Multimodal Disease Progression Modeling via Spatio-temporal Disentanglement and Multiscale Alignment

Chen Liu[1], **Wenfang Yao**[1], **Kejing Yin**[2,*], **William K. Cheung**[2], **Jing Qin**[1]
[1]School of Nursing, The Hong Kong Polytechnic University
[2]Department of Computer Science, Hong Kong Baptist University

Abstract

Longitudinal multimodal data, including electronic health records (EHR) and sequential chest X-rays (CXRs), is critical for modeling disease progression, yet remains underutilized due to two key challenges: (1) redundancy in consecutive CXR sequences, where static anatomical regions dominate over clinically-meaningful dynamics, and (2) temporal misalignment between sparse, irregular imaging and continuous EHR data. We introduce `DiPro`, a novel framework that addresses these challenges through region-aware disentanglement and multi-timescale alignment. First, we disentangle static (anatomy) and dynamic (pathology progression) features in sequential CXRs, prioritizing disease-relevant changes. Second, we hierarchically align these static and dynamic CXR features with asynchronous EHR data via local (pairwise interval-level) and global (full-sequence) synchronization to model coherent progression pathways. Extensive experiments on the MIMIC dataset demonstrate that `DiPro` could effectively extract temporal clinical dynamics and achieve state-of-the-art performance on both disease progression identification and general ICU prediction tasks.

1 Introduction

Accurately modeling disease progression, i.e., the temporal evolution of a disease, is critical for personalized clinical care decision-making [1–3]. By capturing progression dynamics, predictive models can enable early identification of deterioration, improve precise risk stratification, and inform individualized treatment planning [2, 4–6]. In ICU settings, for instance, tracking sepsis progression through multimodal data (e.g., vitals, labs, and organ functions) is essential for identifying early deterioration and initiating lifesaving treatments [7, 8].

Modern healthcare increasingly relies on longitudinal clinical data to track disease progression. Sequential chest X-rays (CXRs) capture valuable visual evidence of anatomical and pathological changes over time, while electronic health records (EHRs) provide continuous physiological metrics and treatment responses [9, 10]. The complementary nature of these modalities offers a unique opportunity: a fusion of imaging and clinical time-series data could enable a more holistic modeling of disease trajectories [11, 12]. Despite a growing number of studies that explored disease progression modeling and multimodal fusion using longitudinal clinical data, two key challenges remain insufficiently addressed:

Redundancy in clinical image sequences. Static anatomical features (e.g., chronic cardiac enlargement or stable skeletal deformities) dominate sequential CXR scans, often obscuring subtle but clinically important pathological changes (e.g., new infiltrates or evolving edema). Without explicit disentanglement, the signals of disease progression become diluted, reducing the utility

*Correspondence to: Kejing Yin <cskjyin@comp.hkbu.edu.hk>

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

of sequential imaging analysis. Recent CXR-based progression models (e.g., [13, 14]) primarily focus on extracting symptom-related features and anatomical-disease co-occurrences, treating all imaging features uniformly. However, distinguishing long-term anatomical baselines from evolving pathological changes is the key to improving the precision of disease progression modeling, yet is largely overlooked [15, 16].

Temporal misalignments across modalities. While EHRs provide continuous, high-frequency measurements (e.g., hourly vitals or lab tests), CXRs offer only sparse, irregular snapshots, creating an intrinsic misalignment in timescales. This discrepancy in temporal granularity complicates the alignment of cross-modal trends and may obscure short-term clinical deteriorations between imaging intervals. Existing multimodal fusion approaches prioritized the latest CXR for simplicity [17, 18], neglecting the temporal CXR context, while longitudinal fusion methods [19, 20] rely on rigid imputation or fixed time embeddings, lacking adaptive mechanisms to synchronize fine-grained cross-modal patterns. Thus, strategically integrating these progression dynamics to model cross-modal clinical trajectories remains underexplored.

To address the aforementioned challenges, we propose *Disease Progression-Aware Clinical Prediction*, `DiPro`[2], a novel framework that disentangles time-variant and time-invariant information from longitudinal CXRs and integrates these representations with EHR data across multiple temporal granularities. Specifically, our framework contains three modules: **(1) Spatiotemporal Disentanglement**: We disentangle dynamic pathological changes from static anatomical structures in different spatial regions across consecutive CXRs. This separation helps reduce redundancy and prevents interference between features with different clinical semantics, allowing the model to focus on meaningful progression signals for more effective temporal representation learning. **(2) Progression-Aware Enhancement**: To improve the model's sensitivity to progression direction, we reverse the order of CXR pairs and train the model to produce reversed dynamic features while keeping static features consistent. This strategy further separates the two types of features and emphasizes their distinct clinical semantics. **(3) Multimodal Fusion via Multiscale Alignment**: To fully synergize the multimodal features, we introduce a multi-grained fusion module that achieves alignment in multiple scales: the local fusion aligns consecutive CXR pairs with EHR data, while the global fusion aligns the semantics of the entire CXR and EHR sequences. This bridges fine-grained EHR-CXR interactions with global disease progression patterns.

Our contributions are summarized as follows:

- We present a framework to disentangle dynamic and static information from longitudinal CXRs, with dedicated architectural constraints to capture representations in line with disease progressions.
- We propose a multiscale multimodal fusion approach that facilitates multi-grained interactions between temporally misaligned CXR dynamics and EHR time-series data.
- Extensive experiments demonstrate that `DiPro` achieves state-of-the-art performance on both disease progression modeling and ICU-related prediction tasks. Quantitative evaluation shows that the model aligns well with existing clinical knowledge.

2 The `DiPro` Approach

To address the challenges of redundancy in longitudinal CXRs and temporal misalignment with EHR data, we propose `DiPro`, to systematically disentangle static and dynamic features from CXRs and align multimodal data across hierarchical timescales. The approach is motivated by two key observations: (1) disease progression in CXR sequences unfolds via region-localized pathological changes [21, 22], and (2) EHR and imaging data exhibit complementary dynamics at different temporal granularities. `DiPro` integrates three cohesive modules: (1) Spatiotemporal Disentanglement (STD) to isolate pathology-sensitive features from sequential CXRs, (2) Progression-Aware Enhancement (PAE) to enforce temporal consistency in learned dynamics, and (3) Multiscale Multimodal Fusion (MMF) to bridge local EHR-CXR interactions with global progression trends. Figure 1 illustrate the `DiPro` framework. We next detail each module.

[2]The code is available at `https://github.com/Chenliu-svg/DiPro`.

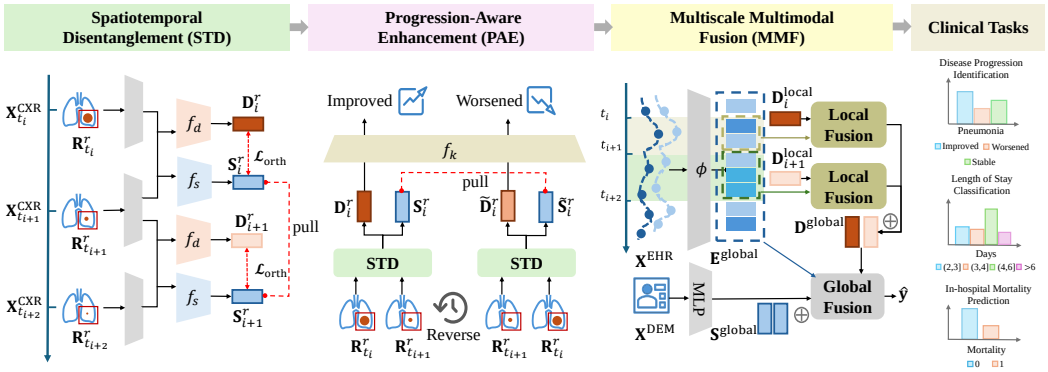


Figure 1: Overview of the DiPro framework. The model comprises three main modules: (1) Spatiotemporal Disentanglement (STD) separates dynamic pathological features ($\mathbf{D}_i^r$) from static anatomical structures ($\mathbf{S}_i^r$) in region-level chest X-rays across time; (2) Progression-Aware Enhancement (PAE) strengthens the model's understanding of progression direction by reversing CXR pair order and enforcing the reversed dynamic features $\widetilde{\mathbf{D}}_i^r$ to predict the reversed progression, while maintaining consistency in static features; (3) Multiscale Multimodal Fusion (MMF) integrates CXR features with temporally misaligned EHR data via local (interval-level) and global (sequence-level) fusion, enabling accurate predictions across multiple clinical tasks, including disease progression identification, length-of-stay classification, and in-hospital mortality prediction.

2.1 Notations and Preliminaries

For each patient, let $\mathbf{X}^{\text{CXR}} = \{\mathbf{X}_{t_i}^{\text{CXR}}\}_{i=1}^{T}$ denote a set of T CXR images during an ICU stay, where each $\mathbf{X}_{t_i}^{\text{CXR}} \in \mathbb{R}^{H\times W\times C}$ is the CXR image taken at time t_i. Each CXR image contains R anatomical regions $\{\mathbf{R}_{t_i}^r\}_{r=1}^{R}$. Each consecutive image pair $(\mathbf{X}_{t_i}^{\text{CXR}}, \mathbf{X}_{t_{i+1}}^{\text{CXR}})$ is associated with a label set $\mathbf{Y}_i^{\text{CXR}} = \{y_i^{r,k}\}_{r,k=1}^{R,K}$ where $y_i^{r,k} \in \{-1, 0, 1\}$ indicates whether the progression status of disease k in the r-th region has worsened, remained stable (no change), or improved. The whole EHR time serie recorded with M timestamps is $\mathbf{X}^{\text{EHR}} = [\mathbf{x}_{t_0}, \mathbf{x}_{t_1}, \ldots, \mathbf{x}_{t_M}]$ with $\mathbf{x}_t \in \mathbb{R}^N$ being the N-dim variables recorded at time t. We use $\mathbf{X}^{\text{EHR}}$ to denote all available EHR time series. Patient's demographic attributes are denoted as $\mathbf{X}^{\text{DEM}} \in \mathbb{R}^P$ where P is the number of attributes. Our objective is to learn a mapping $f_\theta : (\mathbf{X}^{\text{CXR}}, \mathbf{X}^{\text{EHR}}, \mathbf{X}^{\text{DEM}}) \to \mathbf{y}$, that maps the integrated CXR, EHR time series, and demographic information to the final clinical outcome prediction $\mathbf{y}$.

2.2 Spatiotemporal Disentanglement (STD)

We propose a novel method to disentangle region-based time-variant (dynamic) and time-invariant (static) information from consecutive CXR image pairs. The disentanglement is motivated by their distinct clinical roles: dynamic features reflect disease progression, while static features capture patient-specific anatomical structures. Inspired by prior works demonstrating the effectiveness of feature disentanglement in representation learning [23, 24], our method targets more efficient and structured latent representations.

Feature extraction. For each anatomical region $\mathbf{R}_{t_i}^r$, we first use a shared pretrained ResNet-50 [25] to decode it and obtain the region feature $\mathbf{F}_{t_i}^r \in \mathbb{R}^d$. Then we concatenate the two consecutive feature vectors and then pass them through two separate projection heads, f_s and f_d for static and dynamic information extraction, respectively,:

$$\mathbf{S}_i^r = f_s([\mathbf{F}_{t_i}^r || \mathbf{F}_{t_{i+1}}^r]), \quad \mathbf{D}_i^r = f_d([\mathbf{F}_{t_i}^r || \mathbf{F}_{t_{i+1}}^r]), \tag{1}$$

where $||$ denotes channel-wise concatenation. Here, $\mathbf{S}_i^r$ and $\mathbf{D}_i^r$ represent the static and dynamic features, respectively, for region r at time step pair (t_i, t_{i+1}).

Orthogonal disentanglement loss. To encourage effective disentanglement between static and dynamic representations, we apply an orthogonal constraint [26]. Specifically, we minimize the

squared cosine similarity between static and dynamic features within each region and time pair:

$$\mathcal{L}_{\text{orth}} = \frac{1}{(T-1)R} \sum_{i=1}^{T-1} \sum_{r=1}^{R} \left(\text{sim}\left(\mathbf{S}_i^r, \mathbf{D}_i^r\right)\right)^2, \tag{2}$$

where $\text{sim}(\cdot,\cdot)$ denotes the cosine similarity. This term ensures that the dynamic and static features are as unrelated as possible in the latent space, thereby promoting better disentanglement between time-invariant and time-varying patterns.

Temporal consistency for static features. Anatomical structures are expected to remain stable over time in sequential CXR images [27, 28]. To enforce temporal consistency of static features, we introduce a mean squared error (MSE) loss that encourages static features from consecutive time pairs to remain close:

$$\mathcal{L}_{\text{temp}} = \frac{1}{N} \sum_{r=1}^{R} \sum_{i=1}^{T-2} \left\|\mathbf{S}_i^r - \mathbf{S}_{i+1}^r\right\|_2^2, \tag{3}$$

where $N = (T-2) \times R$ is the number of consecutive static feature pairs considered. This constraint ensures that the learned static features remain stable and coherent over time.

2.3 Progression-Aware Enhancement (PAE)

Intuitively, reversing the order of a CXR pair should invert the progression direction while preserving static anatomical information. Hence, dynamic features should reflect this reversal, providing a more robust and interpretable representation of temporal change. To make dynamic features more sensitive to the direction of disease progression, we introduce a progression-aware enhancement.

By reversing the input order of the region feature pair $(\mathbf{F}_{t_i}^r, \mathbf{F}_{t_{i+1}}^r)$, we obtain the dynamic and static features as:

$$\widetilde{\mathbf{D}}_i^r = f_d([\mathbf{F}_{t_{i+1}}^r || \mathbf{F}_{t_i}^r]),\ \widetilde{\mathbf{S}}_i^r = f_s([\mathbf{F}_{t_{i+1}}^r || \mathbf{F}_{t_i}^r]). \tag{4}$$

We then feed both the original and reversed dynamic features into K disease-specific progression classification heads $\{f_k\}_{k=1}^K$, where each head f_k corresponds to a disease and predicts its progression status $y_i^{r,k}$ for the given region r. Let $\widehat{y}_i^{r,k}$ be the predicted label from the original direction, and $\widetilde{y}_i^{r,k}$ be the prediction from the reversed input, i.e., $\widehat{y}_i^{r,k} = f_k(\mathbf{D}_i^r),\ \widetilde{y}_i^{r,k} = f_k(\widetilde{\mathbf{D}}_i^r)$. We expect that reversing the input order should generate a contrary prediction. Thus, we convert the label into $-y_i^{r,k}$ to indicate a reversed progression direction as the ground truth label.

Training objective. For dynamic features, we supervise predictions using cross-entropy (CE) loss for both original and reversed progression prediction; For static features, we leverage MSE to encourage their consistency over the reversal.

$$\mathcal{L}_{\text{PAE}} = \sum_{r=1}^{R} \sum_{k=1}^{K} \left[\text{CE}(\widehat{y}_i^{r,k}, y_i^{r,k}) + \text{CE}(\widetilde{y}_i^{r,k}, -y_i^{r,k})\right] + \lambda_{\text{static}} \sum_{r=1}^{R} \left\|\mathbf{S}_i^r - \widetilde{\mathbf{S}}_i^r\right\|_2^2, \tag{5}$$

where λ_{static} is a hyperparameter. This objective encourages the model to encode progression-aware dynamic information and time-invariant anatomical information, improving the reliability of progression modeling across time.

2.4 Multiscale Multimodal Fusion (MMF)

To effectively integrate temporally disaligned patient data, we propose a multiscale fusion framework that combines visual and temporal features from longitudinal chest X-rays (CXRs) and electronic health records (EHRs). EHR signals are composed of dynamic time-series measurements and static demographics, while CXRs encode both time-varying imaging biomarkers and static anatomical traits. Our fusion proceeds in three stages: (1) local CXR-EHR fusion within each interval, (2) global hierarchical fusion, and (3) static feature integration and prediction.

Local CXR-EHR fusion within temporal intervals. We first encode the whole EHR time series $\mathbf{X}^{\text{EHR}}$ using a global Transformer encoder ϕ [29] to get the global representation of EHR: $\mathbf{E}^{\text{global}} = \phi(\mathbf{X}^{\text{EHR}})$. Given a time interval $[t_i, t_{i+1}]$ from a consecutive image pair $(\mathbf{X}^{\text{CXR}}_{t_i}, \mathbf{X}^{\text{CXR}}_{t_{i+1}})$, to obtain EHR representations focused on the interval, we first define the time embedding for each EHR timestamp t_j relative to the CXR time interval as: $\mathrm{T}_{t_j} = f_{\text{TE}}\left([t_j - t_i,\ t_{i+1} - t_j,\ \sigma((t_j - t_i)(t_{i+1} - t_j))]\right)$, where $\sigma(\cdot)$ is the sigmoid function to approximate the indicator of whether $t_j \in [t_i, t_{i+1}]$. These time embeddings are then stacked as $\mathbf{T}_i = [\mathrm{TE}_{t_0}, \ldots, \mathrm{TE}_{t_M}]$. we then apply cross-attention [29] between the time embeddings and global EHR features with learnable parameters $\mathbf{W}_Q, \mathbf{W}_K$ and $\mathbf{W}_V$ using a center-focused attention mask:

$$\mathbf{E}^{\text{local}}_i = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}} + \text{AttnMask}\right)\cdot \mathbf{V},\ \mathbf{Q} = \mathbf{W}_Q\mathbf{T}_i,\ \mathbf{K} = \mathbf{W}_K\mathbf{E}^{\text{global}},\ \mathbf{V} = \mathbf{W}_V\mathbf{E}^{\text{global}}. \tag{6}$$

The attention mask is defined as:

$$\text{AttnMask}_{ij} = \begin{cases} -\left|t_j - \frac{t_i + t_{i+1}}{2}\right|, & \text{if } t_j \in [t_i, t_{i+1}], \\ -\infty, & \text{otherwise.} \end{cases} \tag{7}$$

We then fuse dynamic CXR features $\mathbf{D}^{\text{local}}_i = \{\mathbf{D}^r_i\}^R_{r=1}$ and local EHR representations $\mathbf{E}^{\text{local}}_i$ via a cross-attention layer. The keys and values are constructed by concatenating modality-specific projections of both inputs.

$$\mathbf{D}^{\text{fuse}}_i = \text{LayerNorm}(\text{CrossAttn}(\mathbf{D}^{\text{local}}_i, [\mathbf{E}^{\text{local}}_i || \mathbf{D}^{\text{local}}_i]). \tag{8}$$

Global hierarchical fusion We collect all locally fused features $\mathbf{D}^{\text{global}} = \{\mathbf{D}^{\text{fuse}}_i\}^{T-1}_{i=1}$ across CXR intervals. We refine the global EHR representation by attending over $\mathbf{D}^{\text{global}}$:

$$\mathbf{H}^{\text{global}} = \text{LayerNorm}(\text{CrossAttn}(\mathbf{E}^{\text{global}}, \mathbf{D}^{\text{global}})). \tag{9}$$

We include an additional self-attention layer [29] to further enhance global interactions between the two modalities. The resulting enriched sequence then serves as the query in the final cross-attention mechanism with the static features for prediction.

Final static fusion and prediction. We first embed demographic information $\mathbf{X}^{\text{DEM}}$ via an MLP, then concatenate it with static CXR features $\mathbf{S}^{\text{global}} = \{\mathbf{S}_i\}^T_{i=1}$:

$$\mathbf{H}^{\text{static}} = [\mathbf{S}^{\text{global}} \,||\, \text{MLP}(\mathbf{X}^{\text{DEM}})]. \tag{10}$$

A fusion module Φ, composed of a cross-attention layer followed by prediction heads, integrates dynamic and static features to generate predictions:

$$\widehat{\mathbf{y}} = \Phi([\mathbf{D}^{\text{global}} \,||\, \mathbf{H}^{\text{global}}], \mathbf{H}^{\text{static}}). \tag{11}$$

The final training loss combines cross-entropy with all auxiliary objectives:

$$\mathcal{L} = \lambda_{\text{pred}} \cdot \text{CE}(\widehat{\mathbf{y}}, \mathbf{y}) + \lambda_{\text{orth}}\mathcal{L}_{\text{orth}} + \lambda_{\text{temp}}\mathcal{L}_{\text{temp}} + \lambda_{\text{PAE}}\mathcal{L}_{\text{PAE}}, \tag{12}$$

where λ_{pred}, λ_{orth}, λ_{temp}, and λ_{PAE} are hyperparameters. Details of the model architecture, hyperparameters, and training procedure are provided in Appendix A.2.1.

3 Experiments

3.1 Experiment Setting

Datasets. We evaluated `DiPro` on the large-scale, public dataset, MIMIC [30], which contains de-identified health data of intensive care unit (ICU) admissions. Our study leveraged three derived datasets from the MIMIC ecosystem: (1) **MIMIC-IV** [30] provides electronic health records (EHR) including demographic information and time-series physiological measurements per ICU stay; (2) **MIMIC-CXR** [31] contains sequential chest radiographs during ICU hospitalizations; and (3) **Chest ImaGenome** [32] augments imaging with fine-grained annotations: bounding boxes for anatomical regions and localized change labels (improved, worsened, or no change) between consecutive CXRs. To facilitate longitudinal analysis, we selected ICU stays with ≥ 2 CXRs to track disease progression. From MIMIC-IV, we extracted EHR data consisting of 7 demographic variables and 38 physiological time-series variables, including vital signs and laboratory results.

Clinical tasks and evaluation metrics. We evaluate the performance of DiPro on two types of clinical tasks to demonstrate the advantages of our multimodal framework:

(1)*Disease progression identification.* Given two consecutive CXRs, the task is to predict the disease progression status (improved, worsened, or no change) for seven common thoracic conditions: atelectasis, enlarged cardiac silhouette, consolidation, pulmonary edema, lung opacity, pleural effusion, and pneumonia. Following Karwande et al. [14], we derive the progression label of each disease for a CXR pair from the progression labels of annotated regions in the Chest ImaGenome dataset. Notably, our work is the first to integrate EHR data for this task. Specifically, we extract the EHR data recorded within the time interval of two CXRs and integrate them for a richer context for progression. We report macro-precision, macro-recall, macro-F1 score, AUPRC and AUROC following prior works [13, 14, 33].

(2)*General ICU prediction.* We consider two clinically vital tasks: In-hospital mortality prediction and ICU length of stay prediction. Both tasks focus on forecasting patient outcomes leveraging multimodal data: EHR time series and sequential CXRs collected during the first 48 hours after ICU admission. The *In-hospital mortality prediction* task is a binary classification problem: it predicts patient mortality prior to hospital discharge. We evaluate performance using AUROC and AUPRC, following [19, 17]. *Length of stay prediction* task aims to estimate patient ICU stay duration. We frame this as a multi-class classification problem by discretizing stay duration into four intervals: [2, 3), [3, 4), [4, 6), and ≥ 6 days. The counting starts from 2 days as we are using ICU stays longer than 48 hours. Following [11, 34, 35], model performance is evaluated using Cohen's kappa and accuracy.

All experiments are conducted with three random seeds, with results reported as the mean $\pm$ standard deviation across independent runs. Details on cohort selection, label prevalence, data statistics and data processing procedures for each task are provided in Appendix A.1.

Baselines. We compare DiPro with the following three types of baselines:

(1) **Sequential CXR disease progression specialists (unimodal)**: *CheXRelNet* [14] combines local and global visual features with anatomical dependencies to model longitudinal disease changes. *CheXRelFormer* [33] adopts hierarchical Siamese Transformer to capture multi-level feature discrepancies across CXR images. *SDPL* [13] learns symptom-aware embeddings to extract and compare condition-specific features from two radiographs.

(2) **Longitudinal multimodal specialists**: *UTDE* [19] models asynchronous longitudinal data via a gated attention-based imputation framework. *UMSE* [20] uses triplet-structured set embeddings and a modality-aware attention mechanism to handle missing data and fuse multiple modalities.

(3) **Clinical multimodal fusion specialists**: *MedFuse* [17] introduces an LSTM-based module for both uni-modal and multimodal input. *DrFuse* [18] disentangles modality-shared and modality-specific features, and utilizes disease-wise attention for effective fusion. Both models are designed to take the last available CXR for modality input, we extend them to the setting of multiple CXRs with minimal architectural modification. Details of all the baseline models are provided in Appendix A.2.2.

3.2 Prediction Performance

DiPro **excels in modeling disease progression in sequential CXRs.** Table 1 reports the mean performance across seven disease progression identification tasks (per-disease results in Table 11). Compared to CXR-based progression models using only unimodal sequential CXRs, DiPro achieves relative improvements of 15.3% in F1 and 12.2% in AUPRC over the state-of-the-art SDPL [13]. This suggests that explicitly disentangling disease dynamics from static anatomical structures across CXR pairs reduces redundancy and enables more effective progression modeling. Compared to CheXRelNet [14], which models region-disease co-occurrence via graphs, DiPro enhances the disentangled regional progression dynamics using a tailored PAE module, offering a more targeted mechanism to capture disease-region progression. A broader baseline comparison with large vision–language models is presented in Table 17.

DiPro **excels in longitudinal multimodal fusion.** As shown in the *Multimodal Methods* block of Table 1, adding EHR to unimodal DiPro improves performance (relative increase of 2.9% in F1 and 2.1% in AUPRC). This confirms DiPro's ability to effectively leverage complementary EHR features for disease progression prediction. Furthermore, DiPro outperforms all baselines, whether

Table 1: **Performance Comparison on Disease Progression Identification Tasks.** This table reports the macro-average performance (± standard deviation) of various unimodal and multimodal methods across seven disease progression identification tasks. DiPro achieves the best results in both unimodal and multimodal settings, indicating its effectiveness in modeling disease progression from sequential CXRs and its strength in longitudinal multimodal fusion. Detailed results are provided in Table 10. Per-disease results are provided in Table 11. (Numbers in bold indicate the best performance in each column, and those underlined represent the best-performing baseline.)

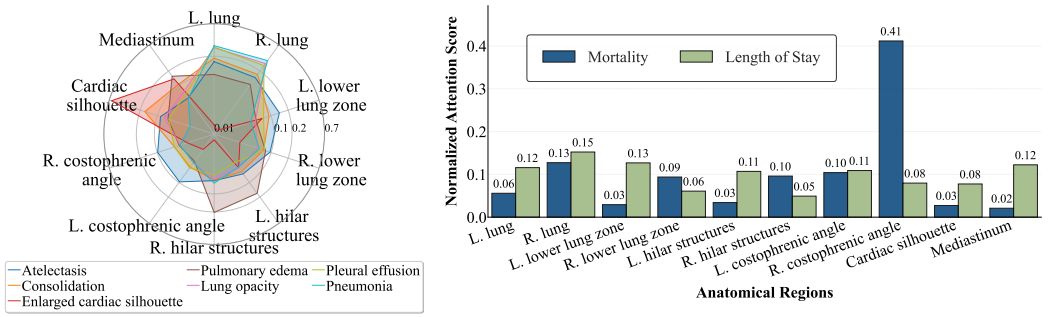
Method	Precision	Recall	F1	AUPRC	AUROC
Unimodal Methods (CXR)					
CheXRelNet [14]	0.395±0.015	0.392±0.010	0.389±0.010	0.394±0.010	0.574±0.011
CheXRelFormer [33]	0.389±0.044	0.379±0.033	0.354±0.032	0.372±0.023	0.551±0.041
SDPL [13]	0.408±0.006	0.406±0.020	0.393±0.010	0.417±0.032	0.609±0.031
DiPro (ours)	**0.475±0.004**	**0.452±0.011**	**0.453±0.009**	**0.468±0.013**	**0.651±0.016**
Multimodal Methods					
UTDE [19]	0.481±0.017	0.462±0.002	0.449±0.005	0.472±0.014	0.659±0.011
UMSE [20]	0.353±0.011	0.361±0.009	0.352±0.013	0.364±0.006	0.544±0.004
MedFuse [17]	0.423±0.049	0.413±0.045	0.409±0.042	0.422±0.040	0.530±0.030
DrFuse [18]	0.442±0.009	0.461±0.007	0.429±0.010	0.438±0.003	0.628±0.002
DiPro (ours)	**0.484±0.008**	**0.471±0.024**	**0.466±0.018**	**0.478±0.018**	**0.664±0.013**

Table 2: **Performance Comparison on General ICU Prediction Tasks.** This table presents the average performance (± standard deviation) on two general ICU prediction tasks: mortality (AUPRC, AUROC) and length of stay (Kappa, ACC). Results are reported for both "Last" and "Long." CXR settings. The "Long." setting incorporates both longitudinal CXR and EHR data, whereas the "Last" setting uses only the most recent CXR together with EHR data. DiPro demonstrates superior performance across both settings, highlighting its effectiveness in longitudinal multimodal fusion. Note that in the "no-CXR" setting (second-to-last row), DiPro effectively reduces to a Transformer-based [29] EHR encoder. More results are provided in Table 12.

	CXR Used		Mortality		Length of Stay	
Method	Last	Long.	AUPRC	AUROC	Kappa	ACC
UTDE [19]	✓		0.717±0.019	0.887±0.004	0.160±0.016	0.381±0.013
		✓	0.710±0.019	0.887±0.012	0.195±0.031	0.400±0.021
UMSE [20]	✓		0.722±0.039	0.896±0.012	0.217±0.013	0.419±0.010
		✓	0.712±0.028	0.891±0.011	0.204±0.019	0.410±0.013
MedFuse [17]	✓		0.686±0.018	0.869±0.011	0.213±0.012	0.413±0.004
		✓	0.716±0.018	0.881±0.005	0.210±0.039	0.412±0.027
DrFuse [18]	✓		0.709±0.012	0.865±0.014	0.114±0.048	0.338±0.041
		✓	0.684±0.008	0.854±0.017	0.142±0.014	0.360±0.011
DiPro (Ours)			0.712±0.009	0.885±0.003	0.226±0.019	0.427±0.014
		✓	**0.742±0.003**	**0.897±0.002**	**0.248±0.008**	**0.440±0.007**

using the last available CXR or longitudinal CXR in multimodal settings, or using unimodal EHR, across all tasks, including disease progression identification (Table 1) and general ICU prediction (Table 2). Specifically, it achieves relative gains of 3.8% in F1 score (disease progression), 2.8% in AUPRC (mortality prediction), and 3.0% in accuracy (length of stay) compared to the respective best-performing baselines. Notably, DiPro consistently outperforms longitudinal multimodal specialists like UTDE [19] and UMSE [20]. These results highlight the superiority of DiPro's multiscale fusion strategy in addressing temporal misalignment compared to unimodal imputation or unified time-embedding methods.

DiPro alleviates redundancy and misalignment in sequential CXRs. Table 2 presents a comparison of multimodal models that integrate EHR data with either the most recent CXR (*Last*) or longitudinal CXRs (*Long.*). For UMSE [20], using longitudinal CXRs do not consistently improve general ICU prediction performance compared to using only the latest image, suggesting that unified time embeddings and bottleneck attention inadequately address temporal redundancy in sequential



(a) Disease Progression Prediction

(b) Mortality and Length-of-stay Prediction

Figure 2: Averaged attention weights of CXR regions in different downstream tasks. The radial axis in (a) is log-scaled to enhance distribution visibility. Mean attention weights across CXR regions reveal DiPro's clinical alignment: (a) overlapping distributions for pneumonia, lung opacity, and pleural effusion reflect shared pathologies, while (b) ICU tasks show divergent patterns: higher weights for right-sided regions in mortality (linked to higher risk) versus diffuse attention in length-of-stay (reflecting multifactorial ICU conditions).

imaging data. Similarly, DrFuse [18] experiences a slight performance drop in mortality prediction when naively concatenating sequential CXR features, highlighting the limitations of direct aggregation without temporal alignment. Alternatively, DiPro explicitly disentangles disease progression dynamics from time-invariant CXR features, reducing redundancy. It aligns multimodal trajectories by fusing progression-aware features with EHR data at both interval and sequence levels, enabling more efficient use of longitudinal data. This approach yields superior performance across tasks.

DiPro **echoes with clinical knowledge.** To analyze the anatomical basis of DiPro's decision-making, we visualize the normalized attention weight of CXR regions of each disease in the disease progression task. As shown in Figure 2a, DiPro demonstrates high attention scores on the cardiac silhouette when identifying cardiomegaly, which aligns with the radiographic diagnostic criteria where cardiothoracic ratio (CTR) > 0.5 on posteroanterior chest radiographs indicates cardiac enlargement [36, 37]. Similarly, the model highlights hilar structures for pulmonary edema detection, corroborating the pathophysiological mechanism that pulmonary venous congestion in left ventricular failure manifests as perihilar vascular redistribution and interstitial edema [38]. Notably, the radar plot reveals overlapping attention weight distributions for pneumonia, lung opacity, and pleural effusion, suggesting shared radiographic features due to common pathways [39–41]. This suggests that DiPro captures clinically meaningful correlations in radiographic patterns.

In Figure 2b, we further analyze DiPro's normalized regional attention weights for two ICU prediction tasks. For in-hospital mortality prediction, DiPro assigns notably high attention to the right costophrenic angle, a region clinically associated with pleural effusions and lower lobe pathologies, both of which are common in critically ill patients and have been linked to increased mortality risk [42, 43]. Furthermore, the model consistently prioritizes right-sided anatomical structures over left ones (e.g., right lung: 0.13 vs. left: 0.06; right hilar region: 0.10 vs. left: 0.03). This pattern aligns with the predominance of right-sided pulmonary complications (e.g., aspiration pneumonia and pleural effusion), linked to increased risk of mortality [44–46]. For length-of-stay prediction, however, the model exhibits a more distributed attention pattern across multiple thoracic regions, including bilateral lungs, hilar structures, and mediastinum. This scattered pattern suggests that predicting the length-of-stay requires a broader view of radiographic features, which is consistent with the understanding that hospital stay duration aggregates diverse and multifactorial conditions, such as pulmonary congestion, atelectasis, and cardiomegaly [47, 48].

3.3 Ablation Study

To better each component's contribution in DiPro, we ablate key modules (results in Table 3). The variant "A1" replaces MMF with a simple fusion strategy that concatenates CXR and EHR features, followed by a multi-head self-attention layer. The variant "A2" removes the PAE module. The

Table 3: **Results of the ablation study.** This table presents the results of ablating major modules to assess their contribution to overall performance. Variants "A1"–"A4" correspond to variants of DiPro with progressively removed modules, while "DiPro-" denotes the variant using automated bounding boxes generated by the RGRG model [49] instead of Chest ImaGenome annotations.

	Components			Disease Progression	Mortality	Length of stay
ID	STD	PAE	MMF	F1	AUPRC	ACC
DiPro	✓	✓	✓	**0.466±0.018**	**0.742±0.003**	**0.440±0.007**
DiPro-	✓	✓	✓	0.457±0.010	0.736±0.021	0.430±0.006
A1	✓	✓	✗	0.460±0.014	0.724±0.015	0.416±0.027
A2	✓	✗	✓	0.433±0.017	0.730±0.029	0.432±0.018
A3	✓	✗	✗	0.439±0.007	0.694±0.016	0.404±0.014
A4	✗	✗	✗	0.362±0.016	0.721±0.036	0.425±0.031

variant "A3" removes both MMF and PAE, leaving only a basic aggregation using self-attention. The "A4" variant removes the STD module and disables MMF and PAE, reducing DiPro to a plain encode-concatenate-attention baseline.

As shown in Table 3, removing any component of DiPro results in performance drop, underscoring the necessity of each module. Notably, contributions vary across tasks. Different components contribute variably across tasks. For disease progression identification, the STD module is critical, yielding relatively F1 improvements ("A3" vs. "A4") by 21.3% through disentangling progression dynamics from static anatomical features. The PAE module further enhances performance (7.6% relative F1 gain, DiPro vs. "A2"), highlighting the benefit of progression-aware feature enhancement. However, the MMF module contributes less to the same task, possibly due to the constraint of using only EHRs linked to CXR pairs, restricting the multiscale fusion capability. Conversely, for general ICU prediction, MMF notably improves performance, highlighting the value of a well-designed fusion strategy for handling longitudinal multimodal data.Interestingly, incorporating the STD module into a simple encode-concatenate-attention baseline results in performance degradation ("A3" vs. "A4"), suggesting that disentanglement alone, without dedicated dynamic/static modeling and multiscale fusion, is insufficient for complex prediction tasks. More metrics can be found in Tables 13 and 14.

Robustness with Automated Bounding Boxes To assess DiPro's robustness to automated region annotations, we conducted an ablation study replacing Chest ImaGenome bounding boxes with those generated by the automated region detection model from RGRG [49],denoted as "DiPro-" in Table 3. While using automated bounding boxes results in a slight performance drop compared to curated annotations, DiPro consistently outperforms all baselines across disease progression (0.449±0.005 in F1), mortality (0.722±0.039 in AUPRC), and length-of-stay (0.427±0.014 in ACC) tasks. This demonstrates that DiPro remains effective even in the absence of manually curated labels, supporting its applicability to datasets lacking fine-grained anatomical annotations. While more accurate anatomical annotations can improve prediction performance, DiPro still achieves strong and generalizable results even with fully automated region proposals.

The ablation study on loss penalties is presented in Table 15, the final selected penalty weights for each prediction task are summarized in Table 16, and the robustness analysis under missing EHR data is reported in Tables 18 and 19.

4 Related Work

Modeling disease progression in sequential CXRs. Recent years have seen growing interest in leveraging longitudinal CXRs for clinical prediction, as they are routinely used to monitor disease progression and naturally provide sequential imaging data [13, 14, 33, 50–54]. Most methods focus on capturing temporal differences using deep learning architectures. For instance, Karwande et al. [14] used a graph attention network to model region-level temporal changes, while Eshraghi Dehaghani et al. [54] adopted a Transformer-based detection model for localized progression signals. Wang et al. [50] introduced time-aware causal attention, and Mbakwe et al. [33] proposed a hierarchical Transformer for multi-scale comparison. Other approaches enhance clinical relevance via auxiliary

tasks, such as symptom prediction [13] or spatiotemporal contrastive learning with radiology reports [55]. Despite these efforts, effectively addressing redundancy in sequential CXR modeling remains a fundamental challenge.

Leveraging multimodal data for clinical prediction. Multimodal data offers rich temporal and semantic information for clinical prediction tasks [53, 56–61]. Several methods combine the latest CXR with EHR data to improve performance using sophisticated fusion strategies [17, 18, 62, 63]. More recent efforts target the challenges of heterogeneous, misaligned longitudinal data. For example, Li et al. [58] applied ICA to extract latent EHR signals and aligned them with CT scans using time-aware Transformers. Susman et al. [57] proposed ensemble models to integrate multimodal sequences and highlight salient features for dementia prediction. Others address temporal irregularity directly: Lee et al. [20] introduced unified time embeddings and modality-aware attention, while Zhang et al. [19] imputed sparse clinical notes with temporal attention and fused them with multivariate time series. Yet, the core challenge of capturing disease progression across misaligned modalities remains underexplored [53, 63, 64], limiting our ability to fully leverage cross-modal synergy in clinical contexts.

5 Impacts and Limitations

`DiPro` advances multimodal disease progression modeling through its efficient integration of regional progression-aware feature disentanglement and multi-timescale alignment. The approach demonstrates significant potential for generalization to asynchronous clinical workflows, such as Alzheimer's disease monitoring using longitudinal MRI/PET scans with cognitive test records, or heart failure progression tracking using periodic echocardiograms with continuous vital signs. However, the current implementation relies on anatomical annotations (bounding boxes) to localize progression-specific features. While effective, this requirement may limit scalability in practice. Future work could reduce dependency on manual labels by adopting emerging segmentation tools, such as medical SAMs or weakly supervised localization methods [65], which can better align the framework with real-world clinical workflows. Meanwhile, Our study excludes visits with only a single CXR. This selection may introduce sampling bias and reduce the overall cohort size. However, it is necessary for modeling disease progression between consecutive CXRs, as the longitudinal task inherently requires at least two images per patient. Developing methods that can handle single-CXR visits remains an important direction for future work.

6 Conclusion

In this paper, we propose `DiPro`, a novel framework that tackles critical challenges in fusing longitudinal multimodal data for clinical tasks: redundancy in sequential CXRs and temporal misalignment across modalities. By explicitly disentangling disease progression dynamics from static anatomical features via dedicated constraints, `DiPro` extracts clinical meaningful and discriminative dynamic/static patterns. To further enhance temporal alignment, we propose a multiscale multimodal fusion strategy that bridges CXR-derived progression features with EHR time-series data through interval-wise and full-sequence-level interactions. Extensive experiments demonstrate that `DiPro` achieves state-of-the-art performance on both disease progression identification and general ICU prediction tasks, while providing interpretability consistent with clinical understanding.

Acknowledgments and Disclosure of Funding

This work is partially supported by an Innovation and Technology Fund of Hong Kong Innovation and Technology Commission (project no. ITS/202/23), a Collaborative Research Fund of Hong Kong Research Grants Council (project no. C5055-24G), the National Natural Science Foundation of China (62302413), the Health and Medical Research Fund (23220312), the General Research Fund RGC/HKBU12202621 from the Research Grant Council, and the Research Matching Grant Scheme RMGS2021_8_06 from the Hong Kong Government.

References

[1] Diane R Mould, Nicholas HG Holford, and Carl C Peck. Disease progress models. In *Atkinson's Principles of Clinical Pharmacology*, pages 389–403. Elsevier, 2022.

[2] Michael Moor, Bastian Rieck, Max Horn, Catherine R Jutzeler, and Karsten Borgwardt. Early prediction of sepsis in the ICU using machine learning: a systematic review. *Frontiers in Medicine*, 8:607952, 2021.

[3] Jianyong Zhong, Hai-Chun Yang, and Agnes B Fogo. A perspective on chronic kidney disease progression. *American Journal of Physiology-Renal Physiology*, 312(3):F375–F384, 2017.

[4] Mostafa Mehdipour Ghazi, Mads Nielsen, Akshay Pai, M Jorge Cardoso, Marc Modat, Sébastien Ourselin, Lauge Sørensen, Alzheimer's Disease Neuroimaging Initiative, et al. Training recurrent neural networks robust to incomplete data: Application to Alzheimer's disease progression modeling. *Medical Image Analysis*, 53:39–46, 2019.

[5] Xiang Wang, David Sontag, and Fei Wang. Unsupervised learning of disease progression models. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 85–94, 2014.

[6] Monica A Konerman, Lauren A Beste, Tony Van, Boang Liu, Xuefei Zhang, Ji Zhu, Sameer D Saini, Grace L Su, Brahmajee K Nallamothu, George N Ioannou, et al. Machine learning models to predict disease progression among veterans with hepatitis C virus. *PloS One*, 14(1):e0208141, 2019.

[7] Christopher W Seymour, Foster Gesten, Hallie C Prescott, Marcus E Friedrich, Theodore J Iwashyna, Gary S Phillips, Stanley Lemeshow, Tiffany Osborn, Kathleen M Terry, and Mitchell M Levy. Time to treatment and mortality during mandated emergency care for sepsis. *New England Journal of Medicine*, 376(23):2235–2244, 2017.

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

[9] Laura Swinckels, Frank C Bennis, Kirsten A Ziesemer, Janneke FM Scheerman, Harmen Bijwaard, Ander de Keijzer, and Josef Jan Bruers. The use of deep learning and machine learning on longitudinal electronic health records for the early detection and prevention of diseases: scoping review. *Journal of Medical Internet Research*, 26:e48320, 2024.

[10] Anna Cascarano, Jordi Mur-Petit, Jeronimo Hernandez-Gonzalez, Marina Camacho, Nina de Toro Eadie, Polyxeni Gkontra, Marc Chadeau-Hyam, Jordi Vitria, and Karim Lekadir. Machine and deep learning for longitudinal biomedical data: a review of methods and applications. *Artificial Intelligence Review*, 56 (Suppl 2):1711–1771, 2023.

[11] Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6(1):96, 2019.

[12] Huajun Zhou, Fengtao Zhou, Chenyu Zhao, Yingxue Xu, Luyang Luo, and Hao Chen. Multimodal data integration for precision oncology: Challenges and future directions. *arXiv preprint arXiv:2406.19611*, 2024.

[13] Ye Zhu, Jingwen Xu, Fei Lyu, and Pong C Yuen. Symptom disentanglement in chest X-ray images for fine-grained progression learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 598–607. Springer, 2024.

[14] Gaurang Karwande, Amarachi B Mbakwe, Joy T Wu, Leo A Celi, Mehdi Moradi, and Ismini Lourentzou. CheXRelNet: An anatomy-aware model for tracking longitudinal relationships between chest X-rays. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 581–591. Springer, 2022.

[15] Ellen Dicks, Lisa Vermunt, Wiesje M van der Flier, Pieter Jelle Visser, Frederik Barkhof, Philip Scheltens, Betty M Tijms, Alzheimer's Disease Neuroimaging Initiative, et al. Modeling grey matter atrophy as a function of time, aging or cognitive decline show different anatomical patterns in Alzheimer's disease. *NeuroImage: Clinical*, 22:101786, 2019.

[16] Linda K McEvoy, Dominic Holland, Donald J Hagler Jr, Christine Fennema-Notestine, James B Brewer, and Anders M Dale. Mild cognitive impairment: baseline and longitudinal structural MR imaging measures improve predictive prognosis. *Radiology*, 259(3):834–843, 2011.

[17] 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.

[18] 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.

[19] Xinlu Zhang, Shiyang Li, Zhiyu Chen, Xifeng Yan, and Linda Ruth Petzold. Improving medical predictions by irregular multimodal electronic health records modeling. In *International Conference on Machine Learning*, pages 41300–41313. PMLR, 2023.

[20] 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.

[21] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2097–2106, 2017.

[22] Chantal Pellegrini, Matthias Keicher, Ege Özsoy, Petra Jiraskova, Rickmer Braren, and Nassir Navab. Xplainer: From X-ray observations to explainable zero-shot diagnosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 420–429. Springer, 2023.

[23] Paul Pu Liang, Yun Cheng, Xiang Fan, Chun Kai Ling, Suzanne Nie, Richard Chen, Zihao Deng, Nicholas Allen, Randy Auerbach, Faisal Mahmood, et al. Quantifying & modeling multimodal interactions: An information decomposition framework. *Advances in Neural Information Processing Systems*, 36:27351–27393, 2023.

[24] Xinqi Zhu, Chang Xu, and Dacheng Tao. Where and what? Examining interpretable disentangled representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5861–5870, 2021.

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[26] Qian Jiang, Changyou Chen, Han Zhao, Liqun Chen, Qing Ping, Son Dinh Tran, Yi Xu, Belinda Zeng, and Trishul Chilimbi. Understanding and constructing latent modality structures in multi-modal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7661–7671, 2023.

[27] Mengwei Ren, Neel Dey, Martin Styner, Kelly Botteron, and Guido Gerig. Local spatiotemporal representation learning for longitudinally-consistent neuroimage analysis. *Advances in Neural Information Processing Systems*, 35:13541–13556, 2022.

[28] Ziyu Zhou, Haozhe Luo, Jiaxuan Pang, Xiaowei Ding, Michael Gotway, and Jianming Liang. Learning anatomically consistent embedding for chest radiography. In *BMVC: proceedings of the British Machine Vision Conference. British Machine Vision Conference*, volume 2023, page 617, 2023.

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

[30] 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.

[31] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317, 2019.

[32] Joy Wu, Nkechinyere Agu, Ismini Lourentzou, Arjun Sharma, Joseph Paguio, Jasper Seth Yao, Edward Christopher Dee, William Mitchell, Satyananda Kashyap, Andrea Giovannini, et al. Chest ImaGenome dataset (version 1.0.0). *Physio Net*, 2021.

[33] Amarachi B Mbakwe, Lyuyang Wang, Mehdi Moradi, and Ismini Lourentzou. Hierarchical vision Transformers for disease progression detection in chest X-ray images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 685–695. Springer, 2023.

[34] Yanbo Xu, Siddharth Biswal, Shriprasad R Deshpande, Kevin O Maher, and Jimeng Sun. RAIM: Recurrent attentive and intensive model of multimodal patient monitoring data. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2565–2573, 2018.

[35] Huan Song, Deepta Rajan, Jayaraman Thiagarajan, and Andreas Spanias. Attend and diagnose: Clinical time series analysis using attention models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[36] Tarun Agrawal and Prakash Choudhary. Segmentation and classification on chest radiography: a systematic survey. *The Visual Computer*, 39(3):875–913, 2023.

[37] Erdem Yanar and Firat Hardalaç. Clinical decision support for early diagnosis of cardiomegaly by using deep learning techniques on chest X-rays. In *2023 Computing in Cardiology (CinC)*, volume 50, pages 1–4. IEEE, 2023.

[38] Robert L Wilkins, James R Dexter, and Philip M Gold. *Respiratory disease: a case study approach to patient care*. FA Davis, 2006.

[39] Catia Cillóniz, Celia Cardozo, and Carolina García-Vidal. Epidemiology, pathophysiology, and microbiology of community-acquired pneumonia. *Annals of research hospitals*, 2(1), 2018.

[40] M Elmukhtar, Habas Ala, Said Abdusslam, Rayani Amnna, Farfar Kalifa, Habas Eshrak, Alfitori Gamal, Errayes Almehdi, Habas Aml, Naser Abdel, et al. Diagnostic approach to pleural effusion based on pathogenesis and radiological findings: A narrative review. *Yemen Journal of Medicine*, 3(2):102–113, 2024.

[41] Karen Marcdante and Robert M Kliegman. *Nelson essentials of pediatrics E-book*. Elsevier Health Sciences, 2014.

[42] BTS Guideline. Non-invasive ventilation in acute respiratory failure. *Thorax*, 57(3):192–211, 2002.

[43] Jean-Louis Vincent, Serdar Akça, Arnaldo De Mendonça, Philip Haji-Michael, Charles Sprung, Rui Moreno, Massimo Antonelli, Peter M Suter, SOFA Working Group, et al. The epidemiology of acute respiratory failure in critically ill patients. *Chest*, 121(5):1602–1609, 2002.

[44] Rahul Lohan. Imaging of ICU patients. *Thoracic Imaging: Basic to Advanced*, pages 173–194, 2019.

[45] Adel Salah Bediwy, Mohammed Al-Biltagi, Nermin Kamal Saeed, Hosameldin A Bediwy, and Reem Elbeltagi. Pleural effusion in critically ill patients and intensive care setting. *World Journal of Clinical Cases*, 11(5):989, 2023.

[46] David M Maslove, Benson Tze-Ming Chen, Helena Wang, and Ware G Kuschner. The diagnosis and management of pleural effusions in the ICU. *Journal of Intensive Care Medicine*, 28(1):24–36, 2013.

[47] SP Wright, D Verouhis, G Gamble, K Swedberg, N Sharpe, and RN Doughty. Factors influencing the length of hospital stay of patients with heart failure. *European Journal of Heart Failure*, 5(2):201–209, 2003.

[48] David A Gruenberg, Wayne Shelton, Susannah L Rose, Ann E Rutter, Sophia Socaris, and Glenn McGee. Factors influencing length of stay in the intensive care unit. *American Journal of Critical Care*, 15(5): 502–509, 2006.

[49] Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. Interactive and explainable region-guided radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7433–7442, 2023.

[50] Fuying Wang, Shenghui Du, and Lequan Yu. HERGen: Elevating radiology report generation with longitudinal data. In *European Conference on Computer Vision*, pages 183–200. Springer, 2024.

[51] Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Perez-Garcia, Maximilian Ilse, Daniel C Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, et al. Learning to exploit temporal structure for biomedical vision-language processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15016–15027, 2023.

[52] Francesco Dalla Serra, Chaoyang Wang, Fani Deligianni, Jeff Dalton, and Alison Q O'Neil. Controllable chest X-ray report generation from longitudinal representations. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[53] Luoting Zhuang, Stephen H Park, Steven J Skates, Ashley E Prosper, Denise R Aberle, and William Hsu. Advancing precision oncology through modeling of longitudinal and multimodal data. *arXiv preprint arXiv:2502.07836*, 2025.

[54] Mehrdad Eshraghi Dehaghani, Amirhossein Sabour, Amarachi B Madu, Ismini Lourentzou, and Mehdi Moradi. Representation learning with a Transformer-based detection model for localized chest X-ray disease and progression detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 578–587. Springer, 2024.

[55] Kang Liu, Zhuoqi Ma, Xiaolu Kang, Yunan Li, Kun Xie, Zhicheng Jiao, and Qiguang Miao. Enhanced contrastive learning with multi-view longitudinal data for chest X-ray report generation. *arXiv preprint arXiv:2502.20056*, 2025.

[56] Virgilio Kmetzsch, Emmanuelle Becker, Dario Saracino, Daisy Rinaldi, Agnes Camuzat, Isabelle Le Ber, Olivier Colliot, PREV-DEMALS study group, et al. Disease progression score estimation from multimodal imaging and microRNA data using supervised variational autoencoders. *IEEE Journal of Biomedical and Health Informatics*, 26(12):6024–6035, 2022.

[57] Aviad Susman, Rupak Krishnamurthy, Yan Chak Li, Mohammad Olaimat, Serdar Bozdag, Bino Varghese, Nasim Sheikh-Bahaei, and Gaurav Pandey. Longitudinal ensemble integration for sequential classification with multimodal data. *arXiv preprint arXiv:2411.05983*, 2024.

[58] Thomas Z Li, John M Still, Kaiwen Xu, Ho Hin Lee, Leon Y Cai, Aravind R Krishnan, Riqiang Gao, Mirza S Khan, Sanja Antic, Michael Kammer, et al. Longitudinal multimodal Transformer integrating imaging and latent clinical signatures from routine EHRs for pulmonary nodule classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 649–659. Springer, 2023.

[59] Elisa Warner, Joonsang Lee, William Hsu, Tanveer Syeda-Mahmood, Charles E Kahn Jr, Olivier Gevaert, and Arvind Rao. Multimodal machine learning in image-based and clinical biomedicine: Survey and prospects. *International Journal of Computer Vision*, 132(9):3753–3769, 2024.

[60] Yidan Feng, Bohan Zhang, Sen Deng, Zhanli Hu, and Jing Qin. Asynchronous multi-modal learning for dynamic risk monitoring of acute respiratory distress syndrome in intensive care units. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 14–23. Springer, 2025.

[61] Chenyu Lian, Hong-Yu Zhou, Dongyun Liang, Jing Qin, and Liansheng Wang. Efficient medical vision-language alignment through adapting masked vision models. *IEEE Transactions on Medical Imaging*, 2025.

[62] Yidan Feng, Bingchen Gao, Sen Deng, Anqi Qiu, and Jing Qin. Unified multi-modal learning for any modality combinations in Alzheimer's disease diagnosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 487–497. Springer, 2024.

[63] Wenfang Yao, Chen Liu, Kejing Yin, William Cheung, and Jing Qin. Addressing asynchronicity in clinical multimodal fusion via individualized chest X-ray generation. *Advances in Neural Information Processing Systems*, 37:29001–29028, 2024.

[64] Qingyu Zhao, Ehsan Adeli, and Kilian M Pohl. Longitudinal correlation analysis for decoding multi-modal brain development. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 400–409. Springer, 2021.

[65] Jay N. Paranjape, Shameema Sikder, S. Swaroop Vedula, and Vishal M. Patel. S-SAM: SVD-based fine-tuning of segment anything model for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 720–730. Springer, 2024.

[66] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

[67] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[68] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-Flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pages 353–367. PMLR, 2023.

[69] Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. Multi-modal learning with missing modality via shared-specific feature modelling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15878–15887, 2023.

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 clearly describe the two core challenges—temporal redundancy in CXRs and multimodal misalignment—and correspond directly to the proposed solutions: disentanglement of dynamic/static features and hierarchical multimodal fusion. These claims are substantiated through method design and validated by comprehensive experiments.

   Guidelines:

   - The answer NA 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 NA 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 paper discusses the limitations in Section 5. We will address these limitations in future work.

   Guidelines:

   - The answer NA 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:[NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA 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: We provide the implementation settings used to obtain the main experimental results in the appendix and supplementary materials. Detailed results and analyses can be found in Section 3.

Guidelines:

- The answer NA 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: The source code is submitted as supplementary material alongside the paper and will be made publicly available upon acceptance. All datasets used in this study are open-source and publicly accessible.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.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://nips.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, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Details of training and testing, including data splits, preprocessing steps, and hyperparameter search space, are provided in Appendix A.2.1.

   Guidelines:

   - The answer NA 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 results are reported by taking the average of three runs of model training along with the standard deviations.

   Guidelines:

   - The answer NA 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 paper provides detailed information on the compute resources used, including GPU types, memory, and architecture for all experiments in the Appendix A to ensure reproducibility.

   Guidelines:

   - The answer NA 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 research strictly follows the NeurIPS Code of Ethics by ensuring data privacy, avoiding harm to participants, adhering to institutional protocols, and considering potential societal impacts such as fairness, bias, and security throughout the study.

   Guidelines:

   - The answer NA 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 discusses both the potential impacts of the proposed work in Section 5.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA 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., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our work does not involve the release of high-risk models or scraped datasets that could pose misuse risks, and therefore no additional safeguards were required.

    Guidelines:

    - The answer NA 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: We use publicly available datasets such as MIMIC, which is cited properly along with its license terms (PhysioNet Credentialed Health Data License). Additionally, all external codebases and models referenced are appropriately credited with their original citations and licenses mentioned where applicable.

    Guidelines:

    - The answer NA 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: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA 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: [NA]

    Justification: This paper does not involve crowdsourcing experiments or direct research with human subjects. All experiments are conducted on publicly available, de-identified dataset, so participant instructions or compensation details are not relevant.

    Guidelines:

    - The answer NA 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: [NA]

Justification: The study uses publicly available, de-identified data (e.g., MIMIC), which does not involve direct interaction with human participants.

Guidelines:

- The answer NA 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 rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: LLM is only used for editing and formatting purpose of this paper.

Guidelines:

- The answer NA 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 (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Experiment Details

A.1 Details of Data Preprocessing

EHR Data Preprocessing We adapted the EHR data processing pipeline from [17], modifying the sampling frequency from 2-hour to 1-hour intervals. We incorporate vital signs, laboratory measurements, and clinical scores, resulting in a total of 38 clinical time-series variables, including alanine aminotransferase, albumin, alkaline phosphate, anion gap, asparate aminotransferase, bicarbonate, bilirubin, blood urea nitrogen, chloride, creatinine, diastolic blood pressure, fraction inspired oxygen, Glasgow coma scale (eye opening, motor response, and verbal response), glucose, heart rate, height, hematocrit, hemoglobin, magnesium, mean blood pressure, oxygen saturation, partial pressure of carbon dioxide, partial thromboplastin time, platelets, positive end-expiratory pressure, potassium, prothrombin time, respiratory rate, sodium, systolic blood pressure, temperature, troponin-T, urine output, daily weight, white blood cell count, and pH. Following [17], we applied identical discretization and standardization procedures. Demographic data including age, height, admission weight, gender, race, language, and marital status were also incorporated.

CXR Preprocessing CXR studies were temporally aligned with corresponding ICU stays to construct longitudinal imaging sequences. We retained only AP view images, matched them to ICU stays based on study timestamps, and excluded outlier admissions using length-of-stay filtering. Each matched CXR was restricted to those with available bounding-box annotations in Chest ImaGenome. Meanwhile, in this study, we do not perform explicit longitudinal registration of CXR images. To mitigate potential positional misalignment over time, we follow the preprocessing strategy of CheXRelNet [14]: we apply cropping based on anatomical bounding boxes from Chest ImaGenome dataset. This preprocessing allows us to focus on capturing the semantic-level disease progression within consistent anatomical structure, rather than relying on pixel-level alignment. As a result, our model is designed to be more robust to positional variability across different time points.

Data Splitting and Statistics Our analysis focused on ICU stays containing at least two CXRs. The dataset was partitioned into training (70%), validation (10%), and test sets (20%) at the subject level to prevent data leakage. Table 4 summarizes the sample counts for each task, while Table 5 and Table 6 detail the label distributions for disease progression identification and general ICU prediction tasks, respectively. The distribution of CXR examinations per patient is presented in Table 7.

Table 4: Data statistics in training, validation, and testing sets for each task.

Task	Training	Validation	Test
Disease Progression Identification	3982	560	1137
General ICU Prediction	1889	285	546

Table 5: Label distribution for the Disease Progression Identification task.

Category	Atelectasis	Enlarged Cardiac Silhouette	Consolidation	Pulmonary Edema	Lung Opacity	Pleural Effusion	Pneumonia
Improved	328 (15.3%)	83 (4.6%)	124 (14.2%)	572 (31.9%)	784 (19.2%)	317 (13.1%)	98 (12.7%)
Worsened	674 (31.4%)	141 (7.8%)	285 (32.7%)	597 (33.3%)	1273 (31.2%)	702 (28.9%)	387 (50.1%)
No Change	1143 (53.3%)	1575 (87.5%)	463 (53.1%)	624 (34.8%)	2025 (49.6%)	1407 (58.0%)	287 (37.2%)
Total	2145	1799	872	1793	4082	2426	772

Table 6: Label distribution for the General ICU Prediction Tasks.

	Mortality		Length of Stay			
	0	1	$[2, 3)$	$[3, 4)$	$[4, 6)$	> 6
Count (%)	2255 (82.9%)	465 (17.1%)	793 (29.2%)	641 (23.6%)	687 (25.3%)	599 (22.0%)

Table 7: Data statistics of Numbers of CXR in training, validation, and testing sets for the General ICU Prediction Task.

Split	2	3	4	5
Training	1367	465	56	1
Validation	199	76	10	0
Test	409	120	16	1
Total	1975	661	82	2

A.2 Implementation Details

A.2.1 Details of Architectures and Training Procedures of `DiPro`

The training and validation processes are executed on a server equipped with a RTX 3090-24GB GPU card and a 14 vCPU Intel(R) Xeon(R) Gold 6330 CPU. The method is implemented using PyTorch 1.9.1 and PyTorch-Lightning 1.4.2 with CUDA 11.1 environment. AdamW optimizer and CosineAnnealingLR learning rate schedular are used for training.

Model Architecture and Hyperparameters. `DiPro` consists of four major components: (1) a **CXR processing module**, which employs a shared ResNet backbone to extract regional visual features, followed by a multi-layer perceptron (MLP) for feature adjustment and two parallel MLP-based projection heads that encode static and dynamic representations; (2) an **EHR processing module**, which includes a one-layer multivariate transformer encoder for global temporal modeling, a local multi-head attention layer for capturing short-term dependencies, a time-embedding MLP for relative temporal encoding, and a separate MLP for demographic features; (3) a **multimodal fusion module**, which integrates CXR and EHR representations through local and global attention layers, followed by a lightweight transformer block and a group-based static feature fusion mechanism; and (4) a **prediction head**, implemented as a task-specific MLP for downstream classification or regression. The detailed hyperparameter settings and implementation specifics are provided in the released code repository.

Training Configuration The model was trained with base batch sizes of 8 (for disease progression identification) and 4 (for general ICU prediction), using 4-step gradient accumulation to achieve an effective batch size of 32 or 16. Training proceeded for a maximum of 100 epochs with early stopping triggered after 10 epochs without validation improvement. Task-specific selection metrics were employed: macro-F1 for disease identification, accuracy for length-of-stay classification, and AUPRC for mortality prediction. The hyperparameter search spaces for each task are documented in Table 8.

Table 8: Hyperparameter search space used for model tuning.

Hyperparameter	Search Grid
Learning rate	8×10^{-6}, 5×10^{-6}, 1×10^{-5}, 5×10^{-5}
Dropout rate	0.1, 0.2, 0.3
Hidden dimension	64, 128, 256
λ_{temp}	0.01, 0.001, 0.1, 1.0
λ_{pred}	2, 6, 10
λ_{PAE}	0.01, 0.1, 2,
λ_{orth}	0.001, 0.01, 0.1, 10

A.2.2 Implementation of Baselines

Since none of the baselines can handle both disease identification and general ICU prediction tasks within a unified framework like ours, we introduce minimal modifications to adapt existing approaches. For sequential CXR disease progression specialists (e.g., CheXRelNet [14], CheXRelFormer [33], and SDPL [13]), we concatenate EHR time series and demographic data, then integrate the same EHR encoder and attention fusion layer as our model for fair comparison. For multimodal fusion

baselines, we adapt UTDE [19], originally designed for longitudinal clinical notes and EHR, by replacing its text encoder with our image encoder to process longitudinal CXRs. Similarly, we modify MedFuse [17] by concatenating CXR representations at the sequence level before fusion, and extend DrFuse [18] by first disentangling CXR-EHR pairs and then concatenating the CXR features for fusion.

Hyperparameter Search. We conducted a unified hyperparameter search for all baseline models using the following grid: learning rates $\{1 \times 10^{-5}, 4 \times 10^{-7}, 1 \times 10^{-6}, 1 \times 10^{-7}\}$, dropout rates $\{0.1, 0.2, 0.3\}$, and hidden dimensions $\{64, 128, 256, 320\}$. All other hyperparameters were kept consistent with those specified in the original implementations provided by the official source code of each baseline.

Computational Efficiency and Inference Cost Comparison We summarize the comparison of computational efficiency and inference cost between `DiPro` and baseline models in Table 9, evaluated under the multimodal input setting (sequential CXR + EHR) for the disease progression identification task. Compared to the CheXRelNet baseline, which is also an anatomical region-based model, `DiPro` reduces FLOPs by 9.8% and latency by 22.0%, while achieving a relative gain of 16.5% in F1 for disease progression.

Table 9: Computational efficiency and inference cost comparison between `DiPro` and baseline models. (F1 scores are extracted from Table 10.)

Model	Params (M)	FLOPs (G)	MACs (G)	Latency Mean (ms)	Throughput (samples/s)	F1 Score
`DiPro` (Ours)	31.06	82.75	41.37	23.05	43.38	**0.466±0.018**
CheXRelNet [14]	49.15	90.90	45.45	27.34	36.58	0.382±0.016
CheXRelFormer [33]	49.13	19.95	9.98	14.86	67.28	0.352±0.021
SDPL [13]	34.24	9.45	4.73	13.35	74.90	0.393±0.010
UTDE [19]	6.69	1.61	0.80	5.24	190.90	0.449±0.005
UMSE [20]	23.99	8.27	4.14	9.65	103.62	0.352±0.013
MedFuse [17]	27.21	8.28	4.14	9.78	102.23	0.409±0.042
DrFuse [18]	56.40	16.45	8.23	19.93	50.18	0.429±0.010

B Additional Results

Zero-Weight Ablation Study for Loss Penalties We conducted ablation experiments in which each loss component, λ_{orth} (orthogonal disentanglement loss), λ_{temp} (temporal consistency loss for static features), and λ_{PAE} (Progression-Aware Enhancement loss), was individually disabled by setting its corresponding weight to zero. The results are summarized in Table 15. Key observations include:

- **Orthogonal disentanglement loss (λ_{orth}):** Removal caused notable decreases in disease progression and length-of-stay predictions, emphasizing the benefit of disentangling dynamic pathology from static anatomy.
- **Progression-Aware Enhancement loss (λ_{PAE}):** Disabling this loss reduced disease progression prediction performance, indicating its role in enhancing sensitivity to progression direction.
- **Temporal consistency loss (λ_{temp}):** Its removal led to significant performance drops across ICU tasks, highlighting the importance of maintaining longitudinal consistency of static anatomical information.
- These results demonstrate that all three loss components contribute meaningfully and in distinct ways to the model's overall performance across different tasks. We report the final selected penalty weights for each prediction task in Table 16, which reflect the relative contribution of different regularization terms to overall model performance.

Table 10: **Detailed Performance Comparison on Disease Progression Identification Tasks.** Sequential CXR disease progression specialists (e.g., CheXRelNet [14], CheXRelFormer [33], and SDPL [13]), originally designed for unimodal sequential CXR inputs, are extended to multimodal integration by incorporating a transformer-based EHR encoder and applying cross-attention fusion. Methods that can naturally process uni-CXR inputs (e.g., UMSE [20] and MedFuse [17]) are included for comparison under the unimodal CXR setting. Numbers in **bold** indicate the best overall performance, while underlined values denote the top-performing baseline.

Method	Precision	Recall	F1	AUPRC	AUROC	ACC
Unimodal Methods (CXR)						
CheXRelNet [14]	0.395±0.015	0.392±0.010	0.389±0.010	0.394±0.010	0.574±0.011	0.508±0.013
CheXRelFormer [33]	0.389±0.044	0.379±0.033	0.354±0.032	0.372±0.023	0.551±0.041	0.446±0.057
SDPL [13]	0.408±0.006	0.406±0.020	0.393±0.010	0.417±0.032	0.609±0.031	0.538±0.024
UMSE [20]	0.337±0.004	0.337±0.008	0.329±0.008	0.347±0.004	0.513±0.006	0.476±0.004
MedFuse [17]	0.439±0.006	0.440±0.009	0.433±0.010	0.453±0.011	0.643±0.006	0.543±0.024
DiPro (ours)	**0.475±0.004**	**0.452±0.011**	**0.453±0.009**	**0.468±0.013**	**0.651±0.016**	**0.567±0.007**
Unimodal Methods (EHR)						
Transformer [29]	0.412±0.041	0.358±0.010	0.327±0.013	0.384±0.011	0.569±0.008	0.509±0.009
Multimodal Methods						
CheXRelNet [14]	0.391±0.020	0.387±0.019	0.382±0.016	0.390±0.019	0.572±0.028	0.504±0.024
CheXRelFormer [33]	0.359±0.019	0.362±0.026	0.352±0.021	0.371±0.016	0.544±0.022	0.474±0.009
SDPL [13]	0.409±0.009	0.398±0.006	0.393±0.003	0.401±0.008	0.582±0.008	0.529±0.014
UTDE [19]	0.481±0.017	0.462±0.002	0.449±0.005	0.472±0.014	0.659±0.011	0.527±0.016
UMSE [20]	0.353±0.011	0.361±0.009	0.352±0.013	0.364±0.006	0.544±0.004	0.484±0.011
MedFuse [17]	0.423±0.049	0.413±0.045	0.409±0.042	0.422±0.040	0.609±0.05 0	0.530±0.030
DrFuse [18]	0.442±0.009	0.461±0.007	0.429±0.010	0.438±0.003	0.628±0.002	0.475±0.021
DiPro (ours)	**0.484±0.008**	**0.471±0.024**	**0.466±0.018**	**0.478±0.018**	**0.664±0.013**	**0.565±0.013**

Table 11: **Performance Comparison on Disease Progression Identification Tasks across Different Diseases**. This table reports the F1 performance of various unimodal and multimodal methods across seven disease progression identification tasks.

Method	Atelectasis	Enlarged Cardiac Silhouette	Consolidation	Pulmonary Edema	Lung Opacity	Pleural Effusion	Pneumonia
Unimodal Methods (CXR)							
CheXRelNet [14]	0.425	**0.340**	0.381	0.408	0.398	0.366	0.407
CheXRelFormer [33]	0.359	0.318	0.319	0.390	0.354	0.380	0.357
SDPL [13]	0.396	0.362	0.350	0.439	0.431	0.443	0.331
DiPro (Ours)	**0.436**	0.338	**0.388**	**0.527**	**0.523**	**0.509**	**0.452**
Multimodal Methods							
UTDE [19]	0.445	0.338	**0.402**	0.470	0.503	0.478	**0.510**
UMSE [20]	0.352	0.313	0.343	0.346	0.384	0.354	0.368
MedFuse [17]	0.422	0.340	0.392	0.455	0.447	0.433	0.363
DrFuse [18]	0.434	0.310	0.354	0.499	0.464	0.447	0.498
DiPro (Ours)	**0.453**	**0.362**	0.399	**0.530**	**0.509**	**0.500**	**0.510**

Broader Baseline Comparison with Large Vision-Language Models. To provide a comprehensive evaluation, we include recent large-scale vision-language models (VLMs) capable of processing multi-image inputs (e.g., Gemma3 [66], QWen2.5VL [67], and Med-Flamingo [68]) as baselines for disease progression identification. We input consecutive CXR pairs and use a few-shot prompting approach (following [68]) to have the VLMs predict progression status for thoracic conditions (results in Table 17). DiPro achieves notably higher accuracy with lower computational cost, highlighting the advantages of our design: spatialtemporal disentanglement (STD) and progression-aware enhancement (PAE) for clinical prediction.

Robustness to Missing EHR Data We conducted additional experiments on the general ICU tasks, where EHR data were randomly dropped during training at rates of 25%, 50%, and 75% (following [69]), while testing was performed on the complete dataset. Notably, although model performance naturally decreases as the EHR missing rate increases, DiPro consistently outperforms

Table 12: **Detailed Performance Comparison on General ICU Prediction Tasks.** The "Input Modalities" section specifies the data sources used for each method. "Last" and "Long." indicate CXR-only input settings, where "Last" denotes the use of the last available CXR and "Long." represents the use of longitudinal CXRs. The inclusion of EHR data is indicated by the "EHR" column. "Sequential CXR Disease Progression Specialists" are extended to multimodal integration by incorporating a transformer-based EHR encoder and applying cross-attention fusion. For "Clinical Multimodal Fusion Specialists", which were originally designed for single-CXR inputs, we modify MedFuse [17] by concatenating CXR representations at the sequence level before fusion, and extend DrFuse [18] by first disentangling CXR–EHR pairs before concatenating CXR features for multimodal fusion.

	Input Modalities			Mortality		Length of Stay			
Method	Last	Long.	EHR	AUPRC	AUROC	Kappa	ACC	F1	AUPRC
Unimodal Methods (EHR)									
Transformer [29]			✓	0.712±0.009	0.885±0.003	0.226±0.019	0.427±0.014	0.360±0.024	0.386±0.014
Sequential CXR Disease Progression Specialists									
ChexRelNet [14]		✓		0.291±0.050	0.624±0.036	0.039±0.020	0.291±0.014	0.238±0.010	0.275±0.004
		✓	✓	0.697±0.040	0.876±0.015	0.166±0.034	0.380±0.028	0.355±0.009	0.358±0.015
ChexRelFormer [33]		✓		0.218±0.011	0.522±0.019	0.005±0.032	0.267±0.043	0.212±0.019	0.255±0.017
		✓	✓	0.522±0.041	0.766±0.021	0.103±0.014	0.333±0.005	0.306±0.005	0.335±0.007
SDPL [13]		✓		0.261±0.006	0.608±0.035	0.011±0.007	0.267±0.010	0.154±0.022	0.261±0.005
		✓	✓	0.717±0.024	0.878±0.019	0.231±0.009	0.430±0.009	**0.385±0.011**	0.404±0.012
Longitudinal Multimodal Specialists									
UTDE [19]	✓		✓	0.717±0.019	0.887±0.004	0.160±0.016	0.381±0.013	0.324±0.005	0.361±0.012
		✓	✓	0.710±0.019	0.887±0.012	0.195±0.031	0.400±0.021	0.346±0.039	0.365±0.028
UMSE [20]	✓		✓	0.722±0.039	0.896±0.012	0.217±0.013	0.419±0.010	0.350±0.026	0.356±0.018
		✓	✓	0.712±0.028	0.891±0.011	0.204±0.019	0.410±0.013	0.342±0.021	0.357±0.018
Clinical Multimodal Fusion Specialists									
MedFuse [17]	✓		✓	0.686±0.018	0.869±0.011	0.213±0.012	0.413±0.004	0.362±0.025	**0.412±0.010**
		✓	✓	0.716±0.018	0.881±0.005	0.210±0.039	0.412±0.027	0.350±0.006	0.410±0.019
DrFuse [18]	✓		✓	0.709±0.012	0.865±0.014	0.114±0.048	0.338±0.041	0.325±0.035	0.316±0.024
		✓	✓	0.684±0.008	0.854±0.017	0.142±0.014	0.360±0.011	0.348±0.006	0.329±0.004
DiPro (Ours)		✓		0.319±0.018	0.637±0.015	0.029±0.017	0.284±0.010	0.227±0.014	0.278±0.004
		✓	✓	**0.742±0.003**	**0.897±0.002**	**0.248±0.008**	**0.440±0.007**	0.384±0.018	0.409±0.010

Table 13: Results of the ablation study. (Disease Progression Identification Task)

	Precision	Recall	F1	AUPRC	AUROC	ACC
DiPro	**0.484±0.008**	**0.471±0.024**	**0.466±0.018**	**0.478±0.018**	**0.664±0.013**	**0.565±0.013**
w/o MMF	0.481±0.009	0.460±0.013	0.460±0.014	0.472±0.008	0.654±0.017	0.580±0.010
w/o PAE	0.443±0.024	0.446±0.023	0.433±0.017	0.461±0.015	0.646±0.018	0.552±0.024
w/o STD	0.372±0.014	0.371±0.015	0.362±0.016	0.377±0.007	0.556±0.006	0.491±0.027
w/o (PAE+MMF)	0.455±0.017	0.452±0.015	0.439±0.007	0.461±0.008	0.647±0.003	0.536±0.016

Table 14: Results of the ablation study. (Length of Stay Classification)

	kappa	Precision	Recall	F1	AUPRC	AUROC	ACC
DiPro	**0.248±0.008**	**0.431±0.037**	**0.433±0.005**	0.384±0.018	**0.409±0.010**	**0.688±0.005**	**0.440±0.007**
w/o MMF	0.214±0.038	0.396±0.053	0.408±0.031	0.362±0.015	0.388±0.009	0.673±0.008	0.416±0.027
w/o PAE	0.235±0.022	0.405±0.021	0.423±0.016	**0.386±0.007**	0.408±0.018	**0.688±0.005**	0.432±0.018
w/o STD	0.225±0.043	0.375±0.082	0.415±0.031	0.372±0.044	0.380±0.025	0.667±0.025	0.425±0.031
w/o (PAE+MMF)	0.194±0.014	0.367±0.013	0.390±0.010	0.357±0.009	0.380±0.009	0.663±0.006	0.404±0.014

Table 15: Ablation study of loss components by setting individual loss weights to zero. '-' indicates that disease progression prediction could not be computed due to $\mathcal{L}_{\text{temp}}$ requiring at least three CXRs.

	Components			Disease Progression	Mortality	Length of stay
ID	λ_{orth}	λ_{PAE}	λ_{temp}	F1	AUPRC	ACC
DiPro	✓	✓	✓	0.466 ± 0.018	0.742 ± 0.003	0.440 ± 0.007
B1	✗	✓	✓	0.448 ± 0.004	0.749 ± 0.028	0.415 ± 0.008
B2	✓	✗	✓	0.441 ± 0.013	0.708 ± 0.038	0.432 ± 0.005
B3	✓	✓	✗	-	0.737 ± 0.008	0.415 ± 0.025

Table 16: Final selected penalty weights for each prediction task.

Prediction Task	λ_{pred}	λ_{orth}	λ_{temp}	λ_{PAE}
In-hospital Mortality	6	0.1	1	0.1
Length of Stay	10	0.001	0.1	0.1
Disease Progression	2	1	–	2

Table 17: Performance Comparison on Disease Progression Identification Task between VLMs and DiPro.

Model	Precision (P)	Recall (R)	F1 Score	Throughput (samples/s)
DiPro (Uni-CXR) (Ours)	**0.475**	**0.452**	**0.453**	**43.38**
Gemma3 [66]	0.329	0.328	0.279	0.088
QWen2.5VL [67]	0.304	0.330	0.251	0.097
Med-Flamingo [68]	0.355	0.345	0.301	0.234

all established baselines across all missing-data scenarios, demonstrating strong robustness in handling incomplete multimodal inputs (see Tables 18 and 19).

Table 18: Length of Stay Classification (Accuracy) under different EHR missing rates.

Method	75% Missing	50% Missing	25% Missing	0% Missing
DiPro (Ours)	**0.399 ± 0.011**	**0.391 ± 0.021**	**0.415 ± 0.005**	**0.440 ± 0.007**
UMSE [20]	0.339 ± 0.029	0.376 ± 0.005	0.396 ± 0.009	0.410 ± 0.013
UTDE [19]	0.377 ± 0.015	0.386 ± 0.012	0.384 ± 0.034	0.400 ± 0.021
DrFuse [18]	0.340 ± 0.009	0.360 ± 0.010	0.363 ± 0.011	0.360 ± 0.011
MedFuse [17]	0.332 ± 0.039	0.360 ± 0.033	0.360 ± 0.025	0.412 ± 0.027

Table 19: Mortality Prediction (AUPRC) under different EHR missing rates.

Method	75% Missing	50% Missing	25% Missing	0% Missing
DiPro (Ours)	**0.696 ± 0.011**	**0.718 ± 0.012**	**0.751 ± 0.006**	**0.742 ± 0.003**
UMSE [20]	0.648 ± 0.009	0.685 ± 0.010	0.686 ± 0.051	0.712 ± 0.028
UTDE [19]	0.603 ± 0.052	0.673 ± 0.020	0.697 ± 0.021	0.710 ± 0.019
DrFuse [18]	0.602 ± 0.024	0.663 ± 0.024	0.674 ± 0.077	0.684 ± 0.008
MedFuse [17]	0.609 ± 0.057	0.641 ± 0.026	0.705 ± 0.016	0.716 ± 0.018