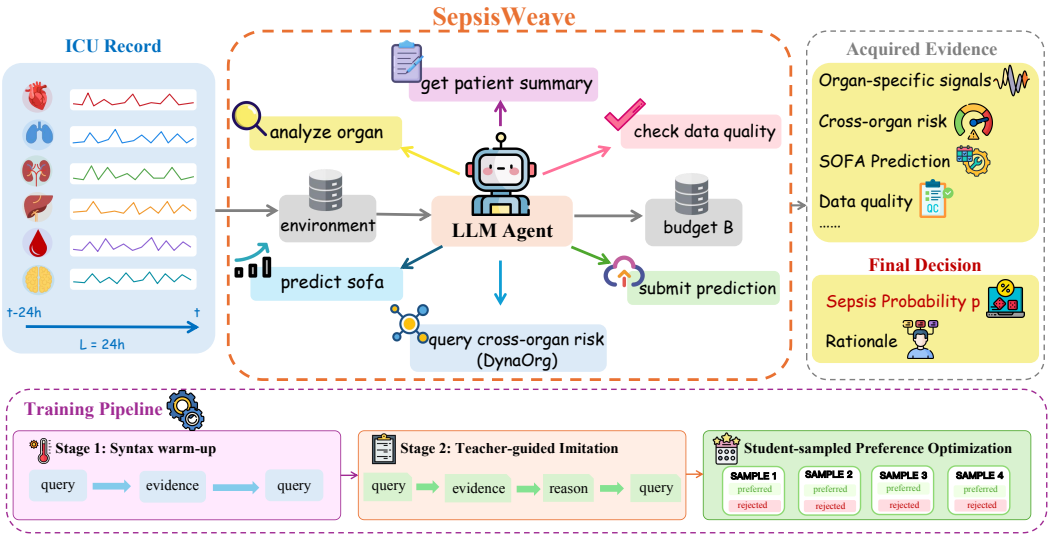



Figure 3: Overview of **SepsisWeave**. Starting from an anchor-bounded ICU record, an LLM agent sequentially selects clinically structured interfaces to acquire evidence, and terminates by submitting a sepsis probability. The acquisition policy is trained progressively through three stages, encouraging accurate, and efficient evidence acquisition.

### 3.3 SepsisWeave: Interactive Evidence Acquisition

Here we introduce SepsisWeave, which formulates early sepsis prediction as a sequential evidence-acquisition process. As shown in figure 3, for each patient–anchor pair $(i, t)$, an agent interacts with the environment through a set of comprehensive, queryable interfaces to acquire sufficient evidence for sepsis prediction.

**Environment Formulation.** At step $k$, the agent receives the current observable evidence set $\mathcal{V}_k$ (with $\mathcal{V}_0$ initialized with the patient's simple context) and selects an action $a_k$. The system executes the corresponding interface $\mathcal{T}_{a_k}$ on the anchor-bounded record $\mathbf{X}_{i,t}$ to return new evidence $e_k$:

$$e_k = \mathcal{T}_{a_k}(\mathbf{X}_{i,t}), \qquad \mathcal{V}_{k+1} = \mathcal{V}_k \cup \{e_k\}.$$

The interaction terminates when the agent invokes the prediction interface $a_{\text{stop}}$, producing a full trajectory:

$$\tau = (\mathcal{V}_0, a_1, e_1, \mathcal{V}_1, a_2, e_2, \ldots, a_T, e_T, \mathcal{V}_T, a_{\text{stop}}) .$$

**Clinically Structured Interfaces.** SepsisWeave provides interfaces for patient-level summarization, organ-specific assessment, DynaOrg-based cross-organ risk, SOFA-component forecasting, and data-quality assessment. The agent terminates through `submit_prediction`, which returns a sepsis probability $\hat{p}_{i,t} \in [0, 1]$ and an evidence-grounded rationale. Among them, `query_cross_organ_risk` explicitly exposes DynaOrg's local and interaction-adjusted deterioration risks together with cross-organ interaction signals. Full interface definitions and output schemas are provided in Appendix C.

### 3.4 Learning Evidence Acquisition

Leveraging the interactive environment and the resulting trajectories, we can train a model into an effective decision-making policy $\pi_\theta$ capable of adaptive clinical evidence acquisition and sepsis prediction through following stages.

#### 3.4.1 Two-Stage Supervised Fine-Tuning

**Stage 1: Syntax & Execution Warm-Up.** This stage teaches the model valid API schemas, argument generation, and environment interaction mechanics. We construct synthetic trajectories where

the agent queries all interfaces in a fixed canonical order before submitting a prediction. Training on runtime-aligned trajectory prefixes, we optimize the action-level cross-entropy loss:

$$\mathcal{L}_{\mathrm{SFT}}^{(1)} = -\mathbb{E}_{(h_k,a_k)\sim\mathcal{D}_1}\left[\sum_k \omega_k^{(1)} \log \pi_\theta(a_k \mid h_k)\right], \tag{5}$$

where loss is computed over target action tokens and normalized per action.

**Stage 2: Teacher-Guided Strategy Imitation.** This stage trains the agent to make patient-conditioned evidence requests using trajectories generated by a stronger teacher policy. We collect teacher rollouts and filter them against quality checks, including schema validity, budget constraints, absence of duplicate calls, and ground-truth prediction accuracy.

For each accepted prefix, Stage 2 jointly supervises the teacher's reasoning trace $r_k$ and the subsequent action $a_k$:

$$\mathcal{L}_{\mathrm{SFT}}^{(2)} = \mathbb{E}_{(h_k,r_k,a_k)\sim\mathcal{D}_2}\left[\ell_{\mathrm{act}}(a_k \mid h_k) + \alpha\, \ell_{\mathrm{reason}}(r_k \mid h_k)\right], \tag{6}$$

where $\ell_{\mathrm{act}}$ and $\ell_{\mathrm{reason}}$ denote length-normalized token cross-entropies for actions and reasoning tokens, respectively.

### 3.4.2 Student-Sampled Preference Optimization

While SFT teaches valid execution and expert strategies, it provides limited supervision for distinguishing better and worse decisions made by the student policy itself. We therefore sample multiple trajectories from the Stage-2 policy for each patient context and evaluate them using

$$R(\tau) = \begin{cases} -(\hat{p}_\tau - y)^2 - \lambda_{\mathrm{cost}} q(\tau) + \lambda_{\mathrm{cross}} \mathbb{I}_{\mathrm{cross}}(\tau), & \text{if } V(\tau) = 1, \\ -C_{\mathrm{inv}} - \lambda_{\mathrm{inv}} N_{\mathrm{invalid}}(\tau), & \text{otherwise}, \end{cases} \tag{7}$$

where $q(\tau)$ measures visible tool-output tokens relative to full-evidence access, $\mathbb{I}_{\mathrm{cross}}(\tau)$ indicates successful acquisition of cross-organ evidence, and $V(\tau)$ requires valid execution and grounded evidence. Thus, prediction quality dominates the ranking of valid trajectories, while evidence cost encourages selective acquisition and a hard penalty separates invalid trajectories.

For each context, the highest- and lowest-reward trajectories form a preference pair $(\tau^+, \tau^-)$ when their reward difference exceeds a fixed margin. Because action lengths and trajectory lengths vary, we define an action-normalized trajectory score:

$$S_\pi(\tau) = \frac{1}{T_\tau} \sum_{k=1}^{T_\tau} \frac{1}{|a_k|} \sum_{t=1}^{|a_k|} \log \pi(a_{k,t} \mid h_k, a_{k,<t}), \tag{8}$$

where only agent-generated action tokens are scored and observations remain context. We optimize

$$\mathcal{L}_{\mathrm{DPO}} = -\mathbb{E} \log \sigma\left(\beta\left[S_{\pi_\theta}(\tau^+) - S_{\pi_{\mathrm{ref}}}(\tau^+) - S_{\pi_\theta}(\tau^-) + S_{\pi_{\mathrm{ref}}}(\tau^-)\right]\right), \tag{9}$$

where $\pi_{\mathrm{ref}}$ is the frozen Stage-2 checkpoint. This objective aligns the agent toward accurate, valid, and evidence-efficient interaction strategies. More details are shown in Appendix G.

## 4 Experiments

### 4.1 Experimental Setup

**Datasets and Task Setup.** We evaluate our framework on MIMIC-IV for internal development and eICU-CRD for external validation. Adult ICU stays are split at the patient level, where early sepsis detection is framed as a sliding 4-hour window prediction task from ICU admission until first Sepsis-3 onset or discharge. The internal test set consists of 994 positive and 2,982 sampled negative anchors; external evaluation follows the exact same partitioning and sampling protocols. Details of sepsis cohort construction and variable harmonization are provided in Appendix A.

Table 1: Early sepsis prediction on MIMIC-IV and eICU-CRD. LLM-based methods use the same backbone under single-summary (BASE-S) and SepsisWeave (OURS) settings. $\Delta$ denotes the relative AUROC improvement of OURS over BASE-S. Top-1/2 results are highlighted in **purple** / pink.

| Model | Setting | MIMIC-IV | | | | eICU-CRD | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | AUROC | $\Delta$ (%) | F1 | Recall | AUROC | $\Delta$ (%) | F1 | Recall |
| *Clinical scores* | | | | | | | | | |
| NEWS | – | 0.535 | – | 0.366 | 0.549 | 0.523 | – | 0.344 | 0.500 |
| MEWS | – | 0.545 | – | 0.375 | 0.562 | 0.511 | – | 0.345 | 0.530 |
| qSOFA | – | 0.540 | – | 0.378 | 0.566 | 0.517 | – | 0.353 | 0.550 |
| SIRS | – | 0.544 | – | 0.370 | 0.554 | 0.487 | – | 0.323 | 0.470 |
| *Deep learning baselines* | | | | | | | | | |
| LSTM | – | 0.579 | – | 0.388 | 0.582 | 0.524 | – | 0.362 | 0.581 |
| GRU | – | 0.626 | – | 0.411 | 0.617 | 0.534 | – | 0.370 | 0.610 |
| DFSP | – | 0.694 | – | 0.494 | 0.741 | 0.545 | – | 0.371 | 0.621 |
| *LLM-based methods* | | | | | | | | | |
| GPT-5.6-Sol | BASE-S | 0.654 | | 0.483 | 0.760 | 0.778 | | 0.577 | 0.827 |
| | OURS | 0.695 | *+6.27* | 0.513 | 0.802 | 0.812 | *+4.37* | 0.611 | 0.800 |
| Gemini-3.8-Flash | BASE-S | 0.731 | | **0.526** | 0.811 | 0.821 | | 0.660 | **0.907** |
| | OURS | **0.747** | *+2.19* | 0.521 | **0.835** | **0.842** | *+2.56* | **0.672** | 0.890 |
| Baichuan4 | BASE-S | 0.487 | | 0.219 | 0.236 | 0.474 | | 0.213 | 0.225 |
| | OURS | 0.551 | *+13.14* | 0.355 | 0.508 | 0.562 | *+18.57* | 0.462 | 0.517 |
| DeepSeek-V4-Flash-0731 | BASE-S | 0.591 | | 0.409 | 0.665 | 0.616 | | 0.423 | 0.710 |
| | OURS | 0.680 | *+15.06* | 0.461 | 0.703 | 0.641 | *+4.06* | 0.451 | 0.898 |
| DeepSeek-V4-Pro-0813 | BASE-S | 0.639 | | 0.428 | 0.753 | 0.598 | | 0.404 | 0.660 |
| | OURS | 0.709 | *+10.95* | 0.496 | 0.769 | 0.636 | *+6.35* | 0.486 | 0.825 |
| GLM-5.3 | BASE-S | 0.602 | | 0.462 | 0.773 | 0.615 | | 0.449 | 0.829 |
| | OURS | 0.681 | *+13.12* | 0.516 | 0.752 | 0.647 | *+5.20* | 0.464 | 0.777 |

**Baselines.** We compare against the following groups of methods:

- **Clinical-score comparators.** We use the widely used clinical calculators (i.e., **NEWS**, **MEWS**, **qSOFA**, **SIRS**) to build the sepsis prediction models.
- **Deep learning models. GRU** (Cho et al., 2014) and **LSTM** (Hochreiter & Schmidhuber, 1997) encode the anchor-bounded clinical time series and produce a final sepsis probability. **Double Fusion Sepsis Predictor (DFSP)** (Duan et al., 2023) is an early sepsis prediction baseline that combines convolutional and recurrent representations with dynamic clinical features to enhance representation learning.
- **LLM evidence settings. BASE-S** is a single-turn baseline provided with a comprehensive, standardized summary of the clinical information available up to the prediction anchor, including longitudinal trends, organ scores, treatments, and derived signals. **BASE-F** receives all available tool outputs simultaneously and therefore represents an upper-bound evidence-access setting without selective acquisition. **OURS** interacts with SepsisWeave and decides which evidence interfaces to query before prediction. These settings use the same underlying LLM and final output schema.
- **Trajectory-trained agents.** We evaluate Qwen3.5-4B (Qwen Team, 2026) and Qwen3.5-9B after canonical interaction warm-up (STAGE 1), teacher-guided strategy imitation (STAGE 1+2), and preference optimization (STAGE 1+2+DPO) (Rafailov et al., 2023).

**Implementation Details.** DynaOrg and all organ-specific predictors are trained on MIMIC-IV training split and held frozen throughout policy training and evaluation. We train Qwen3.5-4B and Qwen3.5-9B using QLoRA across three stages: Stage 1 samples 6,000 prefixes from 3,000 synthesized canonical trajectories (33,000 runtime-aligned prefixes); Stage 2 learns trajectories from DeepSeek-V4-Flash-0731 (DeepSeek-AI, 2026a), retaining 989 trajectories (7,346 reasoning–action prefixes) post-filtering; Stage 3 constructs 800 DPO preference pairs per backbone by scoring four rollouts per patient context using our trajectory reward and pairing the extreme trajectories.

**Evaluation Metrics.** We evaluate system performance using Area Under the Receiver Operating Characteristic Curve (AUROC) as the primary ranking metric, alongside F1 score and Recall. For agent-based methods, we additionally report the *average number of visible tool-output tokens* to quantify evidence-access efficiency. For each patient anchor, we count the tokens in all tool outputs exposed before the final prediction and average this quantity across anchors using a consistent estimator.

### 4.2 Experimental Results of Sepsis Prediction

**Predictive Performance.** Table 1 shows that interaction with SepsisWeave consistently improves LLM-based prediction over the single-summary setting on MIMIC-IV. Across the reported backbones, OURS achieves relative AUROC improvements by an average of $\mathbf{10.12}\%$, accompanied by improvements in F1 and Recall.

The cross-dataset results further demonstrate the robustness of SepsisWeave under distribution shift. We observe that pretrained LLM priors may contribute to the stronger cross-dataset generalization of LLM-based methods on eICU-CRD. Building on this transferable foundation, SepsisWeave further improves prediction through selective acquisition of patient-specific structured evidence. For example, SepsisWeave increases Baichuan4 AUROC from $0.474$ to $0.562$, together with substantial gains in F1 and Recall over BASE-S. More qualitative analysis and ablation studies are deferred to Appendix F.

**Token Efficiency–Performance Trade-off.** Figure 4 examines whether the predictive gains arise simply from exposing the LLM to more structured evidence. Compared with BASE-F, OURS reduces the average number of visible tool-output tokens from 2,092 to 1,777 for DeepSeek-V4-Flash-0731, a reduction of $15.1\%$, while increasing AUROC from $0.669$ to $0.680$. For DeepSeek-V4-Pro-0813, selective acquisition reduces the corresponding token count from 2,079 to 1,698 ($18.3\%$) while retaining comparable predictive performance ($0.709$ versus $0.714$ AUROC).

These results show that the benefit of SepsisWeave does not depend on exhaustively exposing the model to every available tool output. By selecting patient-relevant interfaces, the agent removes a substantial amount of potentially redundant evidence while preserving, and in one case improving, predictive performance.

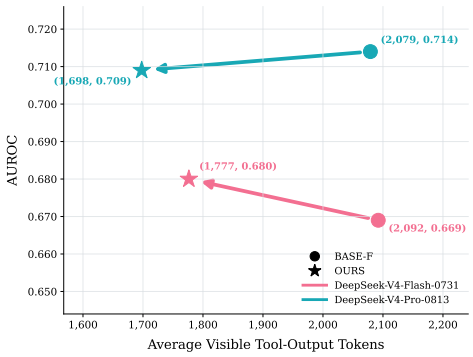



Figure 4: Evidence-access efficiency on MIMIC-IV. Circles denote **BASE-F**, which exposes all tool outputs simultaneously, and stars denote **OURS**, which selectively acquires patient-specific evidence.

### 4.3 Analysis of Explicit Cross-Organ Interaction Modeling

**Gains from Cross-Organ Message Integration.** We evaluate DynaOrg using the Area Under the Precision–Recall Curve (AUPRC), where higher scores reflect superior detection of deteriorating cases. As shown in Figure 5(a), cross-organ message passing offers highly selective utility across physiological systems. Cross-organ communication markedly boosts prediction performance for coagulation and neurological deterioration, whereas local temporal dynamics alone suffice for cardiovascular, respiratory, renal, and hepatic trajectories. This behavior is directly mirrored by the learned adjustment gates (Figure 5b): coagulation and neurological sub-networks exhibit much higher reliance on incoming messages, contrast with the low average gate activations observed in cardiovascular and respiratory modules. Collectively, these findings confirm that DynaOrg adaptively incorporates cross-organ evidence only when it yields complementary predictive value, rather than indiscriminately overriding local dynamics.

**Case-Level Organ Interaction Patterns.** Figure 5(c) further illustrates this behavior at the patient level. 1) In the renal case, a large adjustment gate ($\beta = 0.826$), with the strongest incoming interac-

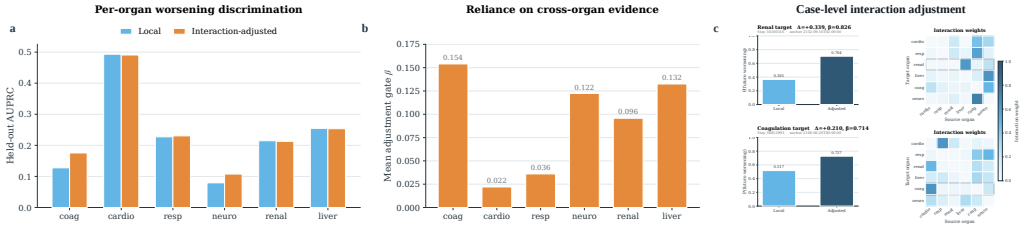



Figure 5: Analysis of cross-organ interaction in DynaOrg. (a) Held-out AUPRC of local and interaction-adjusted predictions. (b) Mean adjustment gate ($\beta$), indicating reliance on cross-organ messages. (c) Case-level organ interaction patterns.

Table 2: Performance across the three-stage training pipeline. Subtle row shading indicates the progressive training stages. Within each dataset, the best result for each metric is shown in **bold**, and the second-best result is underlined.

| Model | Training stage | MIMIC-IV | | | eICU-CRD | | |
|---|---|---|---|---|---|---|---|
| | | AUROC | F1 | Recall | AUROC | F1 | Recall |
| **Qwen3.5-4B** | BASE | 0.482 | 0.249 | 0.354 | 0.560 | 0.403 | 0.612 |
| | STAGE 1 | 0.506 | 0.295 | 0.440 | 0.578 | 0.442 | 0.698 |
| | STAGE 1 + STAGE 2 | 0.576 | 0.378 | 0.569 | 0.614 | 0.459 | 0.802 |
| | STAGE 1 + STAGE 2 + DPO | 0.598 | 0.394 | **0.590** | 0.635 | 0.481 | **0.836** |
| **Qwen3.5-9B** | BASE | 0.459 | 0.218 | 0.255 | 0.528 | 0.355 | 0.537 |
| | STAGE 1 | 0.466 | 0.232 | 0.306 | 0.531 | 0.370 | 0.608 |
| | STAGE 1 + STAGE 2 | 0.558 | 0.326 | 0.473 | 0.597 | 0.414 | 0.682 |
| | STAGE 1 + STAGE 2 + DPO | **0.662** | **0.418** | 0.537 | **0.702** | **0.485** | 0.684 |

tion from the liver, increases the worsening probability from $0.365$ to $0.704$ ($\Delta = +0.339$). Thus, the cross-organ pathway substantially revises an otherwise moderate local renal estimate when the broader multiorgan state provides additional evidence of deterioration. 2) In the coagulation case, a strong cardiovascular contribution combined with $\beta = 0.714$ raises the prediction from $0.517$ to $0.727$ ($\Delta = +0.210$). This pattern is clinically intuitive: systemic circulatory and hemodynamic abnormalities may provide predictive information about subsequent coagulation dysfunction. Additional quantitative analysis of the DynaOrg design are provided in Appendix B.3.

### 4.4 RESULTS OF THREE-STAGE TRAINING

Table 2 shows consistent gains across the training pipeline. Stage 1 mainly improves reliable tool use, while Stage 2 yields larger gains by teaching the agent which evidence to acquire from patient-conditioned teacher trajectories. This trend is consistent across MIMIC-IV and eICU-CRD.

DPO further improves performance using trajectories sampled from the student policy, with the largest gain observed for Qwen3.5-9B. Although the 4B base model starts stronger, the fully trained 9B model achieves the best overall results. This suggests that additional model capacity becomes more useful once the agent is provided with structured trajectory supervision and preference optimization, which better exploit its ability to plan evidence acquisition and integrate heterogeneous clinical signals.

## 5 CONCLUSION

We presented a framework for early sepsis prediction that combines explicit cross-organ modeling with active evidence acquisition. Across multiple backbones and ICU datasets, this approach improves prediction while reducing unnecessary evidence exposure, and interaction trajectories further provide an effective signal for training smaller agents. These results highlight the value of moving beyond fixed end-to-end inference toward structured and selectively acquired clinical evidence.

### AI USE STATEMENT

In this work, we used generative AI tools solely as writing assistants to polish the language of the manuscript. They were not involved in research ideation, experiment design, implementation, analysis, or any other scientific contribution. All AI-refined passages were subsequently reviewed and revised by us. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

### REPRODUCIBILITY STATEMENT

We document all model architectures, training procedures, data processing steps, cohort construction rules, variable mappings, and auxiliary components in the main text and Appendix. As MIMIC-IV and eICU-CRD are credentialed datasets governed by PhysioNet data use agreements, raw and derived patient-level data cannot be redistributed. We therefore provide the exact specifications required for credentialed researchers to reconstruct the experimental pipeline and reproduce our results.

### ETHICS STATEMENT

This work uses de-identified, credentialed-access clinical data from MIMIC-IV and eICU-CRD. The original collection and release of these datasets were conducted under the corresponding institutional review and data-use requirements.

## REFERENCES

Fatemeh Amrollahi, Supreeth P Shashikumar, Fereshteh Razmi, and Shamim Nemati. Contextual embeddings from clinical notes improves prediction of sepsis. In *AMIA annual symposium proceedings*, volume 2020, pp. 197–202, 2021.

Baichuan Intelligence. Baichuan4. https://www.baichuan-ai.com/, 2024. Accessed: 2026-09-24.

Armando D Bedoya, Joseph Futoma, Meredith E Clement, Kristin Corey, Nathan Brajer, Anthony Lin, Morgan G Simons, Michael Gao, Marshall Nichols, Suresh Balu, et al. Machine learning for early detection of sepsis: an internal and temporal validation study. *JAMIA open*, 3(2):252–260, 2020.

Roger C Bone, Robert A Balk, Frank B Cerra, R Phillip Dellinger, Alan M Fein, William A Knaus, Roland MH Schein, and William J Sibbald. Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. *Chest*, 101(6):1644–1655, 1992.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734. Association for Computational Linguistics, 2014. doi: 10.3115/v1/D14-1179.

DeepSeek-AI. DeepSeek-V4-Flash-0731. https://api-docs.deepseek.com/updates/, July 2026a. DeepSeek-V4-Flash update released on July 31, 2026; accessed 2026-09-24.

DeepSeek-AI. DeepSeek-V4-Pro-0813. https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro-0813, August 2026b. Accessed: 2026-09-24.

Thomas Desautels, Jacob Calvert, Jana Hoffman, Melissa Jay, Yaniv Kerem, Lisa Shieh, David Shimabukuro, Uli Chettipally, Mitchell D Feldman, Chris Barton, David J Wales, and Ritankar Das. Prediction of sepsis in the intensive care unit with minimal electronic health record data: A machine learning approach. *JMIR Med Inform*, 4(3):e28, Sep 2016. ISSN 2291-9694. doi: 10.2196/medinform.5909. URL http://medinform.jmir.org/2016/3/e28/.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Yongrui Duan, Jiazhen Huo, Mingzhou Chen, Fenggang Hou, Guoliang Yan, Shufang Li, and Haihui Wang. Early prediction of sepsis using double fusion of deep features and handcrafted features. *Applied Intelligence*, 53(14):17903–17919, 2023. doi: 10.1007/s10489-022-04425-z.

Xue Feng, Lei Sun, Jintao Zhu, Xuan Yao, Zhongheng Zhang, Qing Pan, Luping Fang, and Gangmin Ning. Subphenotyping sepsis based on organ interaction trajectory using a deep temporal graph clustering model: a retrospective cohort study. *EClinicalMedicine*, 90:103691, 2025. doi: 10.1016/j.eclinm.2025.103691.

GLM-5-Team, Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, et al. GLM-5: from vibe coding to agentic engineering, 2026. URL https://arxiv.org/abs/2602.15763.

Kim Huat Goh, Le Wang, Adrian Yong Kwang Yeow, Hermione Poh, Ke Li, Joannas Jie Lin Yeow, and Gamaliel Yu Heng Tan. Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. *Nature communications*, 12(1):711, 2021.

Google. Introducing Gemini 3.8 Flash and 3.8 Flash Cyber. https://blog.google/innovation-and-ai/models-and-research/gemini-models/3-8-flash-and-3-8-flash-cyber/, September 2026. Accessed: 2026-09-24.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735.

Khandaker Reajul Islam, Johayra Prithula, Jaya Kumar, Toh Leong Tan, Mamun Bin Ibne Reaz, Md Shaheenur Islam Sumon, and Muhammad E. H. Chowdhury. Machine learning-based early prediction of sepsis using electronic health records: A systematic review. *Journal of Clinical Medicine*, 12(17):5658, 2023. doi: 10.3390/jcm12175658.

Shihao Ji and Lawrence Carin. Cost-sensitive feature acquisition and classification. *Pattern Recognition*, 40(5):1474–1485, May 2007. ISSN 0031-3203. doi: 10.1016/j.patcog.2006.11.008. URL https://doi.org/10.1016/j.patcog.2006.11.008.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-Wei H. Lehman, Leo Anthony Celi, and Roger G. Mark. Mimic-iv, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023. doi: 10.1038/s41597-022-01899-x.

SM Lauritsen, ME Kalør, EL Kongsgaard, KM Lauritsen, MJ Jørgensen, J Lange, and B Thiesson. Early detection of sepsis utilizing deep learning on electronic health record event sequences. *Artificial Intelligence in Medicine*, 104:101820, 2020. doi: 10.1016/j.artmed.2020.101820.

Shuyue S Li, Vidhisha Balachandran, Shangbin Feng, Jonathan S Ilgen, Emma Pierson, Pang W Koh, and Yulia Tsvetkov. Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning. *Advances in Neural Information Processing Systems*, 37:28858–28888, 2024.

Yang Li and Junier B. Oliva. Active feature acquisition with generative surrogate models, 2021. URL https://arxiv.org/abs/2010.02433.

Yu-sheng Li, Xincen Duan, Beili Wang, Wei Guo, and Han-jia Ye. Disc: Dynamic feature selection for cost-sensitive medical diagnosis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 23283–23291, 2026. doi: 10.1609/aaai.v40i28.39497.

Vincent Liu, Gabriel J. Escobar, John D. Greene, Jay Soule, Alan Whippy, Derek C. Angus, and Theodore J. Iwashyna. Hospital deaths in patients with sepsis from 2 independent cohorts. *JAMA*, 312(1):90–92, 2014.

Vincent X Liu, Vikram Fielding-Singh, John D Greene, Jennifer M Baker, Theodore J Iwashyna, Jay Bhattacharya, and Gabriel J Escobar. The timing of early antibiotics and hospital mortality in sepsis. *American journal of respiratory and critical care medicine*, 196(7):856–863, 2017.

Zhuohan Long and Zhongyu Wei. Strong reasoning isn't enough: Evaluating evidence elicitation in interactive diagnosis. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 27559–27575, 2026.

Chao Ma, Sebastian Tschiatschek, Konstantina Palla, José Miguel Hernández-Lobato, Sebastian Nowozin, and Cheng Zhang. EDDI: efficient dynamic discovery of high-value information with partial VAE. *CoRR*, abs/1809.11142, 2018. URL http://arxiv.org/abs/1809.11142.

Shamim Nemati, Andre Holder, Fereshteh Razmi, Matthew D. Stanley, Gari D. Clifford, and Timothy G. Buchman. An interpretable machine learning model for accurate prediction of sepsis in the icu. *Critical Care Medicine*, 46(4):547–553, 2018. doi: 10.1097/CCM.0000000000002936.

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition. https://openai.com/index/gpt-5-6/, July 2026. Accessed: 2026-09-24.

Tom J. Pollard, Alistair E. W. Johnson, Jesse D. Raffa, Leo A. Celi, Roger G. Mark, and Omar Badawi. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific Data*, 5:180178, 2018. doi: 10.1038/sdata.2018.178.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: your language model is secretly a reward model. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Weijieying Ren, Tianxiang Zhao, Lei Wang, Tianchun Wang, and Vasant G Honavar. DiaLLMs: EHR-enhanced clinical conversational system for clinical test recommendation and diagnosis prediction. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 25622–25635, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1313. URL https://aclanthology.org/2025.findings-acl.1313/.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Ji Woong Kim, Eduardo Pontes Reis, Jeffrey Jopling, and Michael Moor. Agentclinic: a multimodal benchmark for tool-using clinical ai agents. *npj Digital Medicine*, 9:499, 2026. doi: 10.1038/s41746-026-02674-7.

Supreeth P. Shashikumar, Christopher S. Josef, Ashish Sharma, and Shamim Nemati. Deepaise – an interpretable and recurrent neural survival model for early prediction of sepsis. *Artificial Intelligence in Medicine*, 113:102036, 2021. ISSN 0933-3657. doi: https://doi.org/10.1016/j.artmed.2021.102036. URL https://www.sciencedirect.com/science/article/pii/S0933365721000294.

Supreeth P Shashikumar, Sina Mohammadi, Rishivardhan Krishnamoorthy, Avi Patel, Gabriel Wardi, Joseph C Ahn, Karandeep Singh, Eliah Aronoff-Spencer, and Shamim Nemati. Development and prospective implementation of a large language model based system for early sepsis prediction. *NPJ digital medicine*, 8(1):290, 2025.

Wenqi Shi, Ran Xu, Yuchen Zhuang, Yue Yu, Jieyu Zhang, Hang Wu, Yuanda Zhu, Joyce C. Ho, Carl Yang, and May Dongmei Wang. EHRAgent: Code empowers large language models for few-shot complex tabular reasoning on electronic health records. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 22315–22339, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1245. URL https://aclanthology.org/2024.emnlp-main.1245/.

Hajin Shim, Sung Ju Hwang, and Eunho Yang. Joint active feature acquisition and classification with variable-size set encoding. *Advances in neural information processing systems*, 31, 2018.

Mervyn Singer, Clifford S Deutschman, Christopher Warren Seymour, Manu Shankar-Hari, Djillali Annane, Michael Bauer, Rinaldo Bellomo, Gordon R Bernard, Jean-Daniel Chiche, Craig M Coopersmith, et al. The third international consensus definitions for sepsis and septic shock (sepsis-3). *JAMA*, 315(8):801–810, 2016.

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.

C. P. Subbe, M. Kruger, P. Rutherford, and L. Gemmel. Validation of a modified early warning score in medical admissions. *QJM*, 94(10):521–526, 2001. doi: 10.1093/qjmed/94.10.521. URL https://pubmed.ncbi.nlm.nih.gov/11588210/.

Yan Tang, Yu Zhang, and Jiaxi Li. A time series driven model for early sepsis prediction based on transformer module. *BMC Medical Research Methodology*, 24(1):23, 2024.

Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, Elahe Vedadi, Nenad Tomasev, Shekoofeh Azizi, Karan Singhal, Le Hou, Albert Webson, Kavita Kulkarni, S. Sara Mahdavi, Christopher Semturs, Juraj Gottweis, Joelle Barral, Katherine Chou, Greg S. Corrado, Yossi Matias, Alan Karthikesalingam, and Vivek Natarajan. Towards conversational diagnostic artificial intelligence. *Nature*, 642(8067): 442–450, 2025. doi: 10.1038/s41586-025-08866-7.

Juncheng Wu, Letian Zhang, Yuhan Wang, Haoqin Tu, Hardy Chen, Zijun Wang, Cihang Xie, and Yuyin Zhou. Clinseekagent: Automating multimodal evidence seeking for agentic clinical reasoning. *arXiv preprint arXiv:2605.20176*, 2026.

Changchang Yin, Pin-Yu Chen, Bingsheng Yao, Dakuo Wang, Jeffrey Caterino, and Ping Zhang. Sepsislab: Early sepsis prediction with uncertainty quantification and active sensing. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 6158–6168, 2024. doi: 10.1145/3637528.3671586.

Changchang Yin, Shihan Fu, Bingsheng Yao, Thai-Hoang Pham, Weidan Cao, Dakuo Wang, Jeffrey Caterino, and Ping Zhang. Sepsiscalc: Integrating clinical calculators into early sepsis prediction via dynamic temporal graph construction. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*, KDD '25, pp. 2779–2790, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400712456. doi: 10.1145/3690624.3709402. URL https://doi.org/10.1145/3690624.3709402.

Xiang Zhang, Marko Zeman, Theodoros Tsiligkaridis, and Marinka Zitnik. Graph-guided network for irregularly sampled multivariate time series. In *International Conference on Learning Representations, ICLR*, 2022.

# APPENDIX

**A Data Construction and Cohort Definition** ..... 15
A.1 MIMIC-IV Cohort Construction ..... 15
A.2 eICU-CRD Cohort Construction ..... 16

**B DynaOrg: Model and Supplementary Details** ..... 17
B.1 Organ-Specific Inputs and Future Deterioration Targets ..... 17
B.2 Training Objectives and Optimization Schedule ..... 18
B.3 DynaOrg Design Ablations ..... 20

**C SepsisWeave Environment and Interfaces** ..... 21
C.1 Evidence Interfaces ..... 21
C.2 Organ-Agent Implementation ..... 21
C.3 SOFA Component Worsening Predictor ..... 21
C.4 Prompt and Action Protocol ..... 24
C.5 Tool-Output Token Accounting ..... 26

**D Baselines** ..... 26
D.1 Clinical Calculator Baselines ..... 26
D.2 Temporal Deep-Learning Baselines ..... 28
D.3 DFSP-Adapted Baseline ..... 29
D.4 LLM Baseline Settings ..... 31

**E Evaluated Models** ..... 31

**F Additional Experimental Results and Ablations** ..... 32
F.1 Case Studies of SepsisWeave ..... 32
F.2 Cross-Organ Interface Ablations ..... 32

**G Three-Stage Training Configuration** ..... 33

# A Data Construction and Cohort Definition

## A.1 MIMIC-IV Cohort Construction

**Source cohort and stay-level eligibility.** We constructed the development cohort from MIMIC-IV (Johnson et al., 2023) using a preprocessed hourly table containing 60,039 ICU stays from 41,560 adult patients. The table included the first Sepsis-3 onset time, defined by suspected infection temporally associated with an acute increase of at least two SOFA points (Singer et al., 2016). We used only information recorded by the prediction anchor; diagnosis codes from the index admission were excluded from model inputs because their recording time does not reliably establish availability at the point of prediction.

The task required a complete 24-hour post-admission observation period before the first eligible anchor and a fully observable 4-hour future horizon. A Sepsis-3-positive stay was eligible only when its first onset occurred more than 24 hours after ICU admission, no later than both ICU discharge and the last available hourly record, and after the source record had begun. Of the 48,446 Sepsis-3-positive stays in the source cohort, 1,927 had onset after the initial 24-hour period; enforcing the ICU and record boundary yielded 1,326 eligible stays from 1,319 patients. We retained one case stay per patient, selecting the stay with the earliest eligible onset and using the stay identifier as a deterministic tie-breaker. This produced 1,319 case patients.

Controls were intentionally strict. A control patient could not have any Sepsis-3-positive admission in the source cohort, and the selected ICU stay had to remain observable for at least 28 hours after admission, covering both the 24-hour history and the subsequent 4-hour outcome window. These criteria yielded 6,497 eligible stays from 5,539 never-sepsis patients. We selected one eligible stay per patient using a seeded random draw and then sampled 2,638 control patients, producing a fixed 1:2 case-to-control ratio. The final stay-level manifest contained 3,957 patients and 3,957 ICU stays (Figure 6(A)).

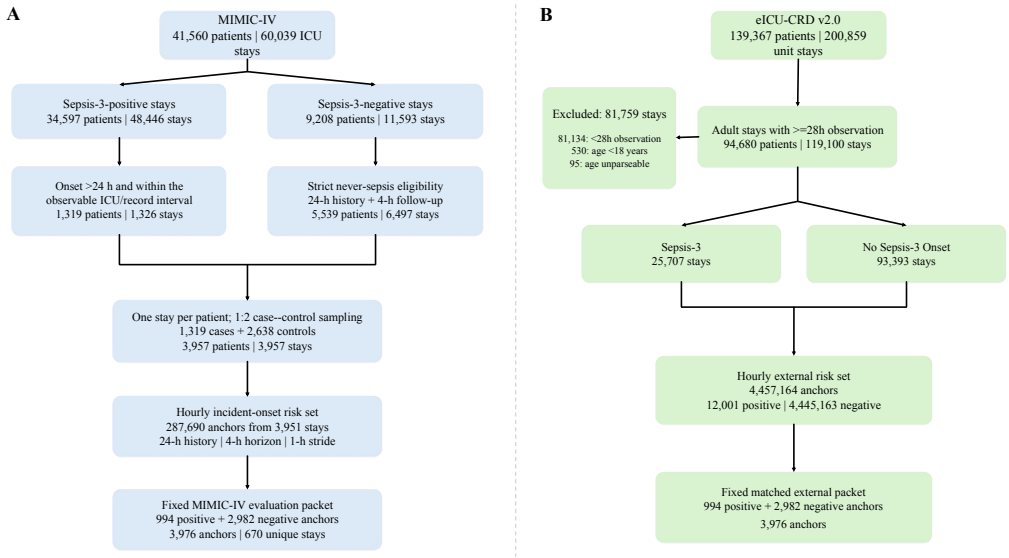



Figure 6: Cohort construction and evaluation sampling for MIMIC-IV and eICU-CRD. (A) MIMIC-IV cases were restricted to patients with a first Sepsis-3 onset after a complete 24-hour observation period, while controls were drawn from strict never-sepsis patients. The resulting patient-level cohort was converted into hourly prediction anchors and a fixed 3,976-anchor evaluation packet. (B) Eligible eICU-CRD stays were labeled using the Sepsis-3 definition and similarly converted into hourly anchors, from which a matched 3,976-anchor external evaluation packet was constructed.

**Patient-level partitioning.** We assigned case and control patients independently to training, validation, and test sets in a 70%/10%/20% ratio randomly. All anchors from the same patient and

Table 3: Frozen patient-level MIMIC-IV cohort split. One ICU stay was retained per selected patient, so patient and stay counts are identical.

| Split | Cases | Never-sepsis controls | Total |
|---|---|---|---|
| Train | 923 | 1,846 | 2,769 |
| Validation | 132 | 264 | 396 |
| Test | 264 | 528 | 792 |
| Total | 1,319 | 2,638 | 3,957 |

Table 4: Hourly anchor distribution and construction of the fixed MIMIC-IV LLM evaluation packet. The packet retained every positive test anchor and sampled negative anchors uniformly without replacement using seed 42.

| Dataset partition | Neg. | Pos. | Total | Stays | Pos. rate |
|---|---|---|---|---|---|
| Train risk set | 194,744 | 3,557 | 198,301 | 2,765 | 1.794% |
| Validation risk set | 28,503 | 500 | 29,003 | 395 | 1.724% |
| Complete test risk set | 59,392 | 994 | 60,386 | 791 | 1.646% |
| Full runtime dataset | 282,639 | 5,051 | 287,690 | 3,951 | 1.756% |
| Fixed LLM evaluation packet | 2,982 | 994 | 3,976 | 670 | 25.000% |

ICU stay inherited the same split. This immutable manifest was reused by DynaOrg, the SOFA-component predictors, the conventional baselines, and the SepsisWeave environment (Table 3).

**Hourly prediction anchors.** For each selected stay, we generated anchors at one-hour intervals, beginning 24 hours after ICU admission. The final anchor was bounded by the earliest of four hours before ICU discharge, four hours before the last available record, and 336 hours after ICU admission. For septic stays, anchor generation also stopped before the first Sepsis-3 onset. Given first onset time $\tau_i$, the label at anchor $t$ was

$$y_{i,t} = \mathbb{I}(t < \tau_i \leq t + 4\,\mathrm{h})\,. \tag{10}$$

Consequently, early anchors from an eventually septic stay remained negative unless onset occurred within the next four hours. This formulation evaluates imminent incident onset rather than the easier task of distinguishing septic admissions from non-septic admissions.

The final runtime dataset contained 287,690 unique anchors from 3,951 stays. Table 4 reports the split-level label distribution. The natural hourly positive rate was 1.756% overall and 1.646% in the complete test risk set.

**Fixed LLM evaluation packet.** To make repeated evaluation across LLMs and evidence settings computationally feasible, we froze a shared anchor-level subset of the test risk set. The packet retained all 994 positive anchors and selected 2,982 of the 59,392 negative anchors uniformly without replacement, resulting in 3,976 anchors with a 1:3 positive-to-negative ratio. The identical `sample_id` set was reused for single-summary (BASE-S), full-evidence (BASE-F), and interactive tool-use (OURS) settings. Sampling was performed at the anchor level, so a stay could contribute more than one anchor.

**Runtime-state audit.** Before formal evaluation, runtime states were materialized using an explicit feature whitelist, non-finite values were normalized, and all identifiers and future/outcome fields were removed from model-visible content. The final artifact contained no duplicate sample identifiers or stay–anchor pairs, no hourly observations after their prediction anchor, and no stay overlap across the three splits. Canonical trajectories further passed visible-message inspection and exact replay. Only this audited runtime artifact was used in the reported agent experiments.

### A.2 eICU-CRD Cohort Construction

**Source cohort and stay-level eligibility.** We used eICU-CRD v2.0 (Pollard et al., 2018) exclusively for external evaluation. The source database contained 200,859 ICU unit stays from 139,367

patients. We required an adult stay with a reliably interpretable age, a valid ICU discharge offset, and at least 28 hours of ICU observation, so that every eligible prediction could use a complete 24-hour history and a complete 4-hour outcome window. Unlike the MIMIC-IV development cohort, we retained all eligible unit stays and used the eICU `uniquepid` identifier to account for repeated stays from the same patient. The complete selection process is summarized in Figure 6(B).

**Cross-database variable harmonization.** All eICU events were mapped to the same canonical variables and organ systems used by the MIMIC-IV environment. The mapping drew on `vitalPeriodic`, `vitalAperiodic`, `lab`, `respiratoryCharting`, `respiratoryCare`, `nurseCharting`, `intakeOutput`, `medication`, and `infusionDrug`. Event offsets were converted to relative ICU hours, and only events whose source offset was no later than the prediction anchor were made visible. Physiologic ranges, feature names, missingness indicators, and freshness rules followed the frozen common schema. Diagnosis and admission-diagnosis tables were not used as dynamic model inputs because their timestamps do not reliably establish availability at the prediction time. Variables that could not be mapped reliably were left unavailable rather than imputed as normal.

**Hourly external risk set.** Prediction anchors began 24 hours after ICU admission and were generated at one-hour intervals. An anchor was retained only when its complete four-hour future window occurred before ICU discharge and, for a stay with an onset, the anchor itself preceded the first onset. For onset offset $\tau_i$ and anchor hour $t$, the external label was

$$y_{i,t} = \mathbb{I}(t < \tau_i \leq t + 4\,\mathrm{h})\,. \tag{11}$$

Stays with onset before the first eligible anchor therefore contributed no post-onset samples. The resulting natural external risk set contained 4,457,164 anchors from 96,598 stays (79,860 patients), including 12,001 positive and 4,445,163 negative anchors. Its natural anchor-level positive rate was 0.269% (Table 5).

Table 5: eICU-CRD cohort, natural hourly risk set, and fixed external execution packet.

| Construction stage | Patients | Stays | Positive anchors | Total anchors |
|---|---|---|---|---|
| Eligible stay-level cohort | 94,680 | 119,100 | – | – |
| Natural hourly risk set | 79,860 | 96,598 | 12,001 | 4,457,164 |
| Fixed external packet | 3,692 | 3,743 | 994 | 3,976 |

**Fixed external execution packet.** To keep the computational budget identical to the MIMIC-IV LLM evaluation, we materialized a fixed 3,976-anchor external packet containing 994 positive and 2,982 negative anchors. The packet covered 3,743 stays, 3,692 patients, and 189 hospitals; the 994 positive anchors came from 872 stays. The same sample identifiers were used for every LLM and evidence-access setting.

**External-evaluation boundary and audit.** eICU-CRD was not used to fit DynaOrg, auxiliary predictors, feature normalizers, agent policies, or prompts. Model and environment components learned on MIMIC-IV were transferred without eICU-specific parameter fitting. The finalized runtime passed checks for unique sample identifiers, complete future horizons, strictly pre-onset anchors, feature-range validity, and the absence of post-anchor observations, future intervention snapshots, or outcome fields in model-visible states.

# B DynaOrg: Model and Supplementary Details

## B.1 Organ-Specific Inputs and Future Deterioration Targets

**Organ-structured temporal inputs.** For each prediction anchor $t$, DynaOrg receives a 24-hour sequence of hourly states for the six organ systems. Each organ is represented by three clinically related temporal signals (Table 6). To preserve measurement provenance, every signal is accompanied

by (i) a binary indicator specifying whether it was directly observed in that hour and (ii) the number of hours since its most recent real observation. Thus, the input for organ $o$ at hour $u$ is

$$\mathbf{x}_u^o = [\mathbf{s}_u^o \parallel \mathbf{b}_u^o \parallel \mathbf{a}_u^o] \in \mathbb{R}^9, \tag{12}$$

where $\mathbf{s}_u^o \in \mathbb{R}^3$ contains the signal values, $\mathbf{b}_u^o \in \{0,1\}^3$ contains exact-hour observation indicators, and $\mathbf{a}_u^o \in \mathbb{R}^3$ records measurement age. Carried-forward values may therefore maintain a continuous temporal representation without being mistaken for newly collected measurements.

In parallel, DynaOrg receives shared clinical context available to all organ nodes. This context includes systemic measurements such as lactate, temperature, platelet count, mean arterial pressure, and shock index; time-aligned treatment indicators such as antibiotic exposure, vasopressor use or escalation, fluid administration, and sedation; overall missingness; and time-auditable comorbidities from prior discharged admissions. These shared variables provide patient-level context but do not replace the organ-specific signal partition.

Table 6: Organ-specific inputs and future deterioration targets used by DynaOrg. Changes are computed between the prediction anchor and the latest newly observed value of the primary signal within the four-hour forecast window.

| **Organ** | **Temporal signals** | **Primary signal** | **Deterioration criterion** |
|---|---|---|---|
| Cardiovascular | MAP, heart rate, vasopressor dose | MAP | future MAP $< 65$ or vasopressor increase |
| Respiratory | $SpO_2$, respiratory rate, P/F ratio | P/F ratio | $\Delta$P/F $\leq -50$ |
| Renal | creatinine, 24-h urine output, pH | creatinine | $\Delta$creatinine $\geq 0.3$ mg/dL |
| Liver | bilirubin, lactate, glucose | bilirubin | $\Delta$bilirubin $\geq 1.0$ mg/dL |
| Coagulation | platelet count, lactate, WBC | platelet count | $\Delta$platelet $\leq -50 \times 10^3/\mu$L |
| Neurologic | GCS, MAP, temperature | GCS | $\Delta$GCS $\leq -2$ |

**Future deterioration targets.** For each organ $o$, we designate a primary signal $v^o$ from which both a continuous change target and a binary deterioration target are constructed. Let $v_{i,t}^o$ be the most recent provenance-backed value available at anchor $t$, subject to a maximum measurement age of 24 hours. We select the latest newly recorded primary-signal value in the forecast interval $(t, t+H]$, denoted by $v_{i,t^+}^o$, where $H = 4$ hours. The continuous auxiliary target is

$$\Delta v_{i,t}^o = v_{i,t^+}^o - v_{i,t}^o. \tag{13}$$

The binary target $z_{i,t}^o$ is obtained by applying the organ-specific rule in Table 6 to this change. For the cardiovascular system, worsening is positive when the newly observed MAP is below $65\,\mathrm{mmHg}$ or when aggregate vasopressor support increases relative to the anchor. For the remaining systems, worsening is defined by the fixed change thresholds shown in the table. The same primary signal is used for the auxiliary change-regression target, including MAP for the cardiovascular system even when its binary label is triggered by increasing vasopressor support.

These targets describe near-term organ deterioration. DynaOrg is therefore trained to extract organ-level predictive structure that can subsequently be exposed as complementary evidence to the SepsisWeave agent.

**Measurement-aware target validity.** An organ target is considered valid only when the anchor value has raw observation provenance, is no more than 24 hours old, and at least one new primary-signal measurement is recorded within $(t, t+4\,\mathrm{h}]$. Forward-filled values alone cannot create a future target. If any of these conditions is not met, the corresponding entry is assigned a zero validity mask and excluded from both the deterioration-classification loss and the change-regression loss.

### B.2 Training Objectives and Optimization Schedule

**Masked organ-macro aggregation.** DynaOrg is supervised only when a valid future target is available for the corresponding patient–anchor and organ pair. Let $M_{i,t}^o \in \{0,1\}$ denote the target-validity mask defined in Appendix B.1. Because measurement frequency varies substantially across

organs, globally averaging all valid entries would allow frequently observed systems to dominate the objective. We instead first average within each organ and then average equally across organs represented in the minibatch. For an entry-wise loss $\ell^o_{i,t}$, this operator is

$$\mathcal{A}_{\text{org}}(\ell) = \frac{1}{|\mathcal{O}_B|} \sum_{o \in \mathcal{O}_B} \frac{\sum_{(i,t)\in B} M^o_{i,t} \ell^o_{i,t}}{\sum_{(i,t)\in B} M^o_{i,t}}, \tag{14}$$

where $\mathcal{O}_B$ contains the organs with at least one valid target in minibatch $B$. Invalid or unobserved targets therefore contribute neither positive nor negative supervision.

**Future-worsening classification.** Let $z^o_{i,t} \in \{0,1\}$ be the future deterioration label and $a^o_{i,t}$ the predicted logit. To address organ-specific class imbalance, we use an $\alpha$-balanced focal loss. For each organ, the positive weight is derived exclusively from valid training entries as $r_o = N^-_o / \max(N^+_o, 1)$. We set $\alpha^+_o = r_o/(1+r_o)$ and $\alpha^-_o = 1/(1+r_o)$. With $p^o_{i,t} = \sigma(a^o_{i,t})$ and

$$p^o_{t,*} = \begin{cases} p^o_{i,t}, & z^o_{i,t} = 1, \\ 1 - p^o_{i,t}, & z^o_{i,t} = 0, \end{cases} \tag{15}$$

the worsening loss is

$$\mathcal{L}_{\text{wors}} = \mathcal{A}_{\text{org}}\left(-\alpha^o_{i,t} \left(1 - p^o_{t,*}\right)^\gamma \log p^o_{t,*}\right), \qquad \gamma = 2, \tag{16}$$

where $\alpha^o_{i,t}$ selects $\alpha^+_o$ or $\alpha^-_o$ according to the target class. We compute this loss separately for the adjusted and local pathways, denoted by $\mathcal{L}^{\text{adj}}_{\text{wors}}$ and $\mathcal{L}^{\text{local}}_{\text{wors}}$.

**Auxiliary signal-change regression.** The continuous target $\Delta v^o_{i,t}$ provides auxiliary supervision for the magnitude and direction of physiological change. To make heterogeneous units comparable, each target is divided by the standard deviation $\sigma_o$ estimated from valid entries in the MIMIC-IV training split and clipped to $[-5, 5]$ in the normalized space. Predicted changes are normalized by the same $\sigma_o$. We apply a masked Huber loss with transition point $\delta = 1$:

$$\mathcal{L}_\Delta = \mathcal{A}_{\text{org}}\left[\text{Huber}_1\left(\frac{\widehat{\Delta v}^o_{i,t}}{\sigma_o}, \text{clip}\left(\frac{\Delta v^o_{i,t}}{\sigma_o}, -5, 5\right)\right)\right]. \tag{17}$$

As above, the loss is evaluated for both adjusted and local outputs during joint backbone training.

**Interaction regularization.** Two auxiliary terms stabilize the learned cross-organ pathway during backbone training. First, an entropy-floor penalty discourages every target organ from collapsing onto a single incoming edge. For interaction matrix $\mathbf{A}_{i,t}$, let

$$\overline{H}(\mathbf{A}) = \frac{1}{|B||\mathcal{O}|} \sum_{(i,t)\in B} \sum_{o \in \mathcal{O}} \left[ - \sum_{j \in \mathcal{O}\setminus\{o\}} A^{o\leftarrow j}_{i,t} \log\left(A^{o\leftarrow j}_{i,t} + \epsilon\right)\right]. \tag{18}$$

The penalty is

$$\mathcal{L}_{\text{ent}} = \max\left(0,\, 0.5 \log(|\mathcal{O}| - 1) - \overline{H}(\mathbf{A})\right). \tag{19}$$

Second, during backbone pretraining only, a provisional residual branch produces an interaction-only correction $c^o_{i,t}$ to the local worsening logit, such that $a^{o,\text{adj}}_{i,t} = a^{o,\text{local}}_{i,t} + c^o_{i,t}$. A margin term encourages this correction to move in the target-consistent direction:

$$\mathcal{L}_{\text{imp}} = \mathcal{A}_{\text{org}}\left[\alpha^o_{i,t} \max\left(0, \mu - (2z^o_{i,t} - 1)c^o_{i,t}\right)\right], \qquad \mu = 0.25, \tag{20}$$

**Stage I: joint backbone training.** We first jointly optimize the six independent temporal encoders, shared context encoder, patient-specific graph, adjustment gate, shared prediction head, and a provisional interaction-correction branch. The Stage-I objective is

$$\begin{aligned} \mathcal{L}_{\text{joint}} = {} & \mathcal{L}^{\text{adj}}_{\text{wors}} + 0.1\mathcal{L}^{\text{adj}}_\Delta \\ & + 0.2\left(\mathcal{L}^{\text{local}}_{\text{wors}} + 0.1\mathcal{L}^{\text{local}}_\Delta\right) + 0.2\mathcal{L}_{\text{imp}} + 0.05\mathcal{L}_{\text{ent}}. \end{aligned} \tag{21}$$

The local auxiliary term ensures that the pre-interaction state remains an independently predictive reference, while the adjusted pathway remains the primary training objective. The provisional correction branch is used only to establish the encoders, graph, gate, and shared head; it is not retained in the deployed DynaOrg architecture.

All continuous inputs are standardized using statistics estimated on the MIMIC-IV training split only. To prevent long ICU stays from dominating an epoch, anchors are sampled with replacement using weights inversely proportional to the number of anchors contributed by their stay. We optimize with AdamW using a learning rate of $3 \times 10^{-4}$, weight decay $10^{-3}$, batch size 64, and seed 42. Training runs for at most 40 epochs. A ReduceLROnPlateau scheduler halves the learning rate after five validation epochs without improvement, down to a minimum of $10^{-6}$; training stops after ten epochs without checkpoint improvement.

**Stage II: latent-message refinement.** We then replace the provisional output-space correction with the deployed latent update

$$\widetilde{\mathbf{l}}^{o}_{i,t} = \mathrm{LN}\big(\mathbf{l}^{o}_{i,t} + \beta^{o}_{i,t}\phi_m\big(\mathbf{m}^{o}_{i,t} - \mathbf{l}^{o}_{i,t}\big)\big)\,. \tag{22}$$

The temporal encoders, context encoder, dynamic graph, gate, feature normalizers, and shared prediction head are frozen. Only the message-update network $\phi_m$ and its output Layer Normalization remain trainable. This preserves the local prediction as an exact fixed reference and ensures that any change in the adjusted prediction must pass through the explicit cross-organ message pathway.

The active Stage-II objective is solely the adjusted-path focal classification loss,

$$\mathcal{L}_{\text{refine}} = \mathcal{L}^{\text{adj}}_{\text{wors}}. \tag{23}$$

The signal-change, local-supervision, and interaction-margin coefficients are all set to zero during this stage. We use AdamW with learning rate $10^{-4}$, weight decay $10^{-3}$, batch size 512, and train with the same stay-balanced sampling scheme and seed.

### B.3 DynaOrg Design Ablations

We compare the finalized DynaOrg model with three design variants.

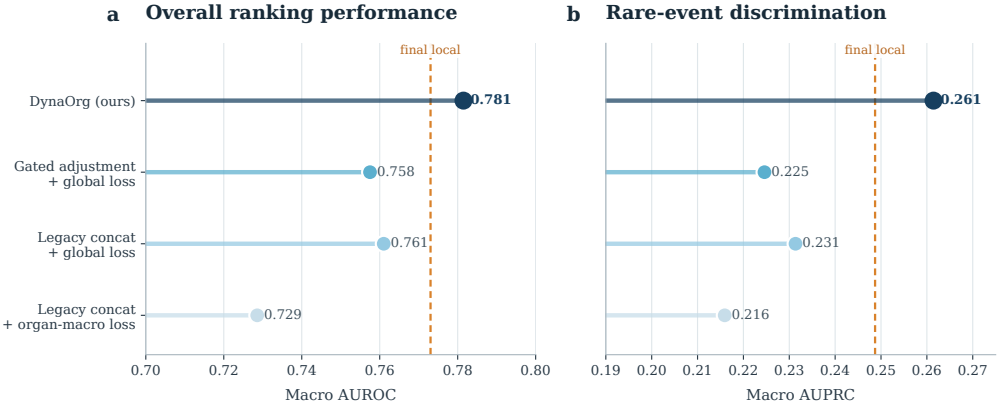



Figure 7: Performance across DynaOrg design variants. (a) Macro AUROC for future organ deterioration. (b) Macro AUPRC. Darker color denotes a later design stage, and the dashed orange line marks the frozen local pathway of the final model. The final gated latent-message model achieves the strongest discrimination and improves over its own pre-interaction reference on both metrics.

As shown in Figure 7, our final model achieves a macro AUROC of 0.7815 and a macro AUPRC of 0.2614. Compared to the best-performing baseline (legacy concatenation with global-valid aggregation), this represents absolute improvements of 0.0301 (13.0% relative) in AUPRC and 0.0205 in AUROC. Similar gains are observed on the validation set, where the model yields 0.2949 macro AUPRC and 0.7839 macro AUROC. In Figure 7, the dashed lines contrast the adjusted prediction against the frozen local pathway within the final model. On the test set, explicit message integration

improves macro AUPRC from 0.2487 to 0.2614 and macro AUROC from 0.7730 to 0.7815. By sharing identical encoders, context, and prediction heads, this within-model comparison isolates the incremental effect of the learned cross-organ update more directly than cross-variant baselines.

# C SepsisWeave Environment and Interfaces

## C.1 Evidence Interfaces

Table 7: Evidence interfaces in SepsisWeave. DynaOrg provides the core cross-organ evidence, while the remaining interfaces supply contextual, organ-local, reference, or reliability evidence under the same temporal cutoff.

| Interface | Query scope | Structured observation | Role |
|---|---|---|---|
| `get_patient_summary` | Current episode | Compressed longitudinal trends, clinical context, interventions, and measurement availability. | General context |
| `analyze_organ` | One of six organs | Current dysfunction level, observed trajectory, supporting and counter-evidence, and missing-data indicators. | Local evidence |
| `query_cross_organ_risk` | All six organ histories | Per-organ local and interaction-adjusted deterioration risks, adjustment gates, and dominant directed interactions. | DynaOrg evidence |
| `predict_sofa` | Available organ trends | Freshness-valid SOFA-component worsening probabilities and reasons for unavailable evidence. | Reference evidence |
| `check_data_quality` | Observation metadata | Measurement sparsity, staleness, missingness, and their associated uncertainty. | Reliability evidence |
| `submit_prediction` | Acquired evidence | Final sepsis probability, supporting evidence, and a concise evidence-grounded rationale. | Terminal action |

## C.2 Organ-Agent Implementation

The `analyze_organ` interface is implemented using six organ-specialist agents corresponding to the cardiovascular, respiratory, renal, liver, coagulation, and neurologic systems. Each agent has a distinct identity, organ-specific system prompt (Figure 8), and isolated evidence packet, but all six share the same frozen Qwen3-8B inference backend. This design keeps the clinical interpretation supplied by the environment fixed across all evaluated main agents.

**Organ-local evidence isolation.** For an organ $o$, the environment constructs an anchor-bounded packet $\mathcal{P}^o_{i,t}$ containing only variables relevant to that organ. Table 8 summarizes the principal contents of the six packets.

## C.3 SOFA Component Worsening Predictor

The `predict_sofa` interface is backed by a lightweight predictor of near-term SOFA-component worsening. It serves as a conventional, non-graph reference within SepsisWeave and is trained independently of DynaOrg and the decision-making agent. The predictor uses the same anchor-bounded clinical history and forecast horizon as the primary task, but predicts organ-specific SOFA progression rather than future sepsis directly.

**Prediction Targets.** Let $s^o_{i,t} \in \{0, \ldots, 4\}$ denote the current SOFA grade for component $o \in \mathcal{O}$ at anchor time $t$. The six components correspond to the cardiovascular, respiratory, renal, liver, coagulation, and neurologic systems. Their supporting observations are summarized in Table 9.

For each component, we locate the latest newly observed component value within the forecast interval $(t, t + H]$, denoted by $s^o_{i,t^+_o}$. The binary worsening target is

**Organ Agent Prompt**

**[SYSTEM PROMPT]**
You are an ICU [organ name] agent.
Your exclusive clinical scope is: [organ-related variables].
Analyze the [organ name] system and only the anchor-bounded evidence in the packet.
organ_dysfunction_level describes CURRENT dysfunction severity only, independent of its direction of change: stable or improving severe dysfunction remains severe. observed_trend separately describes the temporal direction and must not lower organ_dysfunction_level.

Treat missing values as unknown, not normal. GCS must be in [3, 15]. FiO2 is already normalized to a fraction and PF_ratio is already calculated. Distinguish current confirmed intervention status from a recent hourly dose proxy. Return JSON only, with no hidden chain of thought. The missing_data field must contain only items with no valid observation anywhere in the packet; missing_fraction_24h describes partial sampling only, and an item with a non-null latest value or n_obs > 0 MUST NOT be listed as missing regardless of that fraction; copy missing_data items only from completely_missing_items, and return [] when that list is empty.

Required schema:
```
{
 "organ_dysfunction_level": 0,
 "observed_trend": "indeterminate",
 "key_evidence": ["short value-grounded item"],
 "counter_evidence": ["short value-grounded item"],
 "missing_data": ["missing or stale item"],
 "analysis_summary": "one concise sentence"
}
```
organ_dysfunction_level: 0=no credible dysfunction, 1=mild, 2=moderate, 3=severe.
observed_trend: exactly one of "worsening", "stable_or_improving", or "indeterminate".
Use at most 3 items in each evidence list. Do not mention another organ.

Figure 8: Organ agent prompt template.

Table 8: Organ-specific evidence supplied to the six frozen organ agents. All measurements and derived signals are computed using observations available no later than the prediction anchor.

| Agent | Principal evidence | Derived or contextual evidence |
|---|---|---|
| Cardiovascular | MAP, systolic blood pressure, heart rate, lactate, and cardiovascular SOFA | MAP decline, shock index, lactate increase, vasopressor exposure, fluid administration, and consistency between intervention records and recent dose proxies |
| Respiratory | $SpO_2$, respiratory rate, $PaO_2$, $FiO_2$, pH, and respiratory SOFA | Normalized $FiO_2$, derived P/F ratio, and respiratory-support proxy |
| Renal | Creatinine, 24-hour urine output, MAP, and renal SOFA | Recent creatinine increase and fluid administration |
| Liver | Bilirubin and hepatic SOFA | Recent bilirubin increase |
| Coagulation | Platelet count and coagulation SOFA | Relative platelet decline and blood-product administration |
| Neurologic | Valid GCS observations, MAP, temperature, and CNS SOFA | Recent GCS decline, sedation status, and an explicit indicator of possible sedation-related confounding |

$$z^o_{i,t} = \mathbb{I}\left(s^o_{i,t^+_o} - s^o_{i,t} \geq 1\right). \tag{24}$$

Table 9: Clinical observations used to establish the current and future states of each SOFA component.

| SOFA component | Supporting observations |
|---|---|
| Cardiovascular | Mean arterial pressure and vasopressor administration |
| Respiratory | $PaO_2/FiO_2$ ratio |
| Renal | Creatinine and 24-hour urine output |
| Liver | Bilirubin |
| Coagulation | Platelet count |
| Neurologic | Glasgow Coma Scale |

A target is considered valid only if the current component is supported by a real clinical observation no more than 24 hours old and at least one new supporting observation is recorded within the forecast interval. Specifically,

$$M^o_{i,t} = \mathbb{I}\left(a^o_{i,t} \le 24\right) \mathbb{I}\left(t^+_o \in (t, t+H]\right), \tag{25}$$

where $a^o_{i,t}$ is the age, in hours, of the most recent observation supporting the current component. Samples with $M^o_{i,t} = 0$ are excluded from the loss for component $o$. Components already at the maximum SOFA grade of 4 remain valid but have a structurally negative worsening target.

**Input Representation.** For every prediction anchor, we construct a 24-hour hourly history containing 88 variables. These comprise 24 physiological and laboratory measurements, seven intervention features, 50 missingness and freshness descriptors, and the six current SOFA components together with their total. The intervention features summarize antibiotic exposure, vasopressor use and escalation, fluid administration, sedation, and overall intervention intensity. Missingness descriptors explicitly encode whether and how recently each clinical variable was observed, allowing the model to distinguish an observed stable value from a carried-forward value.

We transform the hourly history into a compact trend representation rather than fitting the predictor only on the latest snapshot. Let $\mathbf{x}_{i,t} \in \mathbb{R}^{88}$ be the latest feature vector, $\mathbf{V}_{i,t}$ the trajectories of the 31 clinical and intervention variables together with the six component scores, and $\mathbf{O}_{i,t}$ the observation indicators of the 24 clinical variables. The encoded representation is

$$\boldsymbol{\psi}_{i,t} = \left[\mathbf{x}_{i,t} \parallel \Delta_6 \mathbf{V}_{i,t} \parallel \Delta_{24} \mathbf{V}_{i,t} \parallel \boldsymbol{\mu}_6(\mathbf{V}_{i,t}) \parallel \operatorname{range}_6(\mathbf{V}_{i,t}) \parallel \boldsymbol{\rho}_6(\mathbf{O}_{i,t}) \parallel \boldsymbol{\rho}_{24}(\mathbf{O}_{i,t})\right]. \tag{26}$$

Here, $\Delta_6$ and $\Delta_{24}$ denote changes over the most recent six hours and the full history window, respectively; $\boldsymbol{\mu}_6$ and $\operatorname{range}_6$ capture recent level and variability; and $\boldsymbol{\rho}_6$ and $\boldsymbol{\rho}_{24}$ denote observation rates. This produces a 284-dimensional representation. Every encoded feature is standardized using means and standard deviations estimated exclusively from the training split.

**Component-Specific Models.** We fit an independent logistic model for each SOFA component:

$$\widehat{p}^o_{i,t} = \sigma\left(\mathbf{w}_o^\top \widetilde{\boldsymbol{\psi}}_{i,t} + b_o\right), \tag{27}$$

where $\widetilde{\boldsymbol{\psi}}_{i,t}$ is the standardized trend representation. Each model is optimized only over anchors with a valid target for the corresponding component. To prevent long ICU stays from dominating training, anchors from the same stay are weighted so that each stay contributes equal total weight. The positive and negative classes are subsequently reweighted to contribute equally within each component. The resulting objective is

$$\mathcal{L}^o_{\mathrm{SOFA}} = -\frac{\sum_{i,t} M^o_{i,t} w^o_{i,t} \left[z^o_{i,t} \log \widehat{p}^o_{i,t} + (1 - z^o_{i,t}) \log(1 - \widehat{p}^o_{i,t})\right]}{\sum_{i,t} M^o_{i,t} w^o_{i,t}} + \lambda_{\mathrm{reg}} \|\mathbf{w}_o\|_2^2, \tag{28}$$

where $w^o_{i,t}$ combines stay-level and class-balancing weights. Patient stays are partitioned before anchor construction so that anchors from the same stay cannot occur across training, validation, and test splits.

**Runtime Validity and Tool Output.** At inference time, the trained predictor is frozen. Before returning a forecast, the interface reconstructs the current SOFA component from the same supporting measurements used during target construction. If those measurements lack raw-observation provenance or are more than 24 hours old, the component is marked unavailable rather than assigned a deterioration probability. For a valid component, the tool returns its current grade, the time since its supporting observation, the estimated worsening risk $\widehat{p}^o_{i,t}$, and the thresholded worsening indicator $\mathbb{I}[\widehat{p}^o_{i,t} \geq \eta_o]$. When the current grade is already 4, the reported worsening risk is set to zero and the output is annotated as being at the component ceiling.

Thus, `predict_sofa` provides a freshness-aware forecast of conventional organ-dysfunction progression.

### C.4 Prompt and Action Protocol

The decision-making agent is governed by a shared prompt (Figure 9) and action protocol across model backbones and training stages. The protocol specifies the clinical prediction target, evidence-use principles, valid action format, acquisition budget, and termination behavior. Concrete interface schemas are supplied by the environment at runtime and are therefore kept separate from the model-specific chat template.

**Initial observation.** At the beginning of an episode, the agent receives a compact task observation containing the forecast horizon, relative time since ICU admission, basic demographic context, the evidence-call budget, and the currently available interfaces. The initial observation does not expose the complete patient record. Clinical evidence becomes visible only after the corresponding interface is queried.

**Action representation.** At each turn $k$, the agent produces exactly one action $a_k$. Under the JSON transport used by the environment, every response follows the envelope

$$a_k = \{\texttt{"tool"} : \texttt{tool_name}, \quad \texttt{"arguments"} : \{\ldots\}\}.$$

No additional prose or separate final-answer format is permitted. Hosted models supporting native function calling use the provider's native tool-call representation, but the resulting action is normalized to the same internal $(\texttt{tool}, \texttt{arguments})$ structure. Multiple tool calls in one turn are disallowed.

An evidence-acquisition action invokes one of the currently available clinical interfaces. A terminal action must invoke `submit_prediction` with the following semantic content:

$$a_{\text{stop}} = \left\{ \begin{array}{l} \texttt{"tool"} : \texttt{"submit_prediction"}, \\ \texttt{"arguments"} : \{\texttt{"probability"} : \hat{p}_{i,t}, \\ \qquad\quad \texttt{"evidence"} : [e_1, \ldots, e_m], \\ \qquad\quad \texttt{"reasoning"} : r\} \end{array} \right\}.$$

Here, $\hat{p}_{i,t} \in [0, 1]$ denotes the estimated probability of new Sepsis-3 onset within the forecast horizon, $e_1, \ldots, e_m$ are concise facts grounded in previously returned observations, and $r$ is a brief synthesis. The agent does not submit a binary label. Instead, the environment derives the binary prediction from $\hat{p}_{i,t}$ using the frozen evaluation threshold, ensuring that thresholding is consistent across models.

**Dynamic action availability.** Before each decision, the runtime provides a structured `TOOL STATUS` message containing (i) successfully queried interfaces, (ii) organ systems already examined through `analyze_organ`, and (iii) the interfaces currently available. Successfully completed non-organ queries are removed from the subsequent action set. For `analyze_organ`, only the already queried organ is removed, leaving the remaining organ-specific calls available. `submit_prediction` remains available throughout the episode.

## Decision-Making Agent Prompt

**[SYSTEM PROMPT]**
You are a sepsis prediction agent.
Predict whether NEW Sepsis-3 will occur within the next 4 hours using only the anchor-bounded observations shown in this dialogue. Never infer or request patient identifiers, absolute calendar dates, future outcomes, or hidden labels.

Clinical prediction guidance:
- This is a prospective prediction ofNEW Sepsis-3 onset within 4 hours. The patient need not already satisfy the complete Sepsis-3 definition at the anchor.
- Integrate all observed trajectory, treatment, organ, SOFA, cross-organ, and data-quality evidence. No single present finding is required, and no single model-derived score is a Sepsis diagnosis.
- Absence of current antibiotics, fever, treatment escalation, or a complete SOFA increase is ambiguous rather than strong reassuring evidence: these findings may emerge after the anchor within the prediction horizon. Conversely, missing data are not deterioration.
- Estimate the best calibrated event probability from the evidence as a whole. Use the full probability range when warranted, without predefined probability bands or a requirement that any particular finding be present before assigning a higher probability.

At each turn, either query exactly one available evidence tool or submit. Select the single unused tool whose result is most likely to materially change the estimated probability of new Sepsis-3 within the prediction horizon. Use tools conditionally: get_patient_summary for the global clinical trajectory; analyze_organ when evidence suggests a specific organ may be dysfunctional or changing; query_cross_organ_risk when at least two organ systems appear involved or local evidence is discordant; predict_sofa when future organ deterioration remains important or uncertain; and check_data_quality when missingness, staleness, or conflicting measurements could materially affect the estimate. Do not follow a fixed tool sequence, call a tool merely for coverage, or repeat a successful call. The ten-call budget is an upper bound, not a target. Submit when no unused tool is likely to materially change the probability.

Every assistant response, including the submission, MUST be exactly one action object with no markdown or extra prose:
{"tool": "tool_name", "arguments": {...}}
There is no separate final-answer format. To finish the episode, call submit_prediction as a regular tool action using exactly this envelope:
{"tool":"submit_prediction","arguments":{
"probability":number from 0 to 1,"evidence":["short facts already observed"],
"reasoning":"concise synthesis"}}
Do not output `prediction`; the environment derives it from probability using its frozen decision threshold. Never return `arguments` alone. Never place `probability`, `evidence`, or `reasoning` at the top level. An object without both `tool` and `arguments` is invalid.
The probability must express P(new Sepsis-3 within 4h), not confidence in your prose. Evidence must be grounded in observations already visible in the dialogue.
The evidence-tool budget is 10; submit_prediction does not consume it.

Figure 9: System prompt used by the SepsisWeave decision-making agent.

This dynamic mask prevents redundant successful calls while preserving patient-conditioned acquisition. The accompanying instruction asks the agent to select an unused interface only when its output could materially change the estimated probability; otherwise, it should submit. The agent is explicitly discouraged from following a fixed tool sequence or calling interfaces merely for coverage. Consequently, the maximum of ten evidence calls is an upper bound rather than a target, and `submit_prediction` does not consume this budget. Once the evidence budget is exhausted, submission becomes the only valid action.

### C.5 TOOL-OUTPUT TOKEN ACCOUNTING

We quantify evidence-access cost via the total *visible tool-output tokens* exposed to the agent before prediction.

**Per-episode accounting.** For episode $i$, let $\mathcal{E}_i = (e_{i,1}, \ldots, e_{i,T_i})$ denote the sequence of tool outputs observed prior to prediction. To ensure model-agnostic and deterministic accounting across different tokenizers and APIs, each observation is serialized into a canonical JSON string:

$$s(e) = \texttt{json.dumps}\,(e,\ \texttt{ensure_ascii=False},\ \texttt{sort_keys=True})\,.$$

The total visible character count $C_i^{\text{vis}}$ and the estimated token count $N_i^{\text{vis}}$ are given by:

$$C_i^{\text{vis}} = \sum_{k=1}^{T_i} \left|s(e_{i,k})\right|_{\text{char}}, \qquad N_i^{\text{vis}} = \left\lceil \frac{C_i^{\text{vis}}}{4} \right\rceil. \tag{29}$$

This frozen 4-char/token estimator prevents tokenizer choice from introducing efficiency evaluation bias.

**Accounting across evidence settings.** This uniform accounting applies to all baseline settings: for BASE-S, $N_i^{\text{vis}}$ covers solely `get_patient_summary`; for BASE-F, it spans the complete fixed 10-call sequence (summary, six organ assessments, DynaOrg evidence, SOFA predictions, and data-quality report); for OURS, it accounts only for outputs from interfaces dynamically requested by the agent.

**Full-evidence normalization for policy learning.** For preference optimization, evidence cost is normalized against $N_i^{\text{full}}$—the token count from executing the full 10-call sequence on episode $i$:

$$c_i^{\text{vis}} = \min\left(1,\ \frac{N_i^{\text{vis}}}{N_i^{\text{full}}}\right). \tag{30}$$

By normalizing at the episode level, this formulation accommodates patient-specific text density variations, penalizing the policy based on the relative fraction of acquired evidence rather than raw token length.

## D BASELINES

### D.1 CLINICAL CALCULATOR BASELINES

We construct clinical-calculator baselines from the National Early Warning Score (NEWS), Modified Early Warning Score (MEWS), quick Sequential Organ Failure Assessment (qSOFA), and Systemic Inflammatory Response Syndrome (SIRS) criteria (Smith et al., 2013; Subbe et al., 2001; Singer et al., 2016; Bone et al., 1992). Rather than treating each discrete calculator score as a calibrated sepsis probability, we follow the calculator-based comparison design used in prior early-sepsis work (Yin et al., 2025): each calculator is combined with its underlying components and commonly measured physiological variables in a logistic-regression classifier. We refer to these models as NEWS-LR, MEWS-LR, qSOFA-LR, and SIRS-LR, respectively. Their names are shortened to NEWS, MEWS, qSOFA, and SIRS in the main results table.

**Calculator construction.** At each prediction anchor, calculator components are obtained from direct observations within the preceding 24-hour history. A component is considered current only when its most recent direct measurement is no more than four hours old. Stale or entirely unavailable components contribute zero points rather than being interpreted as physiologically normal or abnormal. Separate availability indicators are retained for the downstream classifier, allowing it to distinguish a true zero-point component from an unavailable one.

Table 10 summarizes the implemented scoring rules. Supplemental oxygen for NEWS is inferred when a current $FiO_2$ measurement exceeds $0.21$. Because the available EHR representation does not contain an explicit AVPU variable, GCS is used as an alertness proxy for NEWS and MEWS.

Table 10: Clinical-calculator rules used by the baseline models. A criterion contributes points only when its source measurement is directly observed within four hours of the prediction anchor.

| Score | Component | Implemented scoring rule |
|---|---|---|
| NEWS | Respiratory rate | $\leq 8:3;\ 9\text{–}11:1;\ 12\text{–}20:0;\ 21\text{–}24:2;\ \geq 25:3$ |
| | $SpO_2$ | $\leq 91:3;\ 92\text{–}93:2;\ 94\text{–}95:1;\ \geq 96:0$ |
| | Supplemental oxygen | $FiO_2 > 0.21:2$; otherwise 0 |
| | Temperature | $\leq 35:3;\ 35.1\text{–}36.0:1;\ 36.1\text{–}38.0:0;\ 38.1\text{–}39.0:1;\ > 39.0:2$ |
| | Systolic blood pressure | $\leq 90:3;\ 91\text{–}100:2;\ 101\text{–}110:1;\ 111\text{–}219:0;\ \geq 220:3$ |
| | Heart rate | $\leq 40:3;\ 41\text{–}50:1;\ 51\text{–}90:0;\ 91\text{–}110:1;\ 111\text{–}130:2;\ > 130:3$ |
| | Consciousness | GCS $< 15:3$; otherwise 0 |
| MEWS | Respiratory rate | $< 9:2;\ 9\text{–}14:0;\ 15\text{–}20:1;\ 21\text{–}29:2;\ \geq 30:3$ |
| | Systolic blood pressure | $\leq 70:3;\ 71\text{–}80:2;\ 81\text{–}100:1;\ 101\text{–}199:0;\ \geq 200:2$ |
| | Heart rate | $< 40:2;\ 40\text{–}50:1;\ 51\text{–}100:0;\ 101\text{–}110:1;\ 111\text{–}129:2;\ \geq 130:3$ |
| | Temperature | $< 35:2;\ 35\text{–}38.4:0;\ > 38.4:2$ |
| | Consciousness | GCS $\geq 15:0;\ 13\text{–}14:1;\ 9\text{–}12:2;\ < 9:3$ |
| qSOFA | Respiratory rate | $+1$ if respiratory rate $\geq 22$ breaths/min |
| | Systolic blood pressure | $+1$ if systolic blood pressure $\leq 100$ mmHg |
| | Mental status | $+1$ if GCS $< 15$ |
| SIRS | Temperature | $+1$ if temperature $> 38^\circ$C or $< 36^\circ$C |
| | Heart rate | $+1$ if heart rate $> 90$ beats/min |
| | Respiratory | $+1$ if respiratory rate $> 20$ breaths/min or $PaCO_2 < 32$ mmHg |
| | Inflammatory | $+1$ if WBC $> 12$ or $< 4 \times 10^3/\mu$L, or band forms $> 10\%$ |

**Available-component SIRS.** WBC, $PaCO_2$, and band-form percentages are not present in the processed comparator tensor used for the formal experiments. The implemented SIRS score therefore uses only the available temperature, heart-rate, and respiratory-rate criteria. We explicitly treat this baseline as an available-component SIRS implementation rather than silently imputing the unobserved criteria. The corresponding availability indicators remain zero for the unavailable variables.

**Calculator-LR feature representation.** For calculator $c$, the final feature vector is

$$\mathbf{x}_{i,t}^{(c)} = \left[ s_{i,t}^{(c)} \parallel \mathbf{v}_{i,t}^{(c)} \parallel \mathbf{v}_{i,t}^{\mathrm{freq}} \parallel \mathbf{m}_{i,t}^{(c)} \parallel \mathbf{m}_{i,t}^{\mathrm{freq}} \right], \tag{31}$$

where $s_{i,t}^{(c)}$ is the discrete calculator score, $\mathbf{v}_{i,t}^{(c)}$ contains its component values, and $\mathbf{v}_{i,t}^{\mathrm{freq}}$ contains commonly observed physiological variables. The latter set comprises heart rate, respiratory rate, systolic, diastolic, and mean blood pressure, temperature, oxygen saturation, and glucose. Duplicate variables appearing in both sets are included only once. The vectors $\mathbf{m}_{i,t}^{(c)}$ and $\mathbf{m}_{i,t}^{\mathrm{freq}}$ indicate whether the corresponding variable was directly observed within four hours of the anchor.

The calculator score itself is always computed from recent direct measurements. In contrast, the numerical covariates supplied to logistic regression use the common non-graph preprocessing pipeline. Missing values are first filled causally through last-observation carry-forward within the 24-hour history. When no earlier observation is available, a cross-sectional IterativeImputer fitted exclusively on the MIMIC-IV training split supplies the value. This is a causal adaptation of the 3D-MICE preprocessing used in prior calculator-based comparisons, not a claim of reproducing an unreleased 3D-MICE implementation. All continuous covariates are standardized using training-set statistics only.

**Model fitting and evaluation.** Each calculator has an independent logistic-regression model,

$$\hat{p}_{i,t}^{(c)} = \sigma\left( \mathbf{w}_c^\top \mathbf{x}_{i,t}^{(c)} + b_c \right).$$

Models are fitted on the MIMIC-IV training split using class-balanced logistic regression with the `liblinear` solver, a maximum of 1,000 iterations. The continuous probability $\hat{p}_{i,t}^{(c)}$ is used directly to calculate AUROC.

For external evaluation, the complete preprocessing pipeline and classifier fitted on MIMIC-IV are applied to eICU-CRD without retraining the logistic regression coefficients. Input variables are mapped to the same feature order, and unavailable components are represented through the same zero-point and availability-indicator mechanism.

### D.2 Temporal Deep-Learning Baselines

We implement recurrent neural-network baselines using gated recurrent units (GRUs) (Cho et al., 2014) and long short-term memory networks (LSTMs) (Hochreiter & Schmidhuber, 1997). Both models receive the same anchor-bounded 24-hour input tensor and differ only in the recurrent cell type.

**Input representation.** For patient $i$ at prediction anchor $t$, the temporal input is

$$\mathbf{X}_{i,t}^{\mathrm{RNN}} \in \mathbb{R}^{24\times 17},$$

where the 24 rows correspond to hourly bins from $t-23$ through $t$. The 17 time-varying variables are listed in Table 11.

Table 11: Inputs shared by the GRU and LSTM baselines. Temporal variables form the 24-hour sequence, while static variables are concatenated with the final recurrent state.

| Input type | Variables |
|---|---|
| Temporal | Heart rate, respiratory rate, systolic blood pressure, diastolic blood pressure, mean arterial pressure, temperature, oxygen saturation, glucose, creatinine, platelet count, bilirubin, lactate, $PaO_2$, $FiO_2$, pH, GCS, and 24-hour urine output. |
| Static | Age, sex, and hours since ICU admission at the prediction anchor. |

We first reconstruct a sparse tensor containing direct observations only, removing values that entered the upstream hourly representation through carry-forward. Missing values are then handled using the same causal preprocessing pipeline across the non-graph baselines. Within each 24-hour history, a measurement is carried forward only from an earlier hour. If a variable has no preceding observation, its value is supplied by a cross-sectional IterativeImputer fitted exclusively on rows from the MIMIC-IV training split. The imputer uses median initialization, at most ten iterations, and a deterministic training-row sample capped at $100{,}000$ rows. This is a causal adaptation of the 3D-MICE preprocessing used in prior non-graph comparisons; it does not use observations occurring after the prediction anchor.

Following imputation, each temporal variable is standardized using its mean and standard deviation over the MIMIC-IV training set:

$$\widetilde{x}_{i,\tau,d} = \frac{x_{i,\tau,d} - \mu_d^{\mathrm{train}}}{\sigma_d^{\mathrm{train}}}.$$

Age, sex, and hours since admission are standardized separately using training-set statistics. No statistics from the validation, internal test, or external eICU-CRD cohorts are used during preprocessing.

**Network architecture.** Each hourly feature vector is first projected into a 64-dimensional latent space:

$$\mathbf{z}_{i,\tau} = \mathrm{ReLU}\left(\mathbf{W}_{\mathrm{in}}\widetilde{\mathbf{x}}_{i,\tau} + \mathbf{b}_{\mathrm{in}}\right).$$

The projected sequence is then processed by a two-layer recurrent encoder,

$$\mathbf{h}_{i,t} = \mathrm{RNN}_\theta\left(\mathbf{z}_{i,t-23:t}\right),$$

where $\mathrm{RNN}_\theta$ is either a GRU or an LSTM with hidden dimension 64. Dropout with probability $0.2$ is applied between recurrent layers.

The final hidden state from the upper recurrent layer is concatenated with the three standardized static covariates:

$$\mathbf{u}_{i,t} = \mathbf{h}_{i,t} \| \widetilde{\mathbf{s}}_{i,t}.$$

The prediction head applies Layer Normalization, dropout, a 32-dimensional hidden layer with ReLU activation, and a scalar output layer:

$$\hat{p}_{i,t} = \sigma\left[\mathbf{W}_2\,\mathrm{ReLU}\left(\mathbf{W}_1\,\mathrm{Dropout}\left(\mathrm{LN}(\mathbf{u}_{i,t})\right) + \mathbf{b}_1\right) + b_2\right].$$

The resulting probability estimates whether new Sepsis-3 onset will occur within the next four hours.

Table 12: Training configuration of the GRU and LSTM baselines.

| Hyperparameter | Value |
|---|---|
| History length | 24 hourly bins |
| Input variables | 17 |
| Input-projection dimension | 64 |
| Recurrent layers | 2 |
| Recurrent hidden dimension | 64 |
| Prediction-head hidden dimension | 32 |
| Dropout | 0.2 |
| Loss | Binary cross-entropy with logits |
| Optimizer | Adam |
| Learning rate | $1 \times 10^{-3}$ |
| Weight decay | $1 \times 10^{-5}$ |
| Batch size | 64 |
| Maximum epochs | 50 |
| Gradient-norm clipping | 5.0 |
| Early-stopping criterion | Validation AUROC |
| Early-stopping patience | 3 epochs |

**Optimization and model selection.** For binary label $y_{i,t}$, each model minimizes the standard binary cross-entropy objective

$$\mathcal{L}_{\mathrm{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_{i,t}\log\hat{p}_{i,t} + (1-y_{i,t})\log(1-\hat{p}_{i,t})\right].$$

We use Adam with learning rate $10^{-3}$, weight decay $10^{-5}$, and a batch size of 64. Gradients are clipped to a maximum norm of 5.0. Training runs for at most 50 epochs and terminates when validation AUROC fails to improve by more than $10^{-5}$ for three consecutive epochs. The checkpoint with the highest validation AUROC is retained.

### D.3 DFSP-Adapted Baseline

We implement an adapted version of the Double Fusion Sepsis Predictor (DFSP) (Duan et al., 2023). The original architecture combines a temporal CNN–GRU encoder with handcrafted clinical features, followed by gradient-boosted decision trees and a late-fusion network. We adapt the model to our 24-hour, 17-variable input and Sepsis-3 prediction setting. To ensure leakage-free stacking, we use stay-level out-of-fold predictions and retain the original six-unit late-fusion structure with non-negative weights.

**Temporal CNN–GRU encoder.** The encoder takes the same causally imputed and standardized input as the GRU and LSTM baselines,

$$\mathbf{X}_{i,t} \in \mathbb{R}^{24\times 17}.$$

Three $2\times1$ convolutional blocks with 64, 32, and 32 channels extract local temporal patterns without mixing clinical variables. A $1 \times 1$ convolution projects the output to a single channel, which is then processed by a two-layer GRU with hidden dimension 40 and dropout 0.2. The final GRU state

$$\mathbf{d}_{i,t}^{(m)} \in \mathbb{R}^{40}$$

is used as the learned temporal representation for ensemble member $m$.

During encoder training, this state is concatenated with age, sex, and hours since ICU admission and optimized with a temporary binary classification head. The head is discarded after training, and the GRU state is retained as the deep feature vector.

Table 13: Configuration of the DFSP-adapted baseline.

| Component | Configuration | Output or role |
|---|---|---|
| Temporal CNN | Three $2 \times 1$ convolutional blocks (64/32/32 channels), BatchNorm, ReLU, followed by $1 \times 1$ projection | Extracts local temporal patterns. |
| Recurrent encoder | Two-layer GRU, hidden dimension 40, dropout 0.2 | 40-dimensional temporal representation. |
| Handcrafted branch | Latest values; 6/12/24-hour statistics; static variables; six clinical indices | 281 handcrafted features. |
| Member predictor | LightGBM, 200 estimators, learning rate 0.05, 31 leaves, balanced weighting | Produces one member probability. |
| Ensemble | Three independently seeded CNN–GRU–GBDT members | Three member probabilities. |
| Stacking | Three-fold GroupKFold by ICU stay | Leakage-free out-of-fold member predictions. |
| Late fusion | Six hidden units with non-negative weights | Final sepsis probability. |

**Handcrafted features.** We construct a 281-dimensional feature vector $\mathbf{h}_{i,t}$ from the same observation history. It includes the latest values of the 17 temporal variables; observation frequency, mean, standard deviation, minimum, and maximum over the preceding 6, 12, and 24 hours; age, sex, and ICU time; and six clinical indices (SOFA, qSOFA, SIRS, MEWS, shock index, and P/F ratio). Observation frequency is computed from the raw observation mask, while the remaining statistics use the causally completed history. Shock index and P/F ratio are computed with lower-bounded denominators.

The deep and handcrafted features are concatenated as

$$\mathbf{q}_{i,t}^{(m)} = \mathbf{h}_{i,t} \| \mathbf{d}_{i,t}^{(m)} \in \mathbb{R}^{321}.$$

**Member models and ensemble.** Each fused representation is fed to an independent LightGBM classifier,

$$p_{i,t}^{(m)} = \mathrm{GBDT}_m(\mathbf{q}_{i,t}^{(m)}).$$

We train three independently seeded members (seeds 42, 43, and 44) with class-balanced weighting. The main architectural and optimization settings are summarized in Table 13.

**Stay-level out-of-fold stacking.** To avoid training the late-fusion network on overfitted member predictions, we generate its training inputs using three-fold GroupKFold with ICU stay as the grouping unit. For each fold, the complete CNN–GRU–GBDT pipeline is trained on two folds and evaluated on the held-out stays. The resulting predictions form

$$\mathbf{P}^{\mathrm{OOF}} = \begin{bmatrix} p_1^{(1)} & p_1^{(2)} & p_1^{(3)} \\ \vdots & \vdots & \vdots \\ p_N^{(1)} & p_N^{(2)} & p_N^{(3)} \end{bmatrix}.$$

Thus, no member prediction used to train the stacker is generated from a model fitted on the same stay. After cross-fitting, the three member pipelines are refitted on the full MIMIC-IV training split and used to obtain predictions for validation and test cohorts.

**Monotonic late fusion.** The three member probabilities are standardized using their out-of-fold training statistics and passed to a six-unit late-fusion network. We constrain both layers to have non-negative weights:

$$\hat{p}_{i,t} = \sigma\left[\mathbf{w}_2^{+\top}\,\mathrm{ReLU}\left(\mathbf{W}_1^{+}\frac{\mathbf{p}_{i,t}-\boldsymbol{\mu}_{\mathrm{OOF}}}{\boldsymbol{\sigma}_{\mathrm{OOF}}}+\mathbf{b}_1\right)+b_2\right],$$

where $\mathbf{W}_1^{+}$ and $\mathbf{w}_2^{+}$ are parameterized with Softplus. This makes the fused prediction non-decreasing with respect to each member probability. The late-fusion network is trained with weighted binary cross-entropy and selected by validation AUROC.

### D.4 LLM Baseline Settings

We evaluate each LLM backbone under three evidence-access settings (Table 14).

Table 14: Evidence-access settings used for LLM evaluation.

| Setting | Evidence access | Protocol |
|---|---|---|
| Base-S | Single standardized summary | The model receives the output of `get_patient_summary` and submits its prediction in one turn. No additional evidence interface is available. |
| Base-F | All interface outputs | A fixed exporter executes the patient summary, six organ-agent queries, the DynaOrg cross-organ query, SOFA prediction, and data-quality check. All ten outputs are exposed simultaneously before a single prediction. |
| Ours | Selective sequential access | The model interacts with SepsisWeave, conditionally selects unused interfaces, and decides when sufficient evidence has been acquired. The maximum of ten evidence calls is an upper bound, and submission does not consume the budget. |

## E Evaluated Models

We evaluate six recent LLMs spanning both proprietary and open-weight model families, covering models with different capability scales, inference characteristics, and agentic reasoning strengths.

**GPT-5.6-Sol.** GPT-5.6-Sol is the flagship model in OpenAI's GPT-5.6 family, designed for complex professional tasks with strong reasoning, long-context understanding, and tool-use capabilities (OpenAI, 2026).

**Gemini-3.8-Flash.** Gemini-3.8-Flash is Google's high-efficiency frontier model designed for long-horizon reasoning, autonomous agents, and complex multi-step tasks while maintaining the latency and cost profile of the Flash family (Google, 2026).

**Baichuan4.** Baichuan4 is a general-purpose large language model developed by Baichuan Intelligence, with an emphasis on Chinese-language understanding and generation, long-context processing, search-assisted reasoning, and multimodal capabilities (Baichuan Intelligence, 2024).

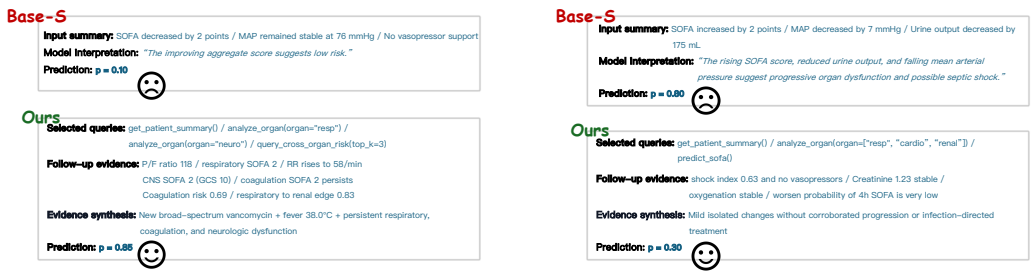
**DeepSeek-V4-Flash-0731.** DeepSeek-V4-Flash-0731 is a re-post-trained checkpoint of DeepSeek-V4-Flash that retains the original V4-Flash architecture while substantially enhancing reasoning and agentic tool-use capabilities (DeepSeek-AI, 2026a).

**DeepSeek-V4-Pro-0813.** DeepSeek-V4-Pro-0813 is the official release of DeepSeek-V4-Pro, a large Mixture-of-Experts model designed for strong reasoning, million-token long-context processing, and complex agentic tasks (DeepSeek-AI, 2026b).

**GLM-5.3.** GLM-5.3 is an open-weight model from Z.ai that builds on the GLM-5 series through enhanced post-training, with particular improvements in complex reasoning, coding, and long-horizon agentic tasks (GLM-5-Team et al., 2026).

# F Additional Experimental Results and Ablations

## F.1 Case Studies of SepsisWeave




(a) Recovering a missed positive through selective evidence acquisition.

(b) Correcting a false positive through organ-specific verification.

Figure 10: Case studies showing how SepsisWeave interaction improves sepsis prediction.

Figure 10 illustrates how SepsisWeave corrects two major failure modes. In the positive case (Figure 10a), Base-S predicts a low probability of 0.10, misled by a decreasing aggregate SOFA score and seemingly stable hemodynamics. In contrast, SepsisWeave analyzes individual organ systems and cross-organ dependencies, identifying impaired oxygenation (P/F=118), persistent neurological dysfunction, and elevated cross-organ risk. Factoring in fever and newly initiated antibiotics, SepsisWeave raises the risk probability to 0.85.

In the negative case (Figure 10b), Base-S misinterprets a two-point SOFA increase, a modest MAP drop, and reduced urine output as progressive sepsis, yielding a probability of 0.80. SepsisWeave's targeted organ assessments reveal these changes to be mild and isolated without systemic deterioration. Coupled with low-risk short-horizon SOFA forecasts, it corrects the predicted probability down to 0.30.

Together, these cases demonstrate that SepsisWeave effectively decomposes aggregate metrics, verifies initial signals, and evaluates whether clinical abnormalities form a coherent trajectory of multi-organ failure—thereby mitigating both underestimation and overestimation inherent in summary-only predictions.

## F.2 Cross-Organ Interface Ablations

We evaluate `query_cross_organ_risk` on the MIMIC-IV test set to assess the contribution of its DynaOrg-derived evidence. . We construct two ablations in addition to the complete SepsisWeave environment. Their differences are summarized in Table 15. In all conditions, the agent uses the same DeepSeek-V4-Flash backbone, prompt formulation, prediction horizon, and maximum interaction budget.

**Predictive Contribution of Cross-Organ Evidence.** As shown in Table 16, the complete cross-organ interface provides the strongest ranking performance. The output-withheld control is particularly informative: retaining the tool without returning patient-specific DynaOrg predictions does not preserve the gain. The result therefore suggests that the useful signal originates from the structured cross-organ content rather than from the mere availability of an additional action.

Recall decreases from 0.703 to 0.315 when the interface is removed and to 0.396 when its output is withheld. The corresponding reduction in F1 indicates that local organ summaries and SOFA-component predictions do not fully recover the cases identified using explicit cross-organ evidence.

Table 15: Design of the cross-organ interface ablations. The output-withheld condition preserves the action affordance but removes its patient-specific information, whereas the no-interface condition removes the tool from both the action space and prompt.

| Condition | Listed | Callable | Returned observation | Purpose |
|---|---|---|---|---|
| **FULL** | ✓ | ✓ | Local and interaction-adjusted risks, gates $\beta$, and dominant directed edges | Complete patient-specific DynaOrg evidence |
| OUTPUT WITHHELD | ✓ | ✓ | Schema-compatible unavailable response with no risks, gates, or interaction edges | Tests whether gains come from the returned interaction-adjusted information |
| NO INTERFACE | – | – | No cross-organ observation | Tests performance when explicit cross-organ evidence cannot be requested |

Table 16: Cross-organ interface ablation on MIMIC-IV test set. $\Delta$AUROC is measured relative to FULL. Tool calls denote the number of evidence-acquisition calls made by the agent. Best predictive results are shown in bold.

| Condition | AUROC | F1 | Recall | Calls | $\Delta$AUROC |
|---|---|---|---|---|---|
| **FULL** | **0.680** | **0.461** | **0.703** | 7.05 | – |
| OUTPUT WITHHELD | 0.653 | 0.348 | 0.396 | 7.73 | $-0.027$ |
| NO INTERFACE | 0.661 | 0.318 | 0.315 | 6.51 | $-0.019$ |

**Compensatory Evidence Acquisition.** Removing the interface decreases the average number of queries from 7.05 to 6.51, but this saving is accompanied by weaker predictive performance. Conversely, when the interface remains callable but returns no information, the agent makes 7.73 calls on average—more than in the complete environment—without recovering the lost discrimination. These results suggest that agents attempt to compensate for missing evidence through additional queries, but repeated querying cannot substitute for informative cross-organ evidence.

## G THREE-STAGE TRAINING CONFIGURATION

We train the Qwen3.5-4B and Qwen3.5-9B agents using QLoRA (Dettmers et al., 2023). Each backbone is trained independently, while the quantization, adapter architecture, sequence length, and effective batch size are held fixed across model scales. The three training stages sequentially update the same LoRA adapter: Stage 1 initializes an adapter from the pretrained backbone, Stage 2 continues training that adapter, and the preference-optimization stage starts from the resulting Stage 2 checkpoint.

**Quantization and Adapter Setup.** The pretrained backbone weights are frozen and quantized to 4-bit precision using the NF4 data type, double quantization, and bfloat16 computation. LoRA adapters are attached to all linear projection modules, including the attention and feed-forward projections. We set the LoRA rank to $r = 32$, the scaling parameter to $\alpha = 64$, and the adapter dropout to $0.05$. No additional bias parameters are trained. Gradient checkpointing is enabled, and the maximum sequence length is $8{,}192$ tokens.

**Stage 1: Action-Level Warm-Up.** The initial dataset contains $33{,}000$ runtime-aligned prefixes expanded from $3{,}000$ canonical trajectories. We deterministically select $6{,}000$ prefixes using phase-

Table 17: QLoRA configuration for the three-stage training pipeline. The same settings are used for Qwen3.5-4B and Qwen3.5-9B unless otherwise specified. Stage 2 and DPO continue updating the adapter obtained from the preceding stage.

| **Configuration** | **Stage 1: Warm-Up** | **Stage 2: Imitation** | **Stage 3: DPO** |
|---|---|---|---|
| Initialization | Pretrained backbone | Stage 1 adapter | Stage 2 adapter |
| Training instances | 6,000 trajectory prefixes | 7,346 trajectory prefixes | 800 preference pairs |
| Objective | Next-action cross-entropy | Reasoning–action cross-entropy | Action-normalized DPO |
| Epochs | 1 | 1 | 1 |
| Learning rate | $2 \times 10^{-4}$ | $1 \times 10^{-4}$ | $1 \times 10^{-5}$ |
| Micro-batch size | 1 | 1 | 1 |
| Gradient accumulation | 8 | 8 | 8 |
| Effective batch size | 8 | 8 | 8 |
| Maximum sequence length | 8,192 | 8,192 | 8,192 |
| Optimizer | Paged AdamW, 8-bit | Paged AdamW, 8-bit | AdamW |
| Learning-rate schedule | Linear | Linear | Linear warm-up and cosine decay |
| Warm-up ratio | 0 | 0 | 0.03 |
| Maximum gradient norm | 1.0 | 1.0 | 1.0 |
| Reasoning loss weight | – | 0.25 | – |
| DPO preference scale | – | – | $\eta = 0.1$ |

dependent sampling weights that emphasize decisions made after the initial summary and after additional evidence has been observed. For every prefix, only the next assistant action is supervised. To prevent longer JSON calls from dominating optimization, token losses are normalized within each target action:

$$\ell_{\text{act}}(a_k \mid h_k) = -\frac{1}{|a_k|} \sum_{m=1}^{|a_k|} \log \pi_\theta \left(a_{k,m} \mid h_k, a_{k,<m}\right).$$

Consequently, each target action contributes equal total loss mass regardless of its serialized length.

**Stage 2: Teacher-Guided Imitation.** Stage 2 continues training the Stage 1 adapter on 7,346 prefixes derived from 989 accepted teacher trajectories. When the teacher provides an explicit reasoning trace, both the reasoning segment and its subsequent action are supervised. The two segments are independently length-normalized and combined as

$$\mathcal{L}_{\text{SFT}}^{(2)} = \ell_{\text{act}} + 0.25\, \ell_{\text{reason}}.$$

This weighting prioritizes executable decisions while retaining a weaker learning signal for the patient-conditioned evidence-acquisition rationale. Prefixes without an available teacher reasoning segment are trained using the action loss alone.

**Stage 3: Preference Optimization.** The final adapter is initialized from Stage 2 and optimized using 800 on-policy preference pairs for each model scale. A frozen copy of the Stage 2 adapter serves as the reference policy. Only assistant action tokens are included in trajectory likelihoods; patient context, tool observations, and previously returned evidence remain visible as conditioning context but do not directly contribute to the loss.

In Equation 7, we set $\lambda_{\text{cost}} = 0.02$ and $\lambda_{\text{cross}} = 0.001$. The cross-organ term is explicitly designed as a calibrated exploration incentive—encouraging the agent to seek meaningful cross-organ evidence while ensuring the bonus remains insufficient to offset the inherent query cost, thereby preventing indiscriminate tool use. In our 3,000-episode MIMIC-IV training pool, a cross-organ call averages approximately 201 tokens, accounting for $9.857\%$ of the full-evidence output. Consequently, its mean acquisition penalty ($0.02 \times 0.09857 = 0.001971$) strictly dominates the exploration bonus ($0.001$). An uninformative query that yields no Brier-score improvement thus incurs a

net penalty of roughly $0.000971$. This partial reimbursement structure incentivizes targeted exploration without rewarding redundant evidence gathering. Note that this scalar reward serves solely to rank same-episode rollouts for constructing unweighted chosen–rejected DPO pairs, rather than being directly added to the preference loss.

To avoid favoring trajectories containing longer actions, we first average token log-probabilities within each action and then average across actions:

$$\log \bar{\pi}_\theta(\tau \mid o_0) = \frac{1}{T_\tau} \sum_{k=1}^{T_\tau} \frac{1}{|a_k|} \sum_{m=1}^{|a_k|} \log \pi_\theta \left( a_{k,m} \mid h_k, a_{k,<m} \right).$$

These normalized trajectory scores are used in the standard DPO objective with preference scale $\eta = 0.1$. Dropout is disabled during this stage to obtain deterministic policy and reference scores. Training uses a $3\%$ linear warm-up followed by cosine learning-rate decay.