Prototype-Aligned Multi-View Graph Learning for EHR Prediction

Jong-Hoon Park
Department of Software
Yonsei University Mirae Campus
Wonju-si, Gangwon-do 26493
jonghoon_park@yonsei.ac.kr

Young-Rae Cho*
Department of Software
Department of Digital Healthcare
Yonsei University Mirae Campus
Wonju-si, Gangwon-do 26493
youngcho@yonsei.ac.kr

Abstract

Clinical prediction from electronic health records (EHRs) requires reasoning over both general medical-event knowledge and visit-specific clinical evidence. However, many graph-based EHR models primarily rely on globally defined event relations derived from cohort statistics or external medical knowledge bases, including medical ontologies. While these relations capture general medical knowledge, they may insufficiently represent visit-specific signals, such as laboratory measurements and locally instantiated event interactions. We propose MVP-EHR, a prototype-aligned multi-view framework that separates global event prior from visit-instantiated evidence. MVP-EHR constructs a lift-thresholded global knowledge graph for population-level event associations and a visit-induced local graph for patient-specific event interactions, then uses shared type-aware prototypes for local evidence refinement and type-level local-global alignment before late-stage fusion. Experiments on MIMIC-IV across five clinical prediction tasks show strong overall performance, including the best results on mortality, readmission, and phenotype prediction. Ablation and prototype analyses show that global prior is most effective when it complements, rather than overrides, visit-specific evidence. These findings suggest that graph-based EHR prediction benefits not simply from stronger global knowledge, but from preserving local clinical evidence before integrating global event prior.

1 Introduction

Clinical prediction from electronic health records (EHRs) is challenging because EHR data contain complex relational structure across medical events, visits, and patient trajectories. Clinical risk does not arise only from which medical events appear in a patient history, but also from how those events interact within each visit and evolve across visits over time [4, 10]. These dependencies arise at multiple levels. Globally, medical events exhibit regularities that may be derived from cohort statistics, curated medical knowledge, ontology structures, or generated event relations [3, 5, 13, 19]. Locally, prediction depends on the subset of events observed in each visit, their interactions, and visit-specific evidence such as laboratory measurements and value-bearing clinical attributes. Temporally, visit-level states evolve along the patient trajectory, requiring reasoning across visits.

Despite this multi-level structure, existing EHR prediction models often do not preserve the distinction between globally defined event knowledge and visit-specific clinical evidence before temporal modeling. Sequence-based models generally operate on pre-aggregated visit representations, leaving event-level relational structure to be learned implicitly [6, 11, 14–16]. Graph-based models improve

*Corresponding author

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

structural awareness by introducing relations among medical events, but many still build their graph structure mainly from globally defined event knowledge, whether obtained from cohort co-occurrence, medical ontologies, external knowledge bases, or generated relations [7]. Such global structures provide useful clinical prior, but they do not directly represent the evidence instantiated in each visit. This limitation is especially important for laboratory events and other visit-dependent signals, whose predictive meaning depends not only on event identity but also on observed values, co-occurring events, and patient-specific local context.

This work is motivated by a simple question: how should global event prior and visit-instantiated evidence be represented and combined before temporal modeling? Previous studies have demonstrated the benefits of temporal modeling, graph-informed event representation, and personalized clinical context, but they often leave the roles of global event prior and visit-specific evidence entangled before temporal prediction. Our empirical finding is that these two sources are effective when represented as separate views: they should first be encoded independently, allowed to develop view-specific event representations, and then aligned through a shared semantic interface.

Based on this insight, we propose MVP-EHR, a multi-view prototype-aligned framework for EHR prediction that aligns global prior and visit-instantiated evidence through shared type-aware prototypes. MVP-EHR constructs a lift-thresholded global knowledge graph (KG) that encodes population-level event associations from the training cohort and, for each visit, derives a local induced subgraph containing only observed events and their supported local interactions. The two views are encoded separately, aligned through shared type-aware prototypes, and integrated only after view-specific event representations have been formed. Our ablations show that this separation is especially important for sparse risk prediction, where removing the global branch, local branch, local event-event relations, or type-aware prototypes substantially degrades mortality prediction.

Our contributions are summarized as follows:

- We propose a multi-view graph formulation for EHR prediction that explicitly separates global event prior from visit-instantiated evidence before temporal modeling, using a cohort-derived global KG and a visit-induced local KG.
- We introduce shared type-aware prototypes for local event refinement and type-level local-global alignment after graph encoding, allowing global prior to complement rather than override visit-specific evidence.
- Our empirical finding is that these two sources are effective when represented as separate views: they should first be encoded independently, allowed to develop view-specific event representations, and then aligned through a shared semantic interface.

2 Related Work

Sequential EHR modeling. Sequence-based EHR models represent a patient as an ordered sequence of visits and apply recurrent, convolutional, attention-based, or Transformer-style architectures for clinical prediction [1, 10–12, 15–17]. While these methods provide strong temporal modeling capacity, they typically operate on pre-aggregated visit representations, leaving event-level relations within each visit to be modeled only implicitly. Our work instead targets the visit-representation stage before temporal modeling, explicitly separating visit-instantiated evidence from global event prior.

Graph-based clinical representation learning. Graph-based EHR models encode relations among medical concepts using ontology structure, co-occurrence statistics, or LLM-generated relations [13, 19, 22]. These models provide useful clinical prior, but their graph structures are often driven by globally defined or external knowledge relations. KerPrint also considers local-global KGs, but its local graph is a patient-specific projection from an external medical KG for diagnosis-oriented interpretation [22]. MVP-EHR instead separates a cohort-derived global KG from a visit-induced local evidence graph and preserves this distinction until prototype-assisted late fusion.

Prototype learning in clinical prediction. Prototype learning has been used in healthcare for interpretability, latent state discovery, and representation regularization [2, 24]. In contrast to approaches that use prototypes as direct patient-level summaries or early representation bottlenecks, MVP-EHR uses shared type-aware prototypes as a post-encoding alignment space between local evidence and global prior.

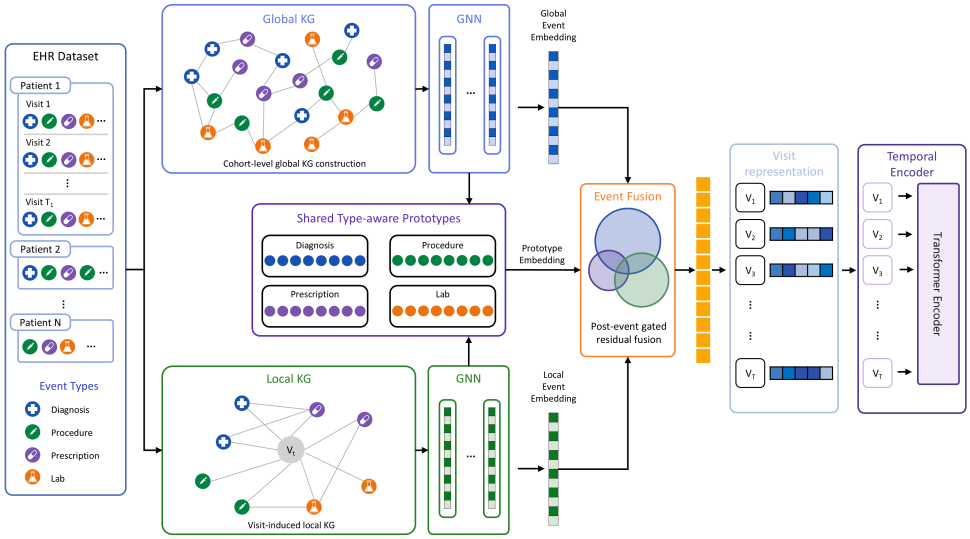

Figure 1: Overview of MVP-EHR. Global and visit-induced local graphs encode population-level priors and patient-specific evidence, respectively. Their representations are integrated through prototype-assisted multi-view fusion and passed to a temporal encoder for clinical prediction.

Multi-view clinical learning. Multi-view learning has been used to combine heterogeneous clinical signals, feature groups, graph structures, and temporal representations [7, 14, 15, 18]. Unlike settings where views mainly differ by modality or feature source, MVP-EHR treats the two graph views as semantically distinct: the global view represents population-level prior, whereas the local view represents visit-instantiated evidence.

3 Method

MVP-EHR represents each patient trajectory through two complementary views. The global view is a cohort-level KG over clinical event concepts, which captures population-level association priors from the training cohort. The local view is a visit-induced graph for each patient visit, which contains only the observed events and their supported within-visit interactions. These two views are encoded separately, aligned through shared type-aware prototypes, and fused only after view-specific event representations have been formed. The resulting visit representations are then processed by a temporal encoder for clinical predictions. Figure 1 summarizes the overall pipeline.

3.1 Problem Definition

Let $\mathcal{D} = \{P_i\}_{i=1}^{N}$ denote the EHR dataset, where each patient $P_i = (V_{i,1}, \ldots, V_{i,T_i})$ is an ordered sequence of visits. Each visit $V_{i,t} = \{e_{i,t,1}, \ldots, e_{i,t,n_{i,t}}\}$ consists of observed clinical events, and each event belongs to one of four types: diagnosis, procedure, prescription, or laboratory. Given a patient trajectory P_i, the goal is to predict a task-specific output $\hat{y}_i$ for patient i from event-level graph evidence and longitudinal visit context.

3.2 Graph Construction

MVP-EHR constructs two graph views that represent different levels of clinical evidence. The global graph summarizes reliable cohort-level event associations, whereas the local graph instantiates visit-specific evidence from the events observed in each patient visit.

Global cohort graph. We construct a global graph $G^{(g)} = (\mathcal{V}^{(g)}, \mathcal{E}^{(g)})$ over the clinical event vocabulary, where nodes are clinical event concepts and edges represent population-level associations

estimated from the training cohort. For each pair of concepts (u, v), we compute the log-lift score

$$s(u, v) = \log \frac{p(u, v)}{p(u)p(v)} \tag{1}$$

where probabilities are estimated from training visits, making $s(u, v)$ an empirical PMI score. Because raw log-lift can overemphasize rare co-occurrences supported by only a few visits, we reweight it by a logarithmic support factor:

$$\tilde{s}(u, v) = s(u, v) \log\big(1 + c(u, v)\big), \tag{2}$$

where $c(u, v)$ is the number of training visits in which u and v co-occur. This scaling preserves the specificity captured by log-lift while reducing the influence of unreliable rare pairs, and the logarithmic form prevents frequent pairs from dominating solely due to large support. The retained global edge set is

$$\mathcal{E}^{(g)} = \{(u, v) \mid \tilde{s}(u, v) \geq \tau\}, \tag{3}$$

with a minimum-support filter applied before thresholding, followed by score normalization. Thus, $G^{(g)}$ serves as a population-level prior graph that captures recurrent cohort structure while suppressing weak or incidental co-occurrences.

Local visit-induced graph. For the t-th visit of patient i, we construct a local graph $G_{i,t}^{(l)} = (\mathcal{V}_{i,t}^{(l)}, \mathcal{E}_{i,t}^{(l)})$ to represent the evidence instantiated in that visit. The node set contains a visit anchor node $v_{i,t}^{\text{visit}}$ and one node for each observed event in $V_{i,t}$. The edge set contains two types of relations: visit-event edges connecting the anchor to all observed events, and event-event edges induced from the retained global edge set. Specifically, an event-event edge $(e_{i,t,j}, e_{i,t,k})$ is included only when both events are observed in $V_{i,t}$ and the corresponding concept pair is retained in $\mathcal{E}^{(g)}$. This construction does not copy the full global prior graph into each visit, but retains only cohort-supported relations that are actually instantiated by the current visit. Accordingly, the local graph remains sparse and patient-specific while still being constrained by reliable population-level associations. A concrete example of this construction is provided in Appendix B.2. In the local graph encoder, edges are associated with learned edge features that indicate relation type and, when available, value-specific attributes such as laboratory measurements.

3.3 Prototype-aligned View Encoding

Given the global graph $G^{(g)}$ and the local visit graph $G_{i,t}^{(l)}$, we encode the two views independently with graph neural networks [21]. The global encoder processes the entire cohort-level graph and produces concept representations $h_u^{(g)}$ for event concepts $u \in \mathcal{V}^{(g)}$. For each observed event $e_{i,t,j}$, its matched global representation $h_{i,t,j}^{(g)}$ is obtained by looking up the corresponding encoded global concept embedding. In parallel, the local encoder processes the visit-induced graph and produces visit-specific event representations $h_{i,t,j}^{(l)}$. The local branch remains graph-structured during encoding, so within-visit relations are modeled through explicit message passing rather than unrestricted dense attention.

To provide a shared semantic basis for event representations, we introduce type-aware prototype banks. For each event type r, we define a prototype bank $P_r = \{p_{r,1}, \ldots, p_{r,K}\}$, where the prototypes are learned jointly with the model. Given a representation h of type r, its assignment to the prototype bank is computed as

$$\alpha_{r,k}(h) = \frac{\exp(\text{sim}(h, p_{r,k})/\gamma)}{\sum_{j=1}^{K} \exp(\text{sim}(h, p_{r,j})/\gamma)}, \tag{4}$$

where γ is a temperature parameter and $\text{sim}(\cdot, \cdot)$ denotes similarity in the prototype space. The corresponding prototype summary is

$$\tilde{p}_r(h) = \sum_{k=1}^{K} \alpha_{r,k}(h) p_{r,k}. \tag{5}$$

In the event-level fusion block, only the local event representation queries the prototype bank. For an observed event $e_{i,t,j}$ of type r, we define the event-level prototype term as

$$\tilde{p}_{i,t,j}^{(e)} = \tilde{p}_r(h_{i,t,j}^{(l)}). \tag{6}$$

The matched global representation $h^{(g)}_{i,t,j}$ does not query the prototype bank at this stage; instead, it enters the subsequent gated residual fusion directly as a population-level prior signal. This design keeps the event-level prototype term grounded in visit-instantiated evidence while allowing the global branch to provide a direct prior correction.

In addition to event-level refinement, the prototype banks are used at the type-summary level for visit-level correction. For each event type r, we form local and global type summaries from the corresponding type-specific event representations within the visit. These summaries query the same type-specific prototype bank, producing prototype-informed summaries in a shared semantic space. These summaries are used to produce a prototype-informed visit-level correction that is injected into the fused visit representation. Thus, prototypes serve two complementary roles: refining local event evidence before event-level fusion and providing type-aware local-global correction at the visit level.

3.4 Multi-view Fusion and Temporal Prediction

For each observed event $e_{i,t,j}$, we fuse three signals: the local event representation $h^{(l)}_{i,t,j}$, the matched global concept representation $h^{(g)}_{i,t,j}$, and the local-derived prototype term $\tilde{p}^{(e)}_{i,t,j}$. We form the fusion descriptor

$$u_{i,t,j} = \left[h^{(l)}_{i,t,j} \,\|\, h^{(g)}_{i,t,j} \,\|\, \tilde{p}^{(e)}_{i,t,j} \,\|\, |h^{(l)}_{i,t,j} - h^{(g)}_{i,t,j}| \,\|\, |h^{(l)}_{i,t,j} - \tilde{p}^{(e)}_{i,t,j}| \right]. \tag{7}$$

A gated residual correction is computed as

$$g_{i,t,j} = \sigma(W_g u_{i,t,j} + b_g), \qquad \Delta h_{i,t,j} = \phi_{\text{event}}(u_{i,t,j}), \tag{8}$$

and the fused event representation is

$$h^{(f)}_{i,t,j} = \text{LN}\left(h^{(l)}_{i,t,j} + g_{i,t,j} \odot \Delta h_{i,t,j} \right). \tag{9}$$

Here, ϕ_{event} is an event-level update MLP. This formulation keeps the local event representation as the residual base, while the global prior and prototype signals contribute only through a gated correction.

The local visit representation $z^{(l)}_{i,t}$ is taken from the visit anchor node produced by local graph message passing, while the fused event representations are used for event- and type-level fusion signals. The global visit representation $z^{(g)}_{i,t}$ is obtained by mean pooling the matched global concept representations associated with the events in $V_{i,t}$. Let $Z^{(l)}_i = (z^{(l)}_{i,1}, \ldots, z^{(l)}_{i,T_i})$ and $Z^{(g)}_i = (z^{(g)}_{i,1}, \ldots, z^{(g)}_{i,T_i})$ denote the local and global visit sequences for patient i. We compute cross-view attention as

$$\hat{Z}^{(g)}_i = \text{CrossAttn}\left(Z^{(l)}_i, Z^{(g)}_i \right), \tag{10}$$

where $Z^{(l)}_i$ provides queries and $Z^{(g)}_i$ provides keys and values with visit masking.

In parallel, local and global type-level summaries query the shared type-specific prototype banks to produce a prototype-informed visit-level signal $Z^{(p)}_i$. The final multi-view visit sequence is produced by a compact visit-level fusion module:

$$Z^{(\text{mv})}_i = \text{Fuse}_{\text{visit}}\left(Z^{(l)}_i, Z^{(g)}_i, \hat{Z}^{(g)}_i, Z^{(p)}_i \right). \tag{11}$$

In implementation, $\text{Fuse}_{\text{visit}}$ is a two-stage module that first mixes the local visit sequence with the cross-attended global context and then applies a prototype-informed residual correction. The detailed formulation is provided in Appendix B.7. The fused visit states $Z^{(\text{mv})}_i$ are finally processed by a Transformer encoder [20] with time-interval embeddings (detailed in Appendix B.8), followed by masked mean pooling and a task-specific prediction head.

3.5 Training Objective

MVP-EHR is optimized with a task-specific supervised loss and two auxiliary local-global regularizers. $\mathcal{L}_{\text{pred}}$ is binary cross-entropy for binary risk prediction, cross-entropy for length-of-stay classification, and multi-label binary cross-entropy for drug recommendation and phenotype prediction. The auxiliary terms include a patient-level consistency loss $\mathcal{L}_{\text{cons}}$ between local and global

patient summaries and a concept-level contrastive alignment loss $\mathcal{L}_{\text{align}}$ between local concept embeddings and matched global KG-encoded concept representations. The compact objective is

$$\mathcal{L} = \mathcal{L}_{\text{pred}} + \lambda_{\text{cons}}\mathcal{L}_{\text{cons}} + \lambda_{\text{align}}\mathcal{L}_{\text{align}}. \tag{12}$$

The auxiliary terms are further balanced using homoscedastic uncertainty weighting [9]; detailed formulations are provided in Appendix A.

4 Experimental Setup

Dataset. We evaluate MVP-EHR on MIMIC-IV [8] using a fixed patient-level train/validation/test split of $6:2:2$. All compared models are re-evaluated under the same split and metric protocol rather than relying on originally reported numbers from prior papers.

Cohort construction. We retain adult patients aged 18–80 years with valid longitudinal records, sort visits chronologically, and aggregate laboratory measurements using the last observed value within each visit. Numeric laboratory normalization statistics are fit on the training split only and reused for validation and test. After filtering, the cohort contains 62,859 patients, 204,266 visits, and 1,432 local event concepts (Table 1).

Table 1: Filtered MIMIC-IV cohort statistics.

Statistic	Value
Patients	62,859
Visits	204,266
Avg. visits per patient	3.25
Avg. events per visit	47.69
Diagnosis events	278
Procedure events	226
Prescription events	197
Laboratory events	728
Total local event vocabulary	1,429

Prediction tasks. We evaluate five clinical prediction tasks using patient visit trajectories. For binary outcome prediction, the model uses the visit prefix $(V_{i,1}, \ldots, V_{i,T_i-1})$ and predicts the outcome after V_{i,T_i-1}. For final-visit prediction tasks, the model uses the full trajectory $(V_{i,1}, \ldots, V_{i,T_i})$, with target-specific information masked when necessary.

- **Mortality prediction.** Binary prediction of whether the patient dies within 30 days after discharge from V_{i,T_i-1}.
- **Readmission prediction.** Binary prediction of whether the next admission V_{i,T_i} occurs within 30 days after discharge from V_{i,T_i-1}.
- **Length-of-stay prediction.** Multiclass prediction of the discretized length of stay for the final visit V_{i,T_i} using 10 duration classes.
- **Drug recommendation.** Multi-label prediction of the prescription set for V_{i,T_i}, with prescription targets excluded from the final-visit input context.
- **Phenotype prediction.** Multi-label prediction of diagnosis-derived phenotype labels for V_{i,T_i}, with diagnosis targets excluded from the final-visit input context.

Evaluation protocol. We report metrics according to task format: AUPRC and AUROC for binary risk prediction, AUROC and macro F1 for length-of-stay classification, and sample-wise AUPRC, AUROC, and F1 for multi-label drug recommendation and phenotype prediction. AUPRC is used as the primary model selection metric for mortality and readmission due to class imbalance. We include Logistic Regression (LR) and XGBoost as simple non-graph baselines using the same visit-level event identities, laboratory values, and relative-time information as MVP-EHR. We further compare against sequential/temporal models (DEEPR [16], ADACARE [14], STAGENET [6]), graph- or knowledge-enhanced models (GRASP [23], KERPRINT [22], GRAPHCARE [7]), and prototype-based representation learning (PROTOEHR [2]). Hyperparameters for Logistic Regression and XGBoost were selected using the task-specific primary validation metric. For the published baseline models, we used the best-performing settings reported in the original papers whenever available, with model selection performed on the validation split.

Implementation. The selected MVP-EHR configuration uses a hidden dimension of 64, two-layer local and global graph encoders, a two-layer temporal Transformer, and 16 prototypes per event type. The support-scaled log-lift threshold is set to $\tau = 0.5$. We train with Adam using a learning rate of

Table 2: Performance comparison across five clinical prediction tasks. Bold marks the best value in each column and underlines mark the second-best value.

Model	Mortality Prediction		Readmission Prediction		Length of Stay	
	AUPRC	AUROC	AUPRC	AUROC	AUROC	F1
LR	4.88 ± 1.25	62.29 ± 3.23	41.17 ± 0.89	64.02 ± 0.63	64.56 ± 0.38	23.90 ± 0.41
XGBoost	13.95 ± 2.46	90.87 ± 1.00	53.68 ± 0.85	71.40 ± 0.48	77.91 ± 0.27	25.35 ± 0.36
Deepr	10.41 ± 1.75	90.50 ± 0.95	53.30 ± 0.95	71.90 ± 0.52	82.88 ± 0.20	29.11 ± 0.40
GRASP	10.96 ± 1.59	92.13 ± 0.73	53.48 ± 0.90	72.05 ± 0.49	82.05 ± 0.22	27.96 ± 0.36
KerPrint	12.18 ± 1.72	91.48 ± 0.77	53.05 ± 0.92	71.25 ± 0.47	79.48 ± 0.24	26.67 ± 0.38
GraphCare	12.40 ± 2.12	92.01 ± 0.75	52.52 ± 0.95	71.16 ± 0.52	74.61 ± 0.28	20.65 ± 0.35
AdaCare	13.33 ± 2.20	91.78 ± 0.83	54.71 ± 0.94	72.50 ± 0.53	83.71 ± 0.19	29.65 ± 0.37
StageNet	14.19 ± 2.31	91.56 ± 0.92	54.24 ± 0.92	72.50 ± 0.51	81.76 ± 0.22	27.86 ± 0.32
ProtoEHR	14.88 ± 2.31	93.03 ± 0.77	54.65 ± 0.91	72.48 ± 0.52	$\mathbf{84.58 \pm 0.20}$	$\mathbf{32.65 \pm 0.47}$
MVPEHR	$\mathbf{17.60 \pm 2.70}$	$\mathbf{93.36 \pm 0.67}$	$\mathbf{56.08 \pm 0.90}$	$\mathbf{73.00 \pm 0.55}$	82.09 ± 0.22	29.32 ± 0.39

Model	Drug Recommendation			Phenotype Prediction		
	AUPRC	AUROC	F1	AUPRC	AUROC	F1
LR	62.24 ± 0.24	93.00 ± 0.08	51.91 ± 0.23	48.20 ± 0.22	90.03 ± 0.10	40.81 ± 0.21
XGBoost	72.16 ± 0.19	96.20 ± 0.03	56.21 ± 0.24	60.99 ± 0.20	$\mathbf{94.40 \pm 0.05}$	45.23 ± 0.23
Deepr	69.11 ± 0.18	94.84 ± 0.05	54.42 ± 0.22	72.43 ± 0.28	90.48 ± 0.13	51.22 ± 0.35
GRASP	69.92 ± 0.18	95.01 ± 0.05	56.09 ± 0.22	71.11 ± 0.27	89.53 ± 0.13	52.35 ± 0.36
KerPrint	73.87 ± 0.18	96.14 ± 0.04	60.99 ± 0.23	77.74 ± 0.25	92.79 ± 0.10	62.34 ± 0.35
GraphCare	69.88 ± 0.19	94.99 ± 0.06	54.43 ± 0.24	73.26 ± 0.28	90.01 ± 0.13	54.00 ± 0.38
AdaCare	70.74 ± 0.18	95.34 ± 0.04	57.19 ± 0.23	74.49 ± 0.27	91.43 ± 0.11	56.17 ± 0.36
StageNet	69.80 ± 0.18	95.01 ± 0.04	56.68 ± 0.21	71.49 ± 0.26	90.00 ± 0.12	51.89 ± 0.35
ProtoEHR	75.02 ± 0.18	$\mathbf{96.34 \pm 0.03}$	62.58 ± 0.23	78.18 ± 0.26	93.12 ± 0.11	62.61 ± 0.38
MVPEHR	$\mathbf{75.70 \pm 0.16}$	96.24 ± 0.04	$\mathbf{64.80 \pm 0.21}$	$\mathbf{79.51 \pm 0.24}$	93.72 ± 0.09	$\mathbf{65.65 \pm 0.33}$

10^{-3} and batch size 256. Additional hyperparameters, graph construction details, auxiliary objective formulations, and implementation details are provided in Appendices A and B.

5 Results

5.1 Performance Comparison Across Clinical Tasks

Table 2 reports the main comparison across five clinical prediction tasks, with values given as percentage-point estimates and bootstrap standard deviations over 100 resamples. MVP-EHR achieves the best AUPRC and AUROC on mortality and readmission, the best results across all reported metrics on phenotype prediction, and the best AUPRC and F1 on drug recommendation while ranking second to ProtoEHR in AUROC. Its main weakness is length-of-stay prediction, where it remains below ProtoEHR, AdaCare, and Deepr, although its F1 is comparable to Deepr. Overall, MVP-EHR is most beneficial for binary risk prediction and multi-label phenotype modeling, with less pronounced gains for multi-class duration prediction.

5.2 Ablation Study

We evaluate five controlled ablations to isolate the roles of global prior, local evidence, induced within-visit relations, type-aware prototypes, and event-level fusion. The **w/o global KG** and **w/o local KG** variants reduce the model to local-only and global-only branches, respectively. The **w/o local event-event edges** variant removes induced event-event relations while retaining visit-event edges. The **w/o type-aware prototypes** variant disables shared prototype banks, and **w/o post-event fusion** defers local-global interaction to the visit level. All variants use the same split, training protocol, and task-specific prediction heads.

We evaluate six controlled ablations to isolate the roles of global prior, local evidence, induced within-visit relations, type-aware prototypes, event-level fusion, and graph message passing. The **w/o global KG** and **w/o local KG** variants reduce the model to local-only and global-only branches,

Table 3: Ablation results across five clinical prediction tasks. Full-model results are from Table 2; results for ablated variants are test-set scores from a single run.

Model variant	Mortality		Readmission		LOS		Drug Recommendation			Phenotype Prediction		
	AUPRC	AUROC	AUPRC	AUROC	AUROC	F1	AUPRC	AUROC	F1	AUPRC	AUROC	F1
Full model	**17.60**	93.36	**56.08**	73.00	82.09	29.32	75.70	96.24	**64.80**	79.51	93.72	65.65
w/o global KG	12.00	91.68	55.47	72.93	80.73	25.96	**76.11**	**96.48**	64.43	**79.92**	**93.88**	**66.06**
w/o local KG	8.18	89.55	53.94	71.58	80.58	27.86	75.16	96.31	63.30	77.80	93.05	63.14
w/o local event-event edges	13.17	93.05	55.83	72.90	82.02	**29.90**	75.09	96.18	63.76	79.37	93.67	65.55
w/o type-aware prototypes	12.48	91.76	54.95	72.42	81.01	28.50	74.67	96.17	62.62	79.52	93.74	65.35
w/o post-event fusion	16.15	**93.45**	56.01	**73.13**	**82.16**	28.92	75.05	96.16	63.99	79.45	93.73	65.88
w/o GNN	7.91	86.46	53.45	71.32	80.16	26.22	73.62	95.85	61.13	56.75	94.03	41.35

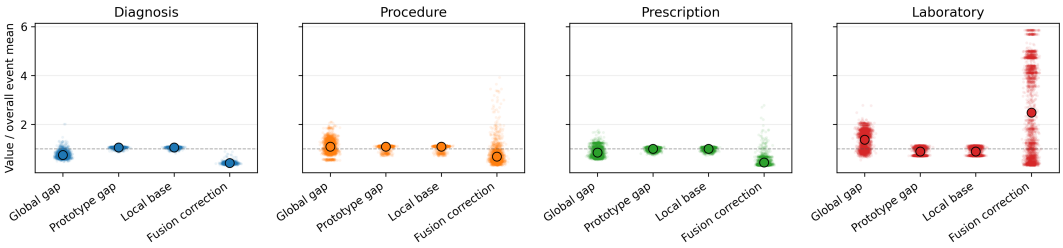

Figure 2: Event-type-specific post-event fusion behavior in the mortality model. Laboratory events show the largest fusion corrections, reflecting stronger correction for value-dependent local evidence.

respectively. The **w/o local event-event edges** variant removes induced event-event relations while retaining visit-event edges. The **w/o type-aware prototypes** variant disables shared prototype banks, and **w/o post-event fusion** defers local-global interaction to the visit level. The **w/o GNN** variant applies contrastive learning between event nodes that were neighbors in the global graph, without graph message passing, type-aware prototypes, or local-global fusion. All variants use the same split, training protocol, and task-specific prediction heads.

Table 3 shows that the effect of each component is strongly task-dependent. The clearest benefit of the full multi-view design appears in mortality prediction, where the full model reaches 17.60 AUPRC, compared with 12.00 for the local-only variant and 8.18 for the global-only variant. Removing type-aware prototypes and local event-event edges also reduces mortality AUPRC to 12.48 and 13.17, respectively, indicating that sparse mortality prediction benefits from the combination of instantiated local structure, cohort-level prior, and prototype-mediated alignment. For readmission and LOS prediction, the same trend is weaker: the full model remains competitive, but the margins are smaller and some AUROC or F1 values favor reduced variants.

The multi-label tasks show a different regime: the local-only variant remains highly competitive and even exceeds the full model on several point estimates, whereas the global-only variant degrades more clearly. This pattern indicates that the global branch should be viewed as a task-dependent prior rather than a universally beneficial signal. Global prior appears most valuable for sparse risk prediction, while high-coverage multi-label tasks can rely more heavily on local evidence and label co-occurrence. Meanwhile, the lower primary-metric performance of the w/o GNN variant indicates that global-neighbor contrastive learning alone does not fully replace explicit relational modeling. Overall, the ablations support MVP-EHR's central design choice: global prior should complement visit-specific evidence rather than override the local branch.

5.3 Post-event Fusion Behavior

Beyond aggregate performance, we examine how the post-event fusion block behaves across event types. We report the mortality model as a representative case. This analysis tests whether fusion acts as a generic smoothing operation or as a selective correction mechanism after local event structure has already been formed.

Figure 2 shows that fusion correction is highly event-type dependent. Laboratory events receive the largest correction by a wide margin (1.3464 absolute norm, 0.1801 relative to the local base), whereas diagnosis and prescription events receive much smaller corrections (0.0215 and 0.0243 relative norms, respectively). Laboratory events also exhibit the largest local-global discrepancy, suggesting that

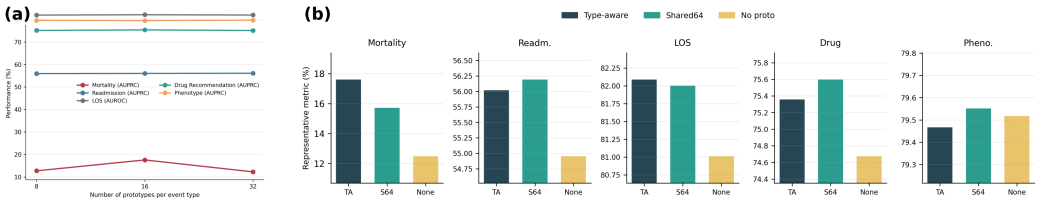

Figure 3: Prototype design analysis. (a) Prototype-count sensitivity. (b) Comparison of type-aware, type-independent, and no-prototype variants. Effects are strongest for mortality and modest for higher-coverage tasks.

post-event fusion intervenes most strongly when visit-specific evidence is harder to reconcile with global and prototype-based semantics.

The average gate strengths remain stable across event types (0.28–0.31), suggesting that the larger laboratory correction is driven more by the magnitude and structure of cross-view disagreement than by gate openness alone. This pattern is consistent with the ablation tradeoffs: excessive smoothing of local evidence may stabilize AUROC while weakening minority-positive retrieval reflected by AUPRC. Thus, post-event fusion appears to act as a selective correction mechanism, applying larger corrections to noisier, value-dependent laboratory events while leaving more stable event types less affected.

5.4 Prototype Design Analysis

We further analyze how prototype capacity and prototype structure affect performance. Figure 3(a) varies the number of prototypes per event type, and Figure 3(b) compares the type-aware prototype bank with a type-independent shared bank and a no-prototype variant.

Figure 3(a) shows that performance is not monotonic in the number of prototypes. The clearest variation is observed for mortality prediction. For this task, too few or too many prototypes reduce AUPRC, while an intermediate prototype count yields the strongest result. In contrast, readmission, LOS, drug recommendation, and phenotype prediction vary only modestly across prototype counts, suggesting that these tasks are less dependent on the exact prototype capacity.

Figure 3(b) shows that prototype structure is task-dependent. The type-aware formulation is most beneficial for mortality prediction, indicating that sparse risk prediction benefits from type-specific partitioning of the prototype space. For the remaining tasks, the type-independent shared bank remains competitive, while removing prototypes is generally worse. Thus, type-aware prototypes are most useful when fine-grained event-type structure is critical, whereas higher-coverage tasks can tolerate weaker prototype partitioning.

Additional analyses. We conduct three additional analyses to examine potential shortcut signals, input contributions, and subgroup-level performance. First, we compare MVP-EHR with a utilization-only XGBoost baseline using healthcare utilization statistics. Second, we remove observed laboratory values and result flags while retaining laboratory event identities to separate their respective contributions. Lastly, we analyze mortality prediction across patient subgroups classified by diagnosis frequency. Detailed settings and results are provided in Appendix C.

6 Discussion

The main finding of this study is that cohort-level prior and visit-specific evidence play complementary roles in EHR graph learning. Across the main comparison, ablation study, and fusion diagnostics, MVP-EHR performs best when global event associations are kept separate from locally instantiated visit evidence and combined through constrained multi-view fusion. This is especially clear in mortality prediction, where removing the global branch, local event-event structure, or type-aware prototypes substantially reduces AUPRC, whereas high-coverage multi-label tasks remain competitive under the local-only variant.

These results support a restrained view of graph-based EHR modeling. Population-level event associations provide useful prior information, but they should not replace visit-specific evidence, laboratory values, and locally instantiated interactions. The fusion diagnostics further suggest that global and prototype-derived signals act as selective corrections rather than uniform smoothing, with larger corrections assigned to laboratory events that show greater local-global disagreement. Overall, our results support the separation of cohort-level prior and visit-level evidence as a useful inductive bias, while a systematic comparison of alternative fusion timing, stronger prior injection, and type-aware temporal modeling remains an important direction for future work.

Limitations. First, all experiments are conducted on MIMIC-IV under a single preprocessing and patient-level split protocol, so external validation on other EHR systems remains necessary. Second, the global and local graphs rely on support-scaled log-lift associations; alternative association measures, external knowledge-guided relations using resources such as UMLS, or causal graph construction may yield different trade-offs between coverage, noise, and interpretability. Third, scaling the global encoder to much larger clinical ontologies may increase memory and computational costs, since the current design encodes the retained global event graph before visit-level matching. Fourth, the prototype mechanism is type-aware but lightweight, and more adaptive prototype allocation may improve performance if it can be stabilized. Finally, the benefits are task-dependent: MVP-EHR is strongest for mortality, readmission, and phenotype prediction, improves AUPRC and F1 for drug recommendation while remaining slightly below ProtoEHR in AUROC, and is weaker for length-of-stay prediction.

7 Conclusion

We proposed MVP-EHR, a prototype-aligned multi-view graph framework that separates population-level event prior from patient-specific visit evidence. By constructing a cohort-level global KG and visit-induced local graphs, encoding the two views separately, and integrating them through prototype-assisted late fusion, MVP-EHR preserves local clinical evidence while using global prior as a constrained correction. Experiments across five MIMIC-IV prediction tasks show that this design is especially effective for mortality, readmission, phenotype prediction, and drug recommendation, while its gains are less consistent for length-of-stay prediction. Overall, our results suggest that explicitly modeling cohort-level event prior and visit-specific evidence as separate views is useful for EHR prediction.

Acknowledgments and Disclosure of Funding

This research was supported by National Research Foundation of Korea (NRF) grant funded by the Ministry of Science and ICT (RS-2025-16067916), IITP grant funded by the Ministry of Science and ICT through the Digital Columbus Project (RS-2025-02304331), and the ANCHOR program through the Gangwon ANCHOR Center funded by the Ministry of Education and the Gangwon State, Republic of Korea (2026-ANCHOR-10-006).

References

[1] I. M. Baytas, C. Xiao, X. Zhang, F. Wang, A. K. Jain, and J. Zhou. Patient subtyping via time-aware LSTM networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 65–74, 2017.

[2] Z. Cai, Y. Liu, Z. Luo, and T. Zhu. ProtoEHR: Hierarchical prototype learning for ehr-based healthcare predictions. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 169–178, 2025.

[3] E. Choi, M. T. Bahadori, E. Searles, C. Coffey, M. Thompson, J. Bost, J. Tejedor-Sojo, and J. Sun. Multi-layer representation learning for medical concepts. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1495–1504, 2016.

[4] E. Choi, M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz, and W. Stewart. RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Advances in neural information processing systems*, volume 29, 2016.

[5] E. Choi, M. T. Bahadori, L. Song, W. F. Stewart, and J. Sun. GRAM: Graph-based attention model for healthcare representation learning. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 787–795, 2017.

[6] J. Gao, C. Xiao, Y. Wang, W. Tang, L. M. Glass, and J. Sun. StageNet: Stage-aware neural networks for health risk prediction. In *Proceedings of the web conference 2020*, pages 530–540, 2020.

[7] P. Jiang, C. Xiao, A. Cross, and J. Sun. GraphCare: Enhancing healthcare predictions with personalized knowledge graphs. In *12th International Conference on Learning Representations*, 2024.

[8] A. E. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T. J. Pollard, S. Hao, B. Moody, B. Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

[9] A. Kendall, Y. Gal, and R. Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491, 2018.

[10] Y. Li, S. Rao, J. R. A. Solares, A. Hassaine, R. Ramakrishnan, D. Canoy, Y. Zhu, K. Rahimi, and G. Salimi-Khorshidi. BEHRT: transformer for electronic health records. *Scientific Reports*, 10(1):7155, 2020.

[11] J. Luo, M. Ye, C. Xiao, and F. Ma. HiTANet: Hierarchical time-aware attention networks for risk prediction on electronic health records. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 647–656, 2020.

[12] F. Ma, R. Chitta, J. Zhou, Q. You, T. Sun, and J. Gao. Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1903–1911, 2017. doi: 10.1145/3097983.3098088.

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

[14] L. Ma, J. Gao, Y. Wang, C. Zhang, J. Wang, W. Ruan, W. Tang, X. Gao, and X. Ma. AdaCare: Explainable clinical health status representation learning via scale-adaptive feature extraction and recalibration. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 825–832, 2020.

[15] L. Ma, C. Zhang, Y. Wang, W. Ruan, J. Wang, W. Tang, X. Ma, X. Gao, and J. Gao. ConCare: Personalized clinical feature embedding via capturing the healthcare context. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 833–840, 2020.

[16] P. Nguyen, T. Tran, N. Wickramasinghe, and S. Venkatesh. Deepr: a convolutional net for medical records. *IEEE Journal of Biomedical and Health Informatics*, 21(1):22–30, 2016.

[17] L. Rasmy, Y. Xiang, Z. Xie, C. Tao, and D. Zhi. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):86, 2021.

[18] T. Shaik, X. Tao, L. Li, H. Xie, and J. D. Velásquez. A survey of multimodal information fusion for smart healthcare: Mapping the journey from data to wisdom. *Information Fusion*, 102:102040, 2024.

[19] J. Shang, T. Ma, C. Xiao, and J. Sun. Pre-training of graph augmented transformers for medication recommendation. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 5953–5959, 2019.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, 2017.

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

[22] K. Yang, Y. Xu, P. Zou, H. Ding, J. Zhao, Y. Wang, and B. Xie. KerPrint: Local-global knowledge graph enhanced diagnosis prediction for retrospective and prospective interpretations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 5357–5365, 2023.

[23] C. Zhang, X. Gao, L. Ma, Y. Wang, J. Wang, and W. Tang. GRASP: Generic framework for health status representation learning based on incorporating knowledge from similar patients. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 715–723, 2021.

[24] Y. Zhu, Z. Wang, L. He, S. Xie, X. Zheng, L. Ma, and C. Pan. PRISM: Mitigating EHR data sparsity via learning from missing feature calibrated prototype patient representations. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 3560–3569, 2024.

A Auxiliary Training Objectives

This section provides the detailed formulation of the auxiliary objectives used in MVP-EHR. The main text uses the compact objective $\mathcal{L} = \mathcal{L}_{\text{pred}} + \lambda_{\text{cons}}\mathcal{L}_{\text{cons}} + \lambda_{\text{align}}\mathcal{L}_{\text{align}}$; here we describe the two auxiliary terms and the uncertainty-based weighting used in the final implementation.

Patient-level consistency loss. The consistency loss encourages the local and global branches to produce compatible patient-level summaries. Let $\bar{h}_i^{(l)}$ and $\bar{h}_i^{(g)}$ denote the masked-mean pooled local and global patient representations for patient i in a minibatch of size B. We define

$$\mathcal{L}_{\text{cons}} = 1 - \frac{1}{B}\sum_{i=1}^{B} \frac{\langle \bar{h}_i^{(l)}, \bar{h}_i^{(g)} \rangle}{\|\bar{h}_i^{(l)}\|_2 \, \|\bar{h}_i^{(g)}\|_2}. \tag{13}$$

This term encourages view-level agreement at the patient-summary level while allowing the local and global encoders to preserve view-specific information before fusion.

Concept-level alignment loss. The concept-level alignment loss aligns local concept embeddings with global KG-encoded concept embeddings over matched concepts from the shared event vocabulary. Importantly, the local-side representation used in this loss is the trainable local concept embedding before local graph message passing, not a contextualized visit-level event representation. For a minibatch of M matched local-global concept pairs

$$\{(c_k^{(l)}, c_k^{(g)})\}_{k=1}^{M}, \tag{14}$$

we project the local-side concept embedding and the global-side encoded concept representation as

$$q_k = \text{Proj}_l(e_{c_k^{(l)}}), \quad \kappa_k = \text{Proj}_g(h^{(g)}_{c_k^{(g)}}), \tag{15}$$

where $e_{c_k^{(l)}}$ denotes the local concept embedding from the shared local event embedding table, and $h^{(g)}_{c_k^{(g)}}$ denotes the corresponding global KG-encoded concept representation. Both projected embeddings q_k and κ_k are ℓ_2-normalized before computing the similarity matrix, so the logits correspond to cosine-style similarities scaled by the alignment temperature. The similarity matrix is computed as

$$S_{ab} = \frac{q_a^\top \kappa_b}{\tau_{\text{align}}}, \tag{16}$$

where τ_{align} is the alignment temperature. We optimize a symmetric contrastive objective:

$$\mathcal{L}_{\text{align}} = \frac{1}{2}\left[\text{CE}(S, I) + \text{CE}(S^\top, I)\right], \tag{17}$$

where I denotes the identity matching over the paired concepts.

Auxiliary weighting. Unless otherwise specified, we use base coefficients $\lambda_{\text{cons}} = 0.05$ and $\lambda_{\text{align}} = 0.01$. In the final model, each auxiliary loss $\mathcal{L}_a$ is further reweighted using a learned uncertainty parameter:

$$\mathcal{L}_a^{\text{uw}} = \lambda_a \left(\exp(-s_a)\mathcal{L}_a + s_a\right), \tag{18}$$

where s_a is a learned scalar parameter for auxiliary loss a. The supervised prediction loss remains unweighted, and the auxiliary objectives are used only to regularize local-global compatibility. The final implemented objective is therefore

$$\mathcal{L} = \mathcal{L}_{\text{pred}} + \mathcal{L}_{\text{cons}}^{\text{uw}} + \mathcal{L}_{\text{align}}^{\text{uw}}. \tag{19}$$

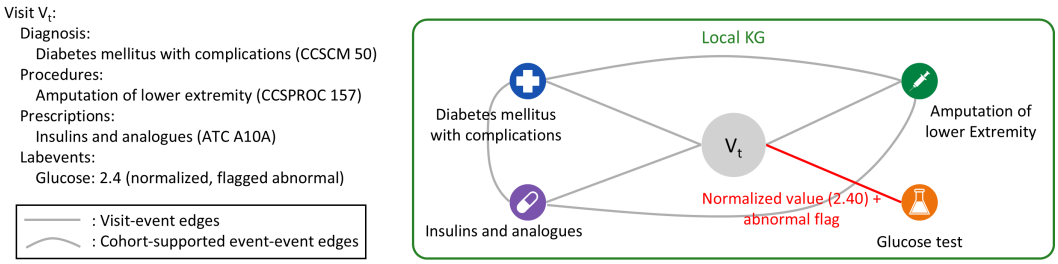

Figure 4: An example of local graph construction. The visit node is connected to every observed event. Event-event edges are included only when supported by the global graph constructed from the training cohort.

B Additional Implementation Details

This section summarizes implementation details that are omitted from the main text for compactness. These details clarify graph construction, visit readout, prototype assignment, laboratory feature handling, and the final selected hyperparameters.

B.1 Graph Edge Construction and Normalization

The global graph is constructed from symmetric concept associations estimated from visit-level co-occurrence statistics. For graph message passing, these retained concept associations are instantiated as directed edges. Similarly, local event-event relations are derived from the retained cohort-level associations and instantiated as directed message-passing edges in each visit-induced graph.

We apply a minimum-support filter before log-lift thresholding. In all final experiments, the minimum support is set to 5, and the support-scaled log-lift threshold is set to $\tau = 0.5$. After thresholding, retained edge scores are normalized per source event by the maximum retained neighbor score:

$$w(u, v) = \frac{\tilde{s}(u, v)}{\max_{v' \in \mathcal{N}^+(u)} \tilde{s}(u, v')}, \tag{20}$$

where $\mathcal{N}^+(u)$ denotes the retained neighbors of source event u after thresholding. This normalization is used in the final model and is not an optional post-processing step.

Threshold sensitivity. As a lightweight sensitivity check, we examined the effect of the support-scaled log-lift threshold on the mortality validation set. As shown in Table 4, increasing τ from 0.5 to 1.0 reduced both AUPRC and AUROC. Based on validation performance, we used $\tau = 0.5$ in the final configuration.

Table 4: Sensitivity of the global graph threshold τ on mortality prediction.

Threshold τ	AUPRC	AUROC
0.5	**17.88**	**92.69**
1.0	16.97	92.45

B.2 A simple example of local graph construction

Consider a visit (V_t) containing diabetes mellitus with complications (diagnosis), lower-extremity amputation (procedure), insulins and analogues (prescription), and a glucose measurement (laboratory event). Figure 4 represents the constructed local graph. The local graph contains (V_t) and one node for each observed event. Every event is connected to (V_t), whereas direct event-event edges are added only for cohort-supported pairs. In this example, no direct glucose-event edge is retained. The glucose node represents the shared laboratory identity, while its measured value and result flag are specific to this occurrence. The local event representation, matched global representation, and type-aware prototype term are subsequently combined through gated residual fusion, and the fused event representations are pooled into the visit representation.

B.3 Visit-level Readout

The local and global visit representations are constructed differently. For the local branch, $z_{i,t}^{(l)}$ is taken from the visit anchor node produced by local graph message passing. In the current implementation, event-level fusion updates event node representations but does not perform an additional graph propagation step from fused event nodes back to the visit anchor. Thus, $z_{i,t}^{(l)}$ is not obtained by mean pooling fused event embeddings.

For the global branch, the entire global KG is first encoded, and the encoded global concept embeddings corresponding to the events in visit $V_{i,t}$ are gathered. The global visit representation $z_{i,t}^{(g)}$ is then computed by mean pooling these matched global concept embeddings over valid events in the visit. Although the fused event embeddings are not propagated back to the visit anchor, they are used to form the local type summaries that define $Z_i^{(p)}$, and therefore remain connected to the final prediction objective through visit-level fusion.

B.4 Prototype Assignment

Prototype assignment uses cosine similarity implemented as normalized dot product. Before computing assignment scores, both the query representation and the prototype vectors are ℓ_2-normalized along the feature dimension:

$$\mathrm{sim}(h, p) = \left\langle \frac{h}{\|h\|_2}, \frac{p}{\|p\|_2} \right\rangle. \tag{21}$$

The resulting similarity scores are divided by the prototype temperature γ before the softmax assignment.

B.5 Laboratory Edge Features

Laboratory values are incorporated through value-aware edge features in the local graph encoder. For each local edge, we concatenate relation-kind embeddings, laboratory flag embeddings, laboratory category embeddings, laboratory type embeddings, and normalized numerical laboratory scalars when available. Numerical laboratory values are normalized using training-split statistics before graph construction and are used as continuous edge scalar attributes. The concatenated feature vector is projected to the hidden dimension by an edge encoder:

$$a_{uv} = \phi_{\mathrm{edge}}\left([e_{\mathrm{kind}} \,\|\, e_{\mathrm{flag}} \,\|\, e_{\mathrm{cat}} \,\|\, e_{\mathrm{type}} \,\|\, s_{uv}]\right), \tag{22}$$

where s_{uv} denotes the normalized scalar laboratory value when available.

The resulting edge representation $a_{uv} \in \mathbb{R}^d$ is passed to the local edge-aware GAT layer as an edge attribute. Accordingly, relation type and laboratory value information participate in attention-based local message passing, rather than being appended as ordinary node features. Categorical laboratory observations are represented through categorical value identifiers and laboratory type embeddings, whereas numerical observations use normalized scalar edge attributes.

B.6 Prototype-informed Visit-level Correction

In addition to the event-level prototype term used in post-event fusion, MVP-EHR uses type-summary prototypes to construct a visit-level prototype signal. For each visit and event type r, we compute a local type summary by mean pooling the fused local event embeddings $h_{i,t,j}^{(f)}$ of that type, and a global type summary by mean pooling the matched global concept embeddings $h_{i,t,j}^{(g)}$. If a type is absent in a visit, its summary is represented by a zero vector and excluded from the final type aggregation using a presence mask.

Both the local and global type summaries query the same type-specific prototype bank P_r. The resulting local and global prototype summaries are fused with their absolute difference through a prototype fusion MLP. We then combine the local type summary, global type summary, fused prototype summary, and local-global absolute difference through a final fusion MLP followed by layer normalization, producing a type-specific fused summary. The visit-level prototype signal $z_{i,t}^{(p)}$ is

obtained by averaging the valid type-specific fused summaries:

$$z_{i,t}^{(p)} = \frac{1}{|\mathcal{R}_{i,t}|} \sum_{r \in \mathcal{R}_{i,t}} \psi_r(i,t), \tag{23}$$

where $\mathcal{R}_{i,t}$ is the set of event types present in visit $V_{i,t}$ and $\psi_r(i,t) \in \mathbb{R}^d$ denotes the fused summary for type r. Stacking $z_{i,t}^{(p)}$ over visits yields $Z_i^{(p)} \in \mathbb{R}^{T_i \times d}$, which is injected into the visit-level fusion module described in Appendix B.7.

B.7 Visit-level Fusion Module

The visit-level fusion module is implemented as a two-stage fusion block. Let $Z_i^{(l)}, Z_i^{(g)} \in \mathbb{R}^{T_i \times d}$ denote the local and global visit sequences for patient i, and let $Z_i^{(p)} \in \mathbb{R}^{T_i \times d}$ denote the prototype-informed visit-level signal. We first compute a cross-attended global context:

$$\hat{Z}_i^{(g)} = \mathrm{CrossAttn}\left(Z_i^{(l)}, Z_i^{(g)}\right), \tag{24}$$

where $Z_i^{(l)}$ provides queries and $Z_i^{(g)}$ provides keys and values with visit masking. A dimension-wise gate then mixes the local visit sequence with the cross-attended global context:

$$\beta_i = \sigma\left(W_\beta[Z_i^{(l)} \,\|\, Z_i^{(g)} \,\|\, \hat{Z}_i^{(g)}] + b_\beta\right), \qquad Z_i^{(\mathrm{mix})} = \beta_i \odot Z_i^{(l)} + (1-\beta_i) \odot \hat{Z}_i^{(g)}. \tag{25}$$

The second stage applies a prototype-informed residual correction. We form the visit-level descriptor

$$R_i = \left[Z_i^{(l)} \,\|\, Z_i^{(g)} \,\|\, Z_i^{(p)} \,\|\, |Z_i^{(l)} - Z_i^{(g)}| \,\|\, |Z_i^{(l)} - Z_i^{(p)}| \,\|\, |Z_i^{(g)} - Z_i^{(p)}|\right]. \tag{26}$$

The final multi-view visit sequence is

$$Z_i^{(\mathrm{mv})} = \mathrm{LN}\left(Z_i^{(\mathrm{mix})} + \phi_{\mathrm{visit}}(R_i)\right), \tag{27}$$

where ϕ_{visit} is an MLP with dropout. Thus, visit-level fusion first controls how much cross-view global context is mixed into the local visit sequence, and then injects prototype-derived information as a residual correction rather than as a replacement.

B.8 Inter-visit Time-gap Encoding

For the first visit, the time-gap index is set to 0. For each subsequent visit V_t, we calculate the interval between the previous visit's discharge and the current admission and convert it into a weekly bucket:

$$g_t = \left\lfloor \frac{\mathrm{admit}(V_t) - \mathrm{discharge}(V_{t-1})}{\text{7 days}} \right\rfloor + 1.$$

The index is clipped to the 64 available bins (0–63). Its learned embedding is added to the corresponding fused visit representation before the temporal Transformer. Consequently, patient trajectories with the same visit order but different inter-visit intervals receive different temporal inputs. This encoding represents elapsed time at weekly resolution, while intervals beyond the maximum bin share the final embedding.

B.9 Hyperparameter Selection

Hyperparameters were selected through a staged validation search over the candidate settings shown in Table 5. Starting from a reference configuration, we varied one hyperparameter group at a time, retained the setting selected by the primary validation metric, and then evaluated the next group. The tuning covered graph construction, encoder capacity, prototype design, fusion, and optimization. We subsequently retained the selected common architecture across all five tasks; only the task output head, batch size, and validation checkpoint criterion were task dependent. The following table summarizes the evaluated candidates and selected settings.

Table 5: Hyperparameter candidates and selected settings for MVP-EHR.

Hyperparameter	Selected value	Search range
Hidden dimension	64	$\{32, 64, 96\}$
Local graph encoder layers	2	$\{1, 2, 3\}$
Global graph encoder layers	2	Fixed
Temporal Transformer layers	2	$\{1, 2, 3\}$
Shared prototypes per event type	16	$\{8, 16, 32\}$
Support-scaled log-lift threshold τ	0.5	$\{0.5, 1.0\}$
Minimum concept-pair support	5	Fixed
Dropout	0.10	$\{0.05, 0.10, 0.15\}$
Weight decay	10^{-5}	$\{10^{-5}, 5 \times 10^{-5}, 10^{-4}\}$
Optimizer	Adam	Fixed
Learning rate	10^{-3}	$\{5 \times 10^{-4}, 10^{-3}, 1.5 \times 10^{-3}\}$
Batch size	256	$\{128, 256\}$
Post-event fusion temperature	1.0	$\{0.7, 1.0, 1.3\}$
Concept alignment temperature τ_{align}	0.2	Fixed
Base consistency weight λ_{cons}	0.05	$\{0.02, 0.03, 0.04, 0.05, 0.10\}$
Base concept alignment weight λ_{align}	0.01	Fixed

Table 6: Comparison with the utilization-only XGBoost baseline.

Task	Metric	MVP-EHR	Utilization-only XGBoost
Mortality	AUPRC	**17.60**	10.65
	AUROC	**93.36**	88.75
Readmission	AUPRC	**56.08**	44.77
	AUROC	**73.00**	65.88
Length of stay	AUROC	**82.09**	70.22
	F1	**29.32**	14.74
Drug recommendation	AUPRC	**75.70**	58.31
	AUROC	**96.24**	92.85
	F1	**64.80**	39.07
Phenotype prediction	AUPRC	**79.51**	34.44
	AUROC	**93.72**	87.66
	F1	**65.65**	8.88

B.10 Compute Resources

All reported experiments were conducted on a single NVIDIA RTX PRO 6000 GPU with 96 GB memory and an AMD Ryzen Threadripper PRO 9975WX CPU. Runtime varied across tasks, model variants, and early stopping schedules, but each MVP-EHR training run required approximately 1–5 GPU-hours on a single GPU.

C Additional Analyses

C.1 Utilization-only Baseline

To assess whether healthcare utilization statistics alone can explain the predictive performance, we compare MVP-EHR with a utilization-only XGBoost baseline. The baseline uses the number of visits, diagnosis, procedure, prescription, and laboratory event counts, the number of abnormal laboratory results, and the mean and maximum inter-visit intervals. It does not use event identities, laboratory values, or demographic variables. We use the same patient split as MVP-EHR and select the model using the task-specific primary validation metric. As represented in Table 6, the utilization-only

Table 7: Effect of removing occurrence-specific laboratory result attributes.

Task	Metric	MVP-EHR	w/o laboratory results
Mortality	AUPRC	**17.60**	15.82
	AUROC	**93.36**	93.05
Readmission	AUPRC	56.08	**56.39**
	AUROC	73.00	**73.36**
Length of stay	AUROC	82.09	**82.18**
	F1	29.32	**30.00**
Drug recommendation	AUPRC	**75.70**	75.34
	AUROC	**96.24**	96.14
	F1	**64.80**	63.82
Phenotype prediction	AUPRC	**79.51**	56.42
	AUROC	93.72	**94.20**
	F1	**65.65**	41.93

Table 8: Mortality prediction across diagnosis-frequency subgroups.

Subgroup	N	Positive	Prevalence	AUPRC	AUROC	P@1%	P@5%
Common	4,104	26	0.63	20.73	91.56	21.43	8.25
Intermediate	4,206	62	1.47	21.16	90.91	27.91	13.74
Rare	4,222	47	1.11	14.38	94.95	16.28	13.21

baseline shows predictive signal but remains below MVP-EHR across all five tasks. This result suggests that healthcare utilization statistics alone do not account for the performance of the full model.

C.2 Effect of Laboratory Result Attributes

To distinguish laboratory identity from occurrence-specific result information, we evaluate a variant that retains laboratory concept nodes and visit-laboratory relations but removes numerical values, categorical results, and abnormal-result flags from the corresponding edge attributes. As shown in Table 7, the effect of laboratory result attributes is task-dependent. Removing these attributes has limited effects on readmission and length-of-stay prediction, while mortality AUPRC and drug-recommendation performance decrease modestly. The largest reduction appears in phenotype prediction, where AUPRC decreases from 79.51 to 56.42 and F1 decreases from 65.65 to 41.93. These results indicate that occurrence-specific laboratory results provide predictive information beyond laboratory identity alone.

C.3 Diagnosis-frequency Subgroup Analysis

To examine the patient subgroups for which the model performs well or not, we analyze whether mortality prediction varies with the frequency of patients' diagnoses in the training cohort. Specifically, we divide the test patients into common, intermediate, and rare diagnosis-frequency subgroups. For each diagnosis, we first calculate its add-one-smoothed patient prevalence using only the training cohort. We then assign each patient a rarity score by averaging the negative log prevalence of their unique diagnoses. Lower scores indicate profiles dominated by frequently observed diagnoses, whereas higher scores indicate profiles containing less frequent diagnoses. The 33rd and 67th percentiles of the training-patient rarity scores define the three subgroups, and these thresholds are applied unchanged to test patients. Patients without a mapped diagnosis are excluded from this analysis.

Table 8 presents the subgroup analysis results, where P@$k\%$ denotes precision among the top-scoring $k\%$ of patients within each subgroup. MVP-EHR achieves similar AUPRC in the common and intermediate groups. The rare subgroup shows the highest AUROC but lower AUPRC and P@1%,

suggesting that the model preserves overall ranking performance but produces more false positives among the highest-scoring patients. Because the subgroup prevalence and number of positive cases differ, these results should be interpreted cautiously.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the proposed global-local graph decomposition, prototype-aligned fusion, and five-task MIMIC-IV evaluation. The claims are supported by the main comparison, ablations, and prototype analyses.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The Discussion section includes a limitations paragraph covering single-dataset evaluation, lift-based graph construction, scalability to larger ontologies, lightweight prototypes, and task-dependent performance.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper does not present theoretical results, theorems, or formal proofs. It is an empirical methods paper.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper specifies the dataset, split, tasks, metrics, architecture, graph construction, losses, and hyperparameters. Additional implementation details are provided in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code and configuration files are provided as an anonymized supplementary package for review. MIMIC-IV is not redistributed and must be obtained through PhysioNet.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper reports the split, task definitions, metrics, optimizer, learning rate, batch size, model dimensions, graph threshold, and selected hyperparameters. Additional details are in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The main comparison reports bootstrap standard deviations over 100 resamples. These are described as estimates of test-set variability under the shared evaluation protocol.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The appendix reports the GPU and CPU hardware and approximate per-run GPU-hours. All experiments were run on a single GPU without distributed training.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The study uses de-identified MIMIC-IV data under its established access requirements. It does not involve deployment, intervention, or prospective decision-making.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: The paper studies clinical prediction as a potential healthcare application, while noting that the model is evaluated only retrospectively on MIMIC-IV and requires external validation before clinical use.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release high-risk models, scraped datasets, or sensitive patient data. MIMIC-IV is not redistributed.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The paper cites MIMIC-IV and all baseline methods used for comparison. MIMIC-IV is used under its official access process and is not redistributed.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.

- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper introduces a new model implementation and provides an anonymized supplementary code package. The restricted MIMIC-IV dataset is not redistributed.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The study does not involve crowdsourcing or direct interaction with human subjects. It uses retrospective, de-identified EHR data.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

Justification: The study uses de-identified MIMIC-IV data and does not involve new human-subject recruitment or intervention. MIMIC-IV is distributed through an established credentialing process.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are not used as part of the proposed method. The core method is based on graph construction, prototype-aligned representation learning, and temporal prediction.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.