MedFiTRG: Jointly Learning Dynamic Temporal and Cross-Patient Graphs for Clinical Outcome Prediction

Shivani Gupta[1*], **Hari Om Kumar**[1], **Joydeep Chandra**[1]
[1]Indian Institute of Technology Patna
{shivani_2221cs18@iitp.ac.in, 2201ai12_hari@iitp.ac.in, joydeep@iitp.ac.in

Abstract

Integrating heterogeneous clinical modalities, structured electronic health records (EHRs), clinical text, and medical imaging is crucial for reliable clinical prediction, yet real-world data are often sparse and imbalanced. Furthermore, prior approaches treat temporal dynamics and inter-patient relationships in isolation, overlooking the dynamic interaction of patient trajectories across populations. We introduce a modality-enhanced dynamic temporal relational graph (MedFiTRG), a unified framework that jointly models sparsity, temporal dynamics, and cross-patient relational dependencies. MedFiTRG leverages modulated graph neural networks (MGNN) to learn modality-aware embeddings, enabling meaningful representation of sparse modalities through adaptive feature modulation. These embeddings are integrated into a temporal relational graph (TRG), where directed intra-patient edges capture longitudinal progression and dynamic inter-patient edges model population-level similarities for synchronized temporal-relational reasoning. Extensive experiments on large-scale real-world datasets across four clinical tasks demonstrate that MedFiTRG achieves superior or comparable performance against state-of-the-art baselines, improving Macro-F1 from 0.155 to 0.310 for length of stay (LOS) classification and achieving an AUROC of 0.939 for mortality prediction (↑6.45%). The code is available at https://anonymous.4open.science/r/MedFiTRG-2714

1 Introduction

Electronic Health Records (EHRs) have become a crucial component in modern healthcare, facilitating accurate and timely clinical decision-making through the integration of heterogeneous patient data. With the advancement of deep learning techniques, EHRs now power predictive models across tasks such as risk assessment [Wang *et al.*, 2024; Bai *et al.*, 2018; Luo *et al.*, 2020], clinical decision support [Rajkomar *et al.*, 2018; Gupta *et al.*, 2025; Chen *et al.*, 2024a], and personalized medicine [Zhang *et al.*, 2023; Ahmed *et al.*, 2023; Zhu *et al.*, 2024]. To achieve this, clinical decision support (CDS) systems increasingly rely on the fusion of diverse modalities, including structured EHR, medical imaging, and unstructured clinical narratives [Al-Zoghby *et al.*, 2025; Jandoubi and Akhloufi, 2025]. Each modality provides a distinct and complementary view of patient health, EHR captures longitudinal physiological measurements and interventions, clinical text encodes diagnostic reasoning, and medical images such as chest X-rays (CXR) reveal latent pathological patterns. While modality-specific models have advanced significantly, real-world clinical decision making requires effective integration across modalities and time [Siam *et al.*, 2025; Teles *et al.*, 2025]. In critical care settings in particular, accurate prediction of outcomes such as multi-disease risk, length of stay (LOS), and mortality depends on jointly modeling temporal dynamics, heterogeneous data types, and complex patient-level interactions. Despite growing interest in multimodal clinical prediction, several fundamental challenges remain:

Modality Heterogeneity and Semantic Misalignment: Multimodal medical data consist of fundamentally different representations, where EHR time-series encode longitudinal clinical events, chest X-rays capture spatial visual patterns, and clinical notes describe subjective and contextual observations. These modalities operate at different semantic granularities and levels of reliability, making direct fusion a challenging task.

Modality Sparsity and Imbalanced Clinical Evidence: In real-world clinical settings, modality availability is highly imbalanced, with structured EHR data present for nearly all visits, while chest X-ray images are available for only a small subset of encounters. As a result, conventional models trained with uniform aggregation or static fusion strategies tend to ignore or underweight sparse modalities, effectively treating them as auxiliary noise rather than informative clinical evidence. This imbalance results in the underuse of critical visual information, particularly in acute scenarios where imaging provides essential diagnostic cues, posing a significant challenge for learning well-calibrated multimodal representations.

*Corresponding author

Dynamic Integration of Temporal and relational Dependency: Existing studies on multimodal clinical prediction typically decouple temporal and relational reasoning, first modeling intra-patient evolution, then applying a static or global patient graph. Such sequential or hierarchical designs overlook synchronous temporal-relational dependencies, where visits from different patients evolve and interact simultaneously over time. The absence of frameworks capturing these visit-level, cross-patient temporal connections limits its current models' ability to reflect real-world clinical dynamics and population-level disease progression.

To address the above-mentioned challenges, we introduce modality-enhanced dynamic temporal relational graph (MedFiTRG), a unified framework to integrate modality-aware GNN encoding with a learnable temporal relational graph (TRG) for dynamic, multimodal patient modeling across time and population. MedFiTRG jointly encodes EHR, text, and imaging into a global context vector and refines each modality using modulated graph neural networks (MGNN) inspired by GNN-FiLM [Brockschmidt, 2020]. Unlike conventional GNN-FiLM, which statically modulates node features, our MGNN dynamically learns FiLM parameters from fused multimodal and relational contexts, enabling context-aware, hierarchical modulation across modalities, time, and patient graphs. This modulation enables context-aware alignment of heterogeneous feature spaces and meaningful utilization of sparse modalities. Subsequently, MedFiTRG builds a TRG, where each node represents a patient visit, with directed intra-patient edges capturing temporal evolution and learnable inter-patient edges encoding dynamic cross-patient similarity. This design allows synchronized reasoning across time and population, capturing multi-dimensional temporal and relational dependencies essential for holistic clinical prediction. The key contributions of this work are summarized as follows:

- **MGNN:** We employ a modulated GNN to learn modality-specific representations for EHR, chest X-rays, and clinical text. This adaptive modulation mechanism performs feature-wise scaling and shifting based on global contextual information, effectively aligning heterogeneous feature spaces and mitigating semantic misalignment across modalities.
- **TRG:** We construct a TRG, where intra-patient edges capture longitudinal evolution and inter-patient edges are dynamically learned through a similarity-based threshold, facilitating adaptive relational reasoning across patients.
- **Contextually Modulated Message Passing:** We perform message passing using MGNN over both temporal and relational edges, jointly capturing temporal progression, population-level relations, and modality interactions in an end-to-end manner.
- **Extensive Experimental Validation:** Extensive experiments on large-scale real-world datasets, MIMIC-IV, MIMIC-CXR-JPG, and MIMIC-IV-NOTE, show that MedFiTRG achieves superior or comparable performance against strong multimodal baselines, improving LOS Macro-F1 from 0.155 to 0.310 and achieving an AUROC of 0.939 for mortality prediction.

2 Related Work

In this section, we conduct a comprehensive review of the literature.

Multimodal Clinical Representation Learning: Recent advances in multimodal learning have demonstrated that jointly leveraging imaging, clinical text, and structured EHR data substantially improves clinical prediction and patient modeling [Xu *et al.*, 2023; Zhong *et al.*, 2024; Zhang *et al.*, 2022; Yao *et al.*, 2024; Ruan *et al.*, 2025]. Early-fusion frameworks such as HAIM [Soenksen *et al.*, 2022] and HEALNet [Hemker *et al.*, 2024] integrate pre-trained modality encoders with attention-based fusion to align heterogeneous features while preserving modality-specific semantics. Subsequent models like VecoCare [Xu *et al.*, 2023], M3Care [Zhang *et al.*, 2022], and DrFuse [Yao *et al.*, 2024] further refine cross-modal learning through adaptive weighting and contrastive consistency objectives, enhancing robustness across diverse clinical outcomes. However, most existing approaches struggle with incomplete or irregularly sampled modalities, a common limitation in real-world EHR systems. MedFuse [Hayat *et al.*, 2022] addresses this by employing LSTM-based fusion to handle partial inputs, while Lee et al. [Lee *et al.*, 2023] design modality-aware attention mechanisms to mitigate EHR data missingness. Similarly, DrFuse [Yao *et al.*, 2024] introduces shared–distinct embedding spaces to improve modality alignment under sparse conditions. Despite these efforts, current methods often fuse modalities at feature level without explicitly modeling temporal progression or inter-patient relationships.

Temporal and Graph-based Clinical Prediction: Temporal and graph-based models have become pivotal for capturing both the dynamic evolution and relational structure of clinical data. Sequence-based architectures such as HiTANet [Luo *et al.*, 2020] and StageNet [Gao *et al.*, 2020] effectively model intra-patient temporal dependencies but lack mechanisms to represent population-level relationships. To jointly capture sequential and structural signals, TRANS [Chen *et al.*, 2024b] formulates heterogeneous temporal graphs that represent visits and medical events as interconnected nodes, while Multi-TA [Sohn *et al.*, 2024] enhances temporal robustness via BERT-based contextual embeddings for early disease detection. Beyond patient timelines, Tang et al. [Tang *et al.*, 2023] introduce multimodal spatiotemporal GNNs that leverage similarity graphs to propagate information across related patients, and Xu et al. [Xu *et al.*, 2024a] design temporal neighbor transformers to handle asynchronous and incomplete data using missingness-aware prompts.

In summary, despite advances in multimodal fusion, temporal modeling, and graph-based learning, existing methods still face challenges with heterogeneous and incomplete clinical data. Most frameworks struggle to generalize under modality imbalance and lack scalable, interpretable architectures for fine-grained cross-modal reasoning, limiting their

effectiveness in real-world healthcare settings.

3 Problem Formulation

Let each patient p_i $(i = 1, 2, \ldots, N)$ have a chronological sequence of clinical visits $V_i = \{v_{i,1}, v_{i,2}, \ldots, v_{i,T_i}\}$, where T_i denotes the total number of visits for patient p_i. Each visit $v_{i,t}$ consists of multimodal data, including structured EHR features $\mathbf{x}_{i,t}^{ehr}$, clinical text embeddings $\mathbf{x}_{i,t}^{text}$, and imaging features $\mathbf{x}_{i,t}^{img}$.

Our goal is to learn a predictive function f_θ that, for each patient p_i and visit $v_{i,t}$, estimates the clinical outcome $\hat{y}_{i,t}$ using all available information up to and including visit t:

$$\hat{y}_{i,t} = f_\theta\big(v_{i,1}, v_{i,2}, \ldots, v_{i,t}\big), \quad \text{for } t = 1, 2, \ldots, T_i.$$

In this formulation, the model performs per-visit prediction, where each prediction at time t leverages the patient's historical trajectory up to that point. For the first visit ($t = 1$), the model relies solely on the current visit's data, whereas subsequent predictions progressively incorporate both past and current multimodal representations. During training, the model learns from all patients $\{V_1, V_2, \ldots, V_N\}$ to optimize θ for accurate visit-level outcome prediction. It predicts a clinical outcome $\hat{y}_{i,t}$ (e.g., in-hospital mortality (IHM), readmission (REA), length of stay (LOS), and phenotyping (PHE)) by jointly modeling temporal evolution and relational dependencies across patients.

4 Methodology

MedFiTRG is a unified multimodal framework that jointly models modality-specific representations, longitudinal patient trajectories, and population-level dependencies. The architecture follows a hierarchical fusion design based on feature-wise linear modulation (FiLM), enabling context-aware interaction across modalities, time, and patients. Figure 1 illustrates the overall architecture.

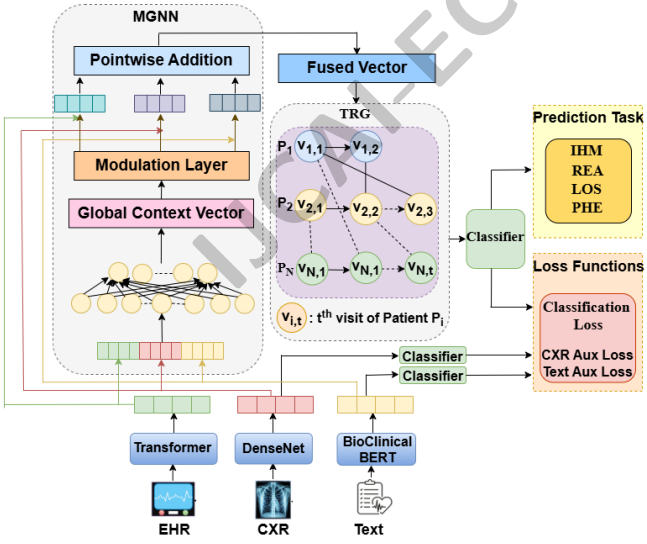


Figure 1: Overview of the MedFiTRG framework.

4.1 Modality-Aware Representation Learning

Consider a patient cohort $\mathcal{P} = \{p_1, \ldots, p_N\}$, where each patient p_i has a sequence of visits $V_i = \{v_{i,1}, v_{i,2}, \ldots, v_{i,T_i}\}$. Each visit consists of structured EHR features x_i^{ehr}, chest X-ray images x_i^{img}, and clinical text x_i^{txt}. Modality-specific encoders transform each input into a latent embedding $\mathbf{e}_{i,t}^m \in \mathbb{R}^{512}$, where $m \in \{\text{ehr}, \text{img}, \text{text}\}$. To robustly handle missing modalities, we replace absent inputs with learnable modality-specific missing tokens, allowing the model to optimize default representations rather than relying on zero-imputation. Auxiliary prediction heads are attached to each unimodal encoder to provide deep supervision and enforce discriminative feature learning prior to fusion.

4.2 Symmetric Modality Fusion via FiLM

To achieve coherent cross-modal alignment, we adopt a symmetric FiLM-based fusion strategy that treats each modality equally while conditioning on the shared multimodal context. For each visit, modality embeddings are concatenated and passed through a lightweight multilayer perceptron (MLP) to produce a global visit context:

$$\mathbf{g}_{i,t} = \text{MLP}\big([\mathbf{e}_{i,t}^{ehr} \| \mathbf{e}_{i,t}^{text} \| \mathbf{e}_{i,t}^{img}]\big). \tag{1}$$

This global context captures interactions and dependencies across modalities, effectively summarizing the visit as a unified latent descriptor. The FiLM generator g_ϕ then produces modality-specific scaling and shifting coefficients (γ_m, β_m) based on this shared context:

$$[\gamma_m, \beta_m] = g_\phi(\mathbf{g}_{i,t}), \quad m \in \{\text{ehr}, \text{text}, \text{img}\}. \tag{2}$$

Each modality embedding is subsequently modulated as:

$$\mathbf{z}_{i,t}^m = \gamma_m \odot \mathbf{e}_{i,t}^m + \beta_m. \tag{3}$$

Through this modulation, the FiLM mechanism adaptively emphasizes or suppresses modality features according to their relevance in the overall clinical context. The key advantage of this symmetric conditioning is that all modalities influence one another through the same contextual signal, promoting balanced contribution and preventing dominance by high-dimensional modalities such as imaging. The modulated embeddings are finally aggregated to yield a fused visit representation $\mathbf{z}_{i,t}$, which encodes both modality-specific evidence and their interdependencies in a compact form suitable for downstream temporal reasoning. The final fused visit representation is obtained via element-wise aggregation of the modulated modality embeddings:

$$\mathbf{z}_{i,t} = \mathbf{z}_{i,t}^{ehr} + \mathbf{z}_{i,t}^{img} + \mathbf{z}_{i,t}^{txt}, \tag{4}$$

where the summation denotes point-wise addition across modalities of equal dimensionality. This formulation preserves modality-specific semantics while enabling joint representation learning in a shared latent space.

4.3 Temporal MGNN for Longitudinal Modeling

While visit-level fusion provides a comprehensive snapshot of multimodal information, patient trajectories evolve dynamically over time. Diseases progress gradually, and their

manifestations in EHR, imaging, and text often exhibit nonlinear temporal dependencies. To model such evolution, we adopt a first-order Markov assumption, where the patient's current clinical state depends primarily on its immediate historical context [Choi *et al.*, 2016]. This assumption aligns with common temporal patterns in longitudinal EHR data, where each visit reflects an incremental transition from the most recent medical state.

Under this Markovian formulation, we introduce a temporal MGNN that performs context-aware reasoning over sequential visits. Specifically, each visit embedding $\mathbf{z}_{i,t}$ is conditioned on its immediate predecessor $\mathbf{z}_{i,t-1}$ through a FiLM-based modulation mechanism:

$$\mathbf{h}_{i,t}^{temp} = \mathrm{FiLM}(\mathbf{z}_{i,t-1}, \mathbf{z}_{i,t}) + \mathbf{z}_{i,t}, \tag{5}$$

where the FiLM operation applies a feature-wise affine transformation conditioned on the previous visit embedding:

$$\mathrm{FiLM}(\mathbf{z}_{i,t-1}, \mathbf{z}_{i,t}) = \gamma(\mathbf{z}_{i,t-1}) \odot \mathbf{z}_{i,t} + \beta(\mathbf{z}_{i,t-1}), \tag{6}$$

with $\gamma(\cdot)$ and $\beta(\cdot)$ being learnable functions (e.g., MLPs) that generate feature-wise scaling and shifting parameters from the context $\mathbf{z}_{i,t-1}$. The previous visit representation $\mathbf{z}_{i,t-1}$ thus serves as a dynamic conditioning context, modulating the current visit embedding $\mathbf{z}_{i,t}$ to capture first-order dependencies between consecutive visits while filtering redundant temporal noise and emphasizing clinically meaningful deviations.

The residual connection preserves current visit information, enabling stable gradient flow and preventing historical overdominance. This formulation disentangles short-term variations from long-term progression, capturing clinically relevant transitions such as treatment response or acute deterioration. Unlike recurrent models (e.g., LSTMs or GRUs) that implicitly encode sequential memory, our Markov-based Temporal MGNN explicitly parameterizes inter-visit influence through learnable modulation functions, enhancing interpretability and robustness to irregular sampling intervals.

4.4 Adaptive Social Relational Fusion

Beyond temporal self-dependence, patients within a population exhibit shared clinical patterns, identified through embedding-level similarities that reflect underlying demographic, phenotypic, and comorbidity factors. To capture these cross-patient relationships, we construct a dynamically learned patient similarity graph that allows information propagation across related individuals. For each pair of patients (i, j) within a mini-batch, we compute an adaptive edge weight:

$$A_{ij} = \sigma\big(\tau \cdot (\cos(\mathbf{h}_i^{temp}, \mathbf{h}_j^{temp}) - \theta)\big), \tag{7}$$

where $\cos(\cdot)$ denotes cosine similarity, θ is a learnable threshold controlling edge sparsity, and τ is a temperature parameter regulating the sharpness of connectivity. This adaptive formulation enables the graph topology to evolve with training, reflecting emergent clinical clusters or patient subtypes. The aggregated social context is then integrated via MGNN:

$$\mathbf{h}_i^{social} = \mathbf{h}_i^{temp} + \tanh\left(\sum_j A_{ij}\right) \cdot \left[\gamma_i \odot \sum_j A_{ij}\mathbf{h}_j^{temp} + \beta_i\right], \tag{8}$$

where $(\gamma_i, \beta_i) = g_\phi([\mathbf{h}_i^{temp} \| \sum_j A_{ij}\mathbf{h}_j^{temp}])$.

In this stage, each patient representation $\mathbf{h}_i^{temp}$ is refined by adaptively modulating information aggregated from its clinically similar neighbors. The FiLM function acts as a contextual gate, determining how social information should alter the patient's individual trajectory. The inclusion of $\tanh(\sum_j A_{ij})$ provides a soft normalization that controls the influence of neighborhood density. This design ensures that population-level patterns, such as treatment responses or risk factors shared among subgroups, inform individual-level predictions without overwhelming personalized signals.

Overall, this adaptive social relational fusion bridges individual and cohort-level learning: temporal FiLM captures intra-patient evolution, while graph-based FiLM models inter-patient associations, resulting in a robust and interpretable framework for multimodal clinical reasoning. The combined effect is a system capable of symmetric multimodal alignment, dynamic temporal conditioning, and socially aware graph reasoning, all unified through FiLM-driven modulation principles.

4.5 Outcome Prediction

The final representation $\mathbf{h}_i^{social}$ integrates information across multiple dimensions, capturing multimodal clinical features, longitudinal temporal evolution, and population-level relational dependencies. This embedding serves as a comprehensive patient state descriptor that reflects both individual health dynamics and contextual similarities within the cohort. To perform downstream clinical tasks such as REA, IHM, PHE, and LOS prediction, we employ a task-specific output head implemented as a linear projection followed by an appropriate activation function:

$$\hat{y}_{i,t} = f(\mathbf{W}_o \mathbf{h}_i^{social} + b_o), \tag{9}$$

where $\mathbf{W}_o$ and b_o denote learnable parameters of the output layer, $\mathbf{h}_i^{social}$ represents the social-aware patient representation, and $f(\cdot)$ denotes the task-specific activation function. Specifically, sigmoid activation is used for binary and multi-label tasks (REA, IHM, and PHE), while softmax activation is employed for LOS classification.

5 Loss Function Formulation

To address severe class imbalance across heterogeneous prediction tasks, we adopt a focal loss–based optimization scheme. The learning objectives are categorized into three groups: *binary*, *multi-label* and *multi-class* tasks.

Binary Tasks (IHM, REA): For binary classification, the model outputs a logit z converted to probability $p = \sigma(z)$. The focal loss is defined as:

$$\mathcal{L}_{\text{bin}}(z, y) = -\lambda_{\text{cls}}\, \alpha_t\, (1 - p_t)^\gamma \log(p_t), \tag{10}$$

where $p_t = p$ if $y = 1$ else $1 - p$, $\gamma = 2$ is the focusing factor, $\alpha_t \in \{\alpha, 1 - \alpha\}$ balances positive/negative samples, and $\lambda_{\text{cls}} = 1 + y(w_{\text{pos}} - 1)$ upweights rare positive cases.

Multi-Label Task (PHE): For phenotyping, each of the C labels is optimized independently using the same focal loss:

$$\mathcal{L}_{\text{phe}} = \sum_{c=1}^{C} \mathcal{L}_{\text{bin}}(z_c, y_c). \tag{11}$$

Multi-Class Task (LOS) For LOS, the task is formulated as a K-class classification problem with $K = 10$ discrete time buckets. Given logits $\mathbf{z} \in \mathbb{R}^K$, class probabilities are computed using the softmax function $p_k = \exp(z_k)/\sum_{j=1}^{K}\exp(z_j)$. With one-hot encoded targets $\mathbf{y} \in \{0,1\}^K$ and a precomputed class-weight vector $\mathbf{w}$, the multi-class robust focal loss is defined as

$$\mathcal{L}_{\text{los}} = -\sum_{k=1}^{K} w_k y_k \, \alpha(1-p_k)^{\gamma}\log(p_k), \tag{12}$$

where w_k denotes the inverse class-frequency weight, computed as $w_k = N/(K \cdot N_k)$, with N representing the total number of training samples, K the number of classes, and N_k the number of samples belonging to class k.

To promote modality-specific discriminative representations, auxiliary classifiers are attached to unimodal encoders and trained via deep supervision. Auxiliary losses are computed only over samples where the corresponding modality is available. The final objective minimizes the weighted sum of the fused prediction loss and auxiliary losses:

$$\mathcal{J} = \mathcal{L}_{\text{task}}(\hat{y}_{\text{fused}}, y) + \lambda_{\text{cxr}}\mathcal{L}_{\text{aux}}^{(\text{cxr})} + \lambda_{\text{text}}\mathcal{L}_{\text{aux}}^{(\text{text})}, \tag{13}$$

where λ_{cxr} and λ_{text} are learnable parameters.

6 Experiments

In this section, we detail the experimental setup and present a comprehensive analysis of the results.

6.1 Experimental Settings

Datasets

In this study, we leverage three complementary components of the MIMIC database ecosystem: MIMIC-IV, MIMIC-CXR-JPG, and MIMIC-IV-Note datasets [Johnson *et al.*, 2019; Johnson *et al.*, 2023]. These datasets share a common patient cohort, enabling seamless alignment across modalities. We extract longitudinal structured EHR data from MIMIC-IV, chest X-ray images from MIMIC-CXR, and clinical narratives from MIMIC-IV-Note to construct a unified, multimodal patient representation that captures temporal, visual, and textual clinical information. The resulting dataset is divided into training, validation, and testing subsets following a 70:10:20 ratio based on unique patient identifiers to ensure patient-level independence across splits. Summary statistics are reported in Table 1.

Task	# Patients	Missing rate per modality		
		EHR	Image	Text
IHM	26,318	0%	76.40%	7.49%
LOS	59,495	0%	85.16%	8.27%
PHE	59,798	0%	81.94%	8.30%
REA	55,712	0%	82.38%	8.06%

Table 1: Statistics of the datasets for multiple prediction tasks, including modality-wise missing rates.

Tasks

To comprehensively evaluate our framework, we benchmark it on four clinically relevant prediction tasks derived from the MIMIC-IV dataset:

- **In-hospital mortality (IHM):** Predicts whether a patient dies during the current ICU stay based on the first 48 hours of data (*binary classification*).
- **Length of stay (LOS):** Estimates the remaining ICU stay, categorized into 10 discrete intervals: one for stays under a day, seven for each subsequent day up to a week, one for stays between one and two weeks, and one for stays exceeding two weeks. (*multi-class classification*).
- **Phenotyping (PHE):** Identifies 25 acute conditions associated with an ICU episode (*multi-label classification*).
- **Readmission (REA):** Predicts hospital readmission within 30 days post-discharge (*binary classification*).

6.2 Baselines

To rigorously evaluate our proposed MedFiTRG framework, we benchmark it against several state-of-the-art multimodal models addressing missing modality and temporal integration. MedFuse [Hayat *et al.*, 2022] introduces an LSTM-based fusion module capable of handling both unimodal and multimodal inputs, enabling flexible integration across heterogeneous data sources. MT-Transformer [Ma *et al.*, 2022] employs a transformer architecture with attention masks to effectively fuse multimodal data under missing modality conditions. M3Care [Zhang *et al.*, 2022] imputes missing modality information within a latent representation space by leveraging neighboring patients with similar clinical patterns. MMF [Lee *et al.*, 2023] adopts a modality-aware attention mechanism combined with a skip bottleneck structure to learn robust EHR representations despite incomplete modalities. MultiModN [Swamy *et al.*, 2023] sequentially processes any subset of available modalities through modality-specific encoders, dynamically skipping absent inputs during inference. FlexCare [Xu *et al.*, 2024b] dynamically weights available modalities to handle missing data with continual learning on different tasks, and MoEHealth [Wang and Yang, 2025] uses a mixture-of-experts design for adaptive modality selection. For consistency, all baselines were adapted or extended to align with our task definitions and multimodal dataset configuration.

6.3 Evaluation Metrics

To comprehensively assess model performance across different clinical prediction tasks, we employ a combination of classification and ranking-based metrics. For binary classification tasks, we report the area Under the receiver operating characteristic curve (AUROC) and the area under the precision–recall curve (AUPRC). For multi-class prediction task, we compute both macro-AUROC and micro-AUROC. For the multi-label prediction task, we evaluate using macro-F1 and micro-F1 scores to capture both per-class and overall balance between precision and recall. Together, these metrics provide a robust evaluation of the model across different prediction tasks.

7 Experimental Results

In this section, we present and analyze the experimental findings to assess the effectiveness of our model.

7.1 Baseline Comparison

Table 2 presents the performance comparison between the proposed MedFiTRG model and various baseline models across various tasks. All experiments are repeated five times with different random seeds, and we report the mean and standard deviation of performance metrics. The best-performing scores are highlighted in bold, while the second-best are underlined for clarity. For LOS classification, it improves Macro-F1 from 0.155 to 0.310, demonstrating superior handling of minority classes. It also achieves an AUROC of 0.939 for mortality prediction (↑6.45%) and maintains competitive performance on readmission and phenotyping tasks, underscoring its strength in leveraging sparse multimodal clinical data for decision support.

7.2 Ablation Study

To evaluate the contribution of individual components in MedFiTRG, we conducted a comprehensive ablation study by systematically removing key modules from the full model. Specifically, we tested four variants: (1) w/o social graph, where inter-patient relational edges were removed, (2) w/o temporal graph, where intra-patient temporal connections were omitted, (3) w/o global context vector, where modality modulation with the global context vector was removed, and (4) w/o temporal and social graph, where both intra- and inter-patient dependencies were excluded. As summarized in Table 3, across all configurations, we observed consistent performance drops compared to the full MedFiTRG model. For LOS the removal of the temporal and social graph caused the largest decline, underscoring the importance of dynamic relational reasoning. For IHM the elimination of global context vector caused the lowest performance, indicating that MGNN-based feature modulation is crucial for aligning and balancing sparse modalities. These results collectively validate the effectiveness of each component in enhancing multimodal clinical prediction.

7.3 Impact of Different GNNs

To further assess the effectiveness of the MGNN architecture, we replaced it with standard GCN and GAT layers in both the global context generation stage and the temporal–social graph message-passing module. A noticeable degradation in performance was observed across different prediction tasks. The results indicate that conventional graph layers fail to capture fine-grained feature-wise dependencies and modality-specific dynamics, leading to weaker fusion and relational reasoning. In contrast, MGNN provides adaptive feature modulation through learnable scaling and shifting, enabling more expressive and context-aware information propagation, which proves essential for robust multimodal clinical prediction (Table 4).

7.4 t-SNE Plot

To further interpret the learned representations, we visualize the modality-fused embeddings of MedFiTRG using t-SNE for the IHM and LOS prediction tasks. The resulting plots reveal well-separated and compact clusters across outcome categories, indicating that the model effectively captures discriminative patient representations. This clear separation highlights MedFiTRG's ability to integrate multimodal and temporal–relational information into a coherent latent space, enhancing both prediction accuracy and interpretability in complex clinical settings. Figure 2 shows the t-SNE plot.

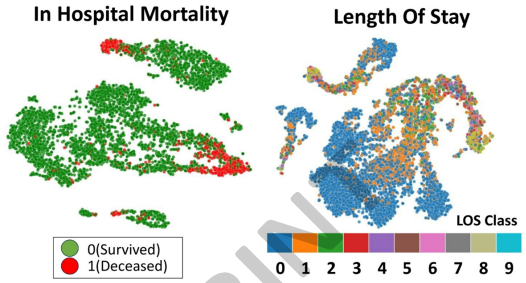


Figure 2: t-SNE visualization of MedFiTRG's learned multimodal patient representations.

7.5 Hyperparameter Sensitivity Analysis

We tune hyperparameters via grid search over learning rate {1e-2, 5e-3, **5e-5**}, batch size {32, **64**, 128}, dropout rate {0.1, **0.2**, 0.3}, transformers layers {1, **2**, 4, 6}, and hidden dimensions {64, 128, **256**}. To evaluate sensitivity, we fix the optimal settings and vary the batch size from 64 to 128. The resulting performance change is modest (0.2-0.5%), suggesting that MedFiTRG is not overly sensitive to hyperparameters. This reflects strong generalization capacity and reduced overfitting risk.

7.6 Scalability and Computational Complexity

We assess the scalability and computational efficiency of MedFiTRG, as summarized in Table 5. The model trains in approximately 770.94 seconds per epoch with 126M parameters, utilizing up to 10.7GB GPU memory. To maintain scalability, inter-patient edges in the social graph are computed within mini-batches rather than the entire patient population. During inference, it achieves an average latency of 1.15 seconds per batch (about 0.014 seconds per patient), demonstrating efficient scalability and suitability for real-time and large-scale clinical applications.

7.7 Explainability

To enhance explainability, MedFiTRG offers multi-level explanations across modalities. For medical imaging, we apply Grad-CAM to visualize salient regions in chest X-rays that most influence model predictions. For structured EHR data, Integrated Gradients are used to quantify feature importance and identify key clinical variables contributing to each outcome. Beyond modality-level attribution, the learned relational graph further improves interpretability: edges connect patients exhibiting similar patterns in EHR, text, or imaging embeddings, revealing clinically meaningful relationships across the population as shown in figure 3.

Model	IHM		REA		LOS		PHE	
	AUROC	AUPRC	AUROC	AUPRC	ma-F1	mi-F1	ma-AUROC	mi-AUROC
MedFuse [Hayat *et al.*, 2022]	0.8772 (0.003)	0.5158 (0.006)	0.7598 (0.002)	0.3618 (0.003)	0.1487 (0.006)	0.6289 (0.005)	0.8340 (0.001)	0.8785 (0.001)
MT [Ma *et al.*, 2022]	0.8726 (0.002)	0.5133 (0.008)	0.7585 (0.002)	0.3481 (0.008)	0.1531 (0.006)	0.6298 (0.003)	0.8362 (0.001)	0.8769 (0.001)
M3Care [Zhang *et al.*, 2022]	0.8732 (0.006)	0.5148 (0.017)	0.7618 (0.001)	0.3562 (0.003)	0.1549 (0.007)	0.6267 (0.004)	0.8429 (0.001)	0.8830 (0.001)
MMF [Lee *et al.*, 2023]	0.8804 (0.001)	0.5136 (0.010)	0.7627 (0.002)	0.3482 (0.006)	0.1554 (0.006)	0.6282 (0.004)	**0.8446 (0.001)**	**0.8845 (0.001)**
MultiModN [Swamy *et al.*, 2023]	0.8751 (0.002)	0.5055 (0.006)	0.7622 (0.001)	0.3526 (0.004)	0.1503 (0.010)	0.6307 (0.006)	0.8424 (0.000)	0.8826 (0.000)
FlexCare [Xu *et al.*, 2024b]	0.8823 (0.002)	0.5372 (0.006)	0.7680 (0.002)	**0.3702 (0.004)**	0.1479 (0.005)	**0.6358 (0.003)**	0.8393 (0.005)	0.8803 (0.004)
MoEhealth [Wang and Yang, 2025]	0.8165 (0.002)	0.3988 (0.006)	0.6377 (0.002)	0.1852 (0.004)	0.1227 (0.010)	0.4842 (0.003)	0.7924 (0.004)	0.8451 (0.006)
MedFiTRG (Ours)	**0.9392 (0.002)**	**0.7481 (0.005)**	**0.7682 (0.001)**	0.3506 (0.003)	**0.3104 (0.005)**	0.6351 (0.002)	0.8432 (0.003)	0.8701 (0.004)

Table 2: Performance comparison on multimodal clinical prediction tasks using MIMIC datasets. Best results are highlighted in bold. IHM: In-hospital mortality; REA: Readmission; LOS: Length of stay; PHE: Phenotyping.

Model	IHM		LOS	
	AUROC	AUPRC	macro-F1	micro-F1
MedFiTRG	0.9392	0.7481	0.3104	0.6351
MedFiTRG w/o social graph	0.9378	0.7235	0.3013	0.6316
MedFiTRG w/o temporal graph	0.9351	0.7287	0.3053	0.6189
MedFiTRG w/o global context vector	0.9303	0.7370	0.2937	0.6224
MedFiTRG w/o temporal and social graph	0.9313	0.7322	0.2827	0.6252

Table 3: Ablation study to show the impact of key architectural components of MedFiTRG model.

Model	IHM		LOS	
	AUROC	AUPRC	macro-F1	micro-F1
MGNN	0.9392	0.7481	0.3104	0.6351
GCN	0.7885	0.2851	0.2234	0.4195
GAT	0.7512	0.26	0.3051	0.6245

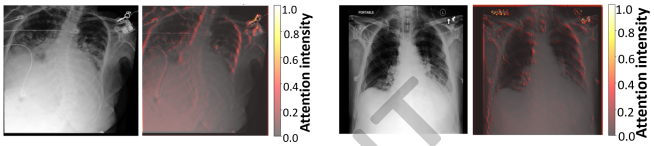
Table 4: Performance comparison of MGNN versus standard GCN and GAT on IHM and LOS prediction tasks.

Metric	Value
Training Time per Epoch	770.94 sec
Number of Parameters	126,392,111
Peak GPU Usage	10.70GB
Inference Latency per batch	1.1483 sec
Inference Latency per patient	0.01439 sec

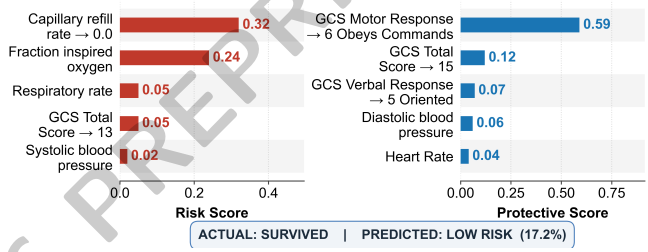
Table 5: Scalability and computation complexity metrics of model.

8 Conclusion and Future Work

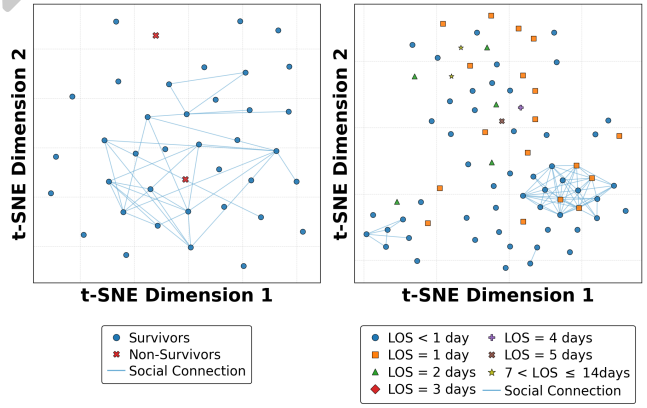
In this work, we introduced MedFiTRG, a unified framework for dynamic multimodal clinical prediction that integrates modality-aware representation learning and temporal–relational reasoning. By employing MGNN to learn adaptive, feature-wise modulated embeddings, MedFiTRG effectively mitigates modality sparsity and aligns heterogeneous feature spaces from EHR, imaging, and text. The subsequent temporal relational graph enables joint modeling of intra-patient temporal evolution and inter-patient relational dynamics, capturing multi-dimensional dependencies that conventional static or unimodal models fail to represent. Extensive experiments across three large-scale MIMIC datasets demonstrate that MedFiTRG achieves superior or comparable performance against strong multimodal baselines, achieving LOS Macro-F1 from 0.155 to 0.310, and IHM AUROC of 0.939 for mortality prediction (↑6.45%) and maintains competitive performance on readmission and phenotyping tasks. While MedFiTRG achieves superior or comparable performance against strong multimodal baselines, future work may extend toward population-level modeling using hierarchical or global graph representations, the integration of additional clinical modalities such as vital signs, lab time-series, and pathology, and prospective evaluations to assess real-world clinical utility.



(a) Grad-CAM visualization highlighting clinically relevant regions in chest X-rays.



(b) Integrated Gradients attribution showing the most influential EHR features for prediction.



(c) Patient similarity graph illustrating relational connections learned from multimodal embeddings for IHM(left) and LOS(right).

Figure 3: Multi-level explainability in MedFiTRG across imaging, EHR, and patient relations.

References

[Ahmed *et al.*, 2023] Imran Ahmed, Misbah Ahmad, Abdellah Chehri, and Gwanggil Jeon. A heterogeneous network embedded medicine recommendation system based on lstm. *Future Generation Computer Systems*, 149:1–11, 2023.

[Al-Zoghby *et al.*, 2025] Aya M Al-Zoghby, Ahmed Ismail Ebada, Aya S Saleh, Mohammed Abdelhay, and Wael A Awad. A comprehensive review of multimodal deep learning for enhanced medical diagnostics. *Computers, Materials & Continua*, 84(3), 2025.

[Bai *et al.*, 2018] Tian Bai, Shanshan Zhang, Brian L. Egleston, and Slobodan Vucetic. Interpretable representation learning for healthcare via capturing disease progression through time. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, page 43–51, New York, NY, USA, 2018. Association for Computing Machinery.

[Brockschmidt, 2020] Marc Brockschmidt. Gnn-film: Graph neural networks with feature-wise linear modulation. In *International Conference on Machine Learning (ICML)*, 2020.

[Chen *et al.*, 2024a] Jiayuan Chen, Yu Wang, Ruining Deng, Quan Liu, Can Cui, Tianyuan Yao, Yilin Liu, Jianyong Zhong, Agnes B Fogo, Haichun Yang, et al. Spatial pathomics toolkit for quantitative analysis of podocyte nuclei with histology and spatial transcriptomics data in renal pathology. In *Proceedings of SPIE–the International Society for Optical Engineering*, volume 12933, page 1293310, 2024.

[Chen *et al.*, 2024b] Jiayuan Chen, Changchang Yin, Yuanlong Wang, and Ping Zhang. Predictive modeling with temporal graphical representation on electronic health records, 2024.

[Choi *et al.*, 2016] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. *Advances in neural information processing systems*, 29, 2016.

[Gao *et al.*, 2020] Junyi Gao, Cao Xiao, Yasha Wang, Wen Tang, Lucas M Glass, and Jimeng Sun. Stagenet: Stage-aware neural networks for health risk prediction. In *Proceedings of The Web Conference 2020*, pages 530–540, 2020.

[Gupta *et al.*, 2025] Shivani Gupta, Saurabh Sharma, Rajesh Sharma, and Joydeep Chandra. Healing with hierarchy: Hierarchical attention empowered graph neural networks for predictive analysis in medical data. *Artificial Intelligence in Medicine*, 165:103134, 2025.

[Hayat *et al.*, 2022] Nasir Hayat, Krzysztof J Geras, and Farah E Shamout. Medfuse: Multi-modal fusion with clinical time-series data and chest x-ray images. In *Machine Learning for Healthcare Conference*, pages 479–503. PMLR, 2022.

[Hemker *et al.*, 2024] Konstantin Hemker, Nikola Simidjievski, and Mateja Jamnik. Healnet: Multimodal fusion for heterogeneous biomedical data. *Advances in Neural Information Processing Systems*, 37:64479–64498, 2024.

[Jandoubi and Akhloufi, 2025] Bassem Jandoubi and Moulay A Akhloufi. Multimodal artificial intelligence in medical diagnostics. *Information*, 16(7):591, 2025.

[Johnson *et al.*, 2019] Alistair Johnson, Matt Lungren, Yifan Peng, Zhiyong Lu, Roger Mark, Seth Berkowitz, and Steven Horng. Mimic-cxr-jpg-chest radiographs with structured labels. *PhysioNet*, 101:215–220, 2019.

[Johnson *et al.*, 2023] Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

[Lee *et al.*, 2023] Kwanhyung Lee, Soojeong Lee, Sangchul Hahn, Heejung Hyun, Edward Choi, Byungeun Ahn, and Joohyung Lee. Learning missing modal electronic health records with unified multi-modal data embedding and modality-aware attention. In *Machine Learning for Healthcare Conference*, pages 423–442. PMLR, 2023.

[Luo *et al.*, 2020] Junyu Luo, Muchao Ye, Cao Xiao, and Fenglong 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.

[Ma *et al.*, 2022] Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18177–18186, 2022.

[Rajkomar *et al.*, 2018] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew Dai, Nissan Hajaj, Peter Liu, Xiaobing Liu, Mimi Sun, Patrik Sundberg, Hector Yee, Kun Zhang, Gavin Duggan, Gerardo Flores, Michaela Hardt, Jamie Irvine, Quoc Le, Kurt Litsch, Jake Marcus, Alexander Mossin, and Jeff Dean. Scalable and accurate deep learning for electronic health records. *npj Digital Medicine*, 1, 01 2018.

[Ruan *et al.*, 2025] Yucheng Ruan, Daniel J Tan, See-Kiong Ng, Ling Huang, and Mengling Feng. Towards accurate and reliable icu outcome prediction: a multimodal learning framework based on belief function theory using structured ehrs and free-text notes. *Journal of Healthcare Informatics Research*, pages 1–42, 2025.

[Siam *et al.*, 2025] Md Kamrul Siam, Md Jobair Hossain Faruk, Bofan He, Jerry Q Cheng, and Huanying Gu. Multimodal models in healthcare: Methods, challenges, and future directions for enhanced clinical decision support. *Information*, 16(11):971, 2025.

[Soenksen *et al.*, 2022] Luis R Soenksen, Yu Ma, Cynthia Zeng, Leonard Boussioux, Kimberly Villalobos Carballo, Liangyuan Na, Holly M Wiberg, Michael L Li, Ignacio

Fuentes, and Dimitris Bertsimas. Integrated multimodal artificial intelligence framework for healthcare applications. *NPJ digital medicine*, 5(1):149, 2022.

[Sohn *et al.*, 2024] Hyunwoo Sohn, Kyungjin Park, Baekkwan Park, and Min Chi. Multi-ta: multilevel temporal augmentation for robust septic shock early prediction. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 6035–6043, 2024.

[Swamy *et al.*, 2023] Vinitra Swamy, Malika Satayeva, Jibril Frej, Thierry Bossy, Thijs Vogels, Martin Jaggi, Tanja Käser, and Mary-Anne Hartley. Multimodn—multimodal, multi-task, interpretable modular networks. *Advances in neural information processing systems*, 36:28115–28138, 2023.

[Tang *et al.*, 2023] Siyi Tang, Amara Tariq, Jared A Dunnmon, Umesh Sharma, Praneetha Elugunti, Daniel L Rubin, Bhavik N Patel, and Imon Banerjee. Predicting 30-day all-cause hospital readmission using multimodal spatiotemporal graph neural networks. *IEEE Journal of Biomedical and Health Informatics*, 27(4):2071–2082, 2023.

[Teles *et al.*, 2025] Ariel Soares Teles, Ivan Rodrigues de Moura, Francisco Silva, Angus Roberts, and Daniel Stahl. Ehr-based prediction modelling meets multimodal deep learning: A systematic review of structured and textual data fusion methods. *Information Fusion*, page 102981, 2025.

[Wang and Yang, 2025] Xiaoyang Wang and Christopher Yang. Moe-health: A mixture of experts framework for robust multimodal healthcare prediction. In *Proceedings of the 16th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 1–9, 2025.

[Wang *et al.*, 2024] Yuanlong Wang, Changchang Yin, and Ping Zhang. Multimodal risk prediction with physiological signals, medical images and clinical notes. *Heliyon*, 10(5), 2024.

[Xu *et al.*, 2023] Yongxin Xu, Kai Yang, Chaohe Zhang, Peinie Zou, Zhiyuan Wang, Hongxin Ding, Junfeng Zhao, Yasha Wang, and Bing Xie. Vecocare: Visit sequences-clinical notes joint learning for diagnosis prediction in healthcare data. In *IJCAI*, volume 23, pages 4921–4929, 2023.

[Xu *et al.*, 2024a] Jingwen Xu, Ye Zhu, Fei Lyu, Grace Lai-Hung Wong, and Pong C Yuen. Temporal neighboring multi-modal transformer with missingness-aware prompt for hepatocellular carcinoma prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 79–88. Springer, 2024.

[Xu *et al.*, 2024b] Muhao Xu, Zhenfeng Zhu, Youru Li, Shuai Zheng, Yawei Zhao, Kunlun He, and Yao Zhao. Flexcare: Leveraging cross-task synergy for flexible multimodal healthcare prediction. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3610–3620, 2024.

[Yao *et al.*, 2024] Wenfang Yao, Kejing Yin, William K Cheung, Jia Liu, and Jing Qin. Drfuse: Learning disentangled representation for clinical multi-modal fusion with missing modality and modal inconsistency. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 16416–16424, 2024.

[Zhang *et al.*, 2022] Chaohe Zhang, Xu Chu, Liantao Ma, Yinghao Zhu, Yasha Wang, Jiangtao Wang, and Junfeng Zhao. M3care: Learning with missing modalities in multimodal healthcare data. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pages 2418–2428, 2022.

[Zhang *et al.*, 2023] Yingying Zhang, Xian Wu, Quan Fang, Shengsheng Qian, and Changsheng Xu. Knowledge-enhanced attributed multi-task learning for medicine recommendation. 41(1), January 2023.

[Zhong *et al.*, 2024] Yuan Zhong, Xiaochen Wang, Jiaqi Wang, Xiaokun Zhang, Yaqing Wang, Mengdi Huai, Cao Xiao, and Fenglong Ma. Synthesizing multimodal electronic health records via predictive diffusion models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4607–4618, 2024.

[Zhu *et al.*, 2024] Fanglin Zhu, Lizhen Cui, Yonghui Xu, Zhe Qu, and Zhiqi Shen. A survey of personalized medicine recommendation. *International Journal of Crowd Science*, 8(2):77–82, 2024.