Multimodal Disease Progression Modeling via Spatiotemporal 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

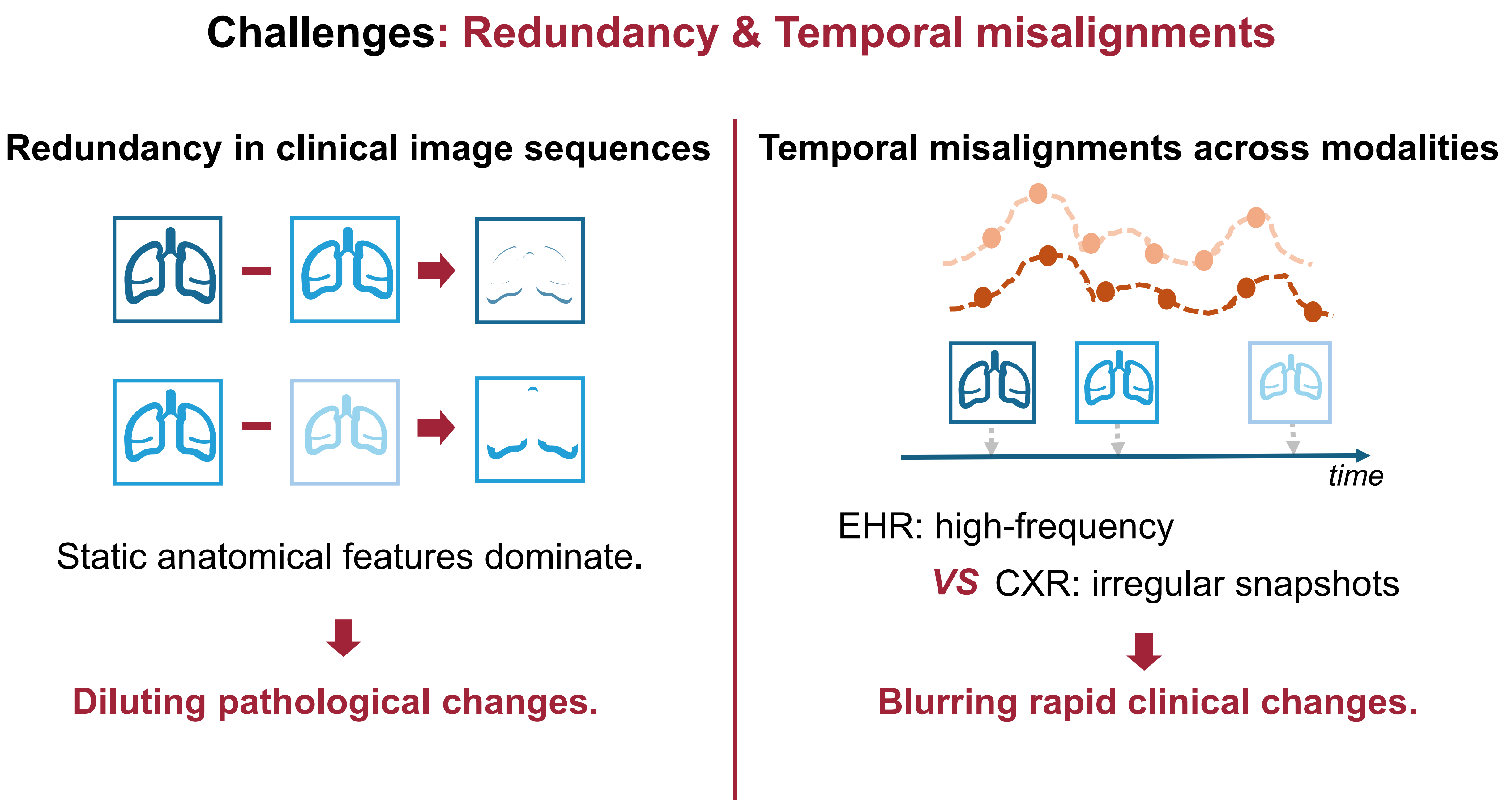
https://github.com/Chenliu-svg/DiPro

Motivation: Multimodal Longitudinal Modeling for Enhanced Diagnosis

A real case from MIMIC:

	Manifestations	Progression?	Diagnosis?
Single-timepoint CXR	Pleural Effusion Pulmonary Edema Lung Opacity	***Unknown!***	Pneumonia? CHF? ARDS? ***High uncertainty!***
Longitudinal CXRs (time)	Pleural Effusion Pulmonary Edema Lung Opacity	Unchanged Worsened Worsened	**Severe Sepsis** ***Enhanced diagnosis***
+ EHR	Heart Rate MAP Respiratory Rate	Increase Decrease Increase	

Multimodal disease progression


Challenges: Redundancy & Temporal misalignments
Redundancy in clinical image sequences
Static anatomical features dominate.
Diluting pathological changes.
Temporal misalignments across modalities
time
EHR: high-frequency
VS CXR: irregular snapshots
Blurring rapid clinical changes.

Our Solution: Disease Progression-Aware Clinical Prediction (DiPro)

Spatiotemporal Disentanglement (STD)

Dynamic pathological changes

Static anatomical structures

Progression-Aware Enhancement (PAE)

Learns progression direction

via **reversal**

Multiscale Multimodal Fusion (MMF)

Local (pairwise interval-level)

Global (full-sequence)

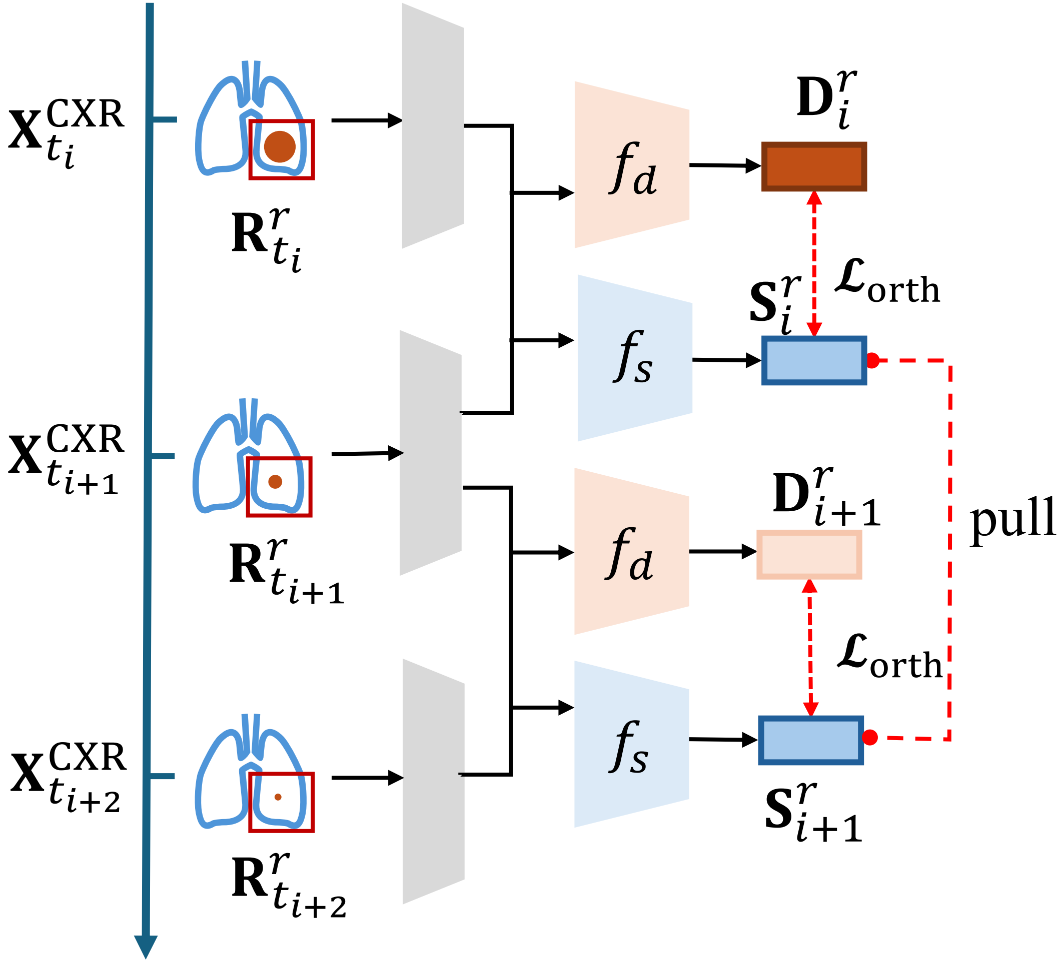
Clinical Tasks

Disease Progression

General ICU Predictions

Our Solution: Spatiotemporal Disentanglement (STD)

Goal: Disentangle region-based **time-invariant (static)** and **time-variant (dynamic)** information.



Feature extraction:

Static feature: $\mathbf{S}_i^r = f_s([\mathbf{F}_{t_i}^r || \mathbf{F}_{t_{i+1}}^r])$

Dynamic feature: $\mathbf{D}_i^r = f_d([\mathbf{F}_{t_i}^r || \mathbf{F}_{t_{i+1}}^r])$

Orthogonal disentanglement loss:

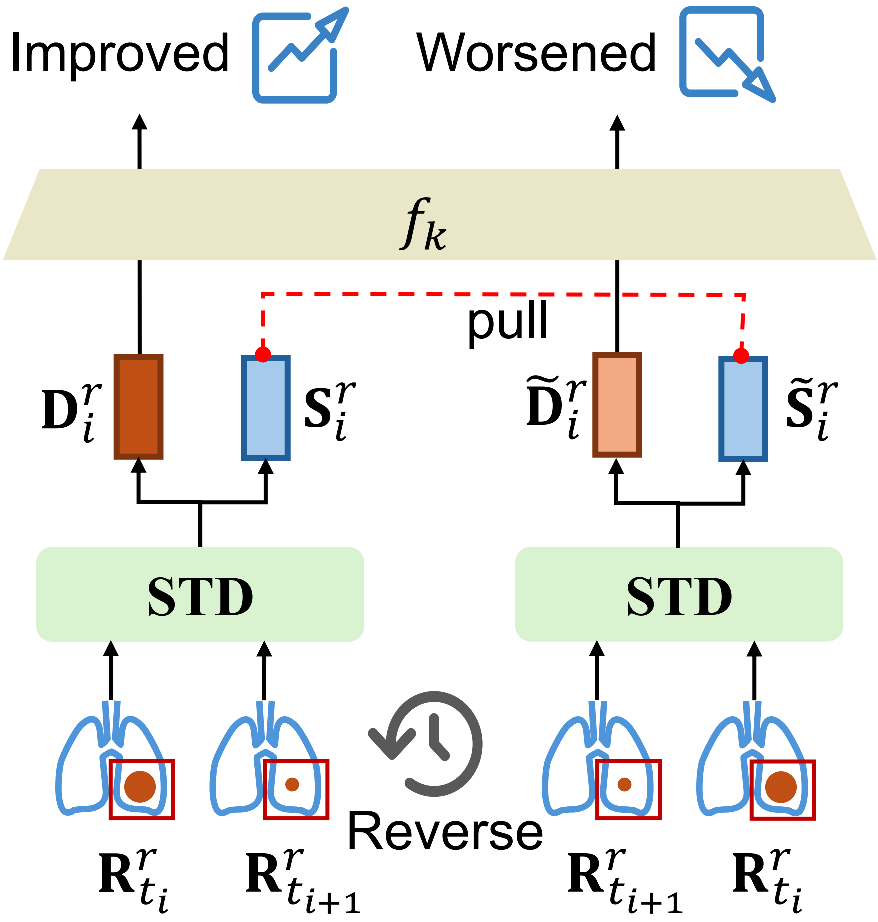
$$\mathcal{L}_{\text{orth}} = \frac{1}{(T-1)R} \sum_{i=1}^{T-1} \sum_{r=1}^{R} (\text{sim}(\mathbf{S}_i^r, \mathbf{D}_i^r))^2$$

Temporal consistency for static features:

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

Our Solution: Progression-Aware Enhancement (PAE)

Goal: Improve the model's sensitivity to progression direction.



Reversed dynamic and static features:

Reversed static feature: $\tilde{\mathbf{S}}_i^r = f_s([\mathbf{F}_{t_{i+1}}^r || \mathbf{F}_{t_i}^r])$

Reversed dynamic feature: $\widetilde{\mathbf{D}}_i^r = f_d([\mathbf{F}_{t_{i+1}}^r || \mathbf{F}_{t_i}^r])$

Region-based disease progression prediction:

Predicted original direction: $\hat{y}_i^{r,k} = f_k(\mathbf{D}_i^r)$

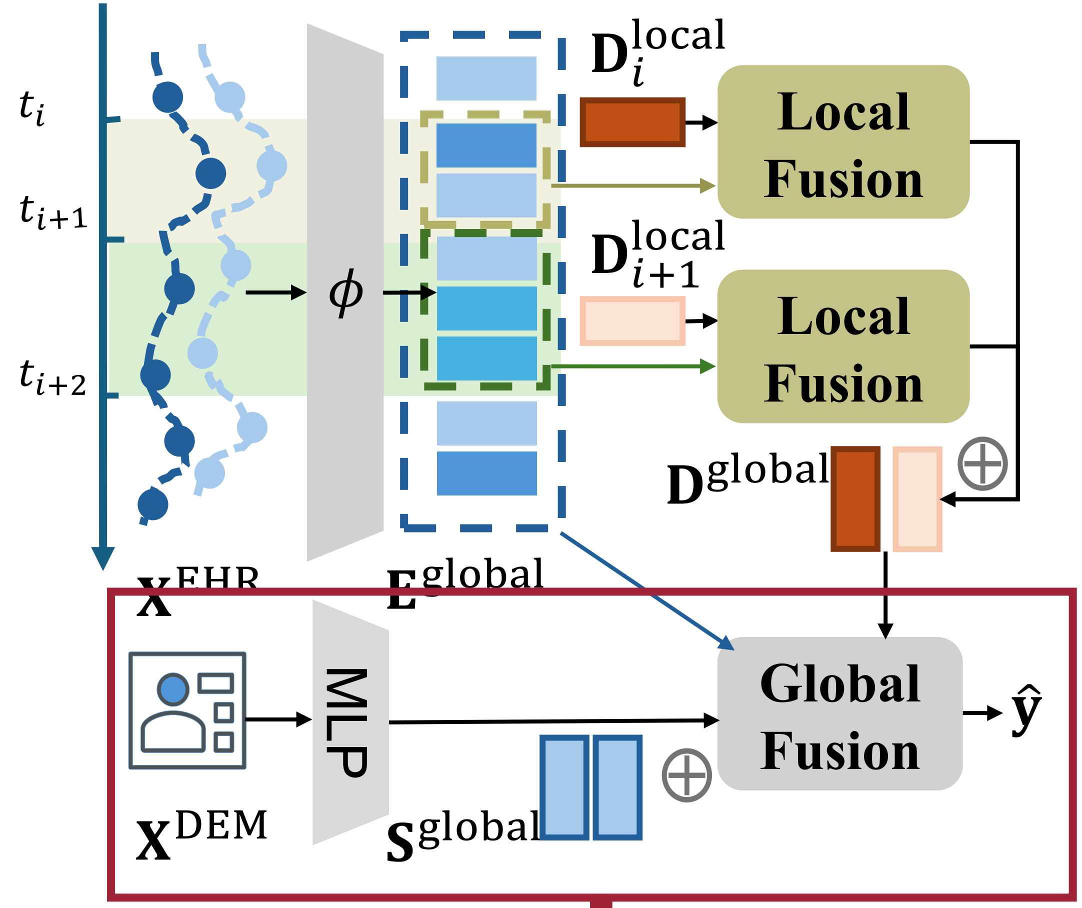
Predicted reversed direction: $\tilde{y}_i^{r,k} = f_k(\widetilde{\mathbf{D}}_i^r)$

Training objective:

$$\mathcal{L}_{\mathrm{PAE}} = \sum_{r=1}^{R} \sum_{k=1}^{K} \Big[\mathrm{CE}(\hat{y}_i^{r,k}, \underbrace{\boxed{y_i^{r,k}}}_{\text{Original label}}) + \mathrm{CE}(\tilde{y}_i^{r,k}, \underbrace{\boxed{-y_i^{r,k}}}_{\text{Reversed label}}) \Big] + \lambda_{\mathrm{static}} \sum_{r=1}^{R} \underbrace{\boxed{\left\| \mathbf{S}_i^r - \tilde{\mathbf{S}}_i^r \right\|_2^2}}_{\text{Static consistency}}$$

Our Solution: Multiscale Multimodal Fusion (MMF)

Goal: Integrate **temporally misaligned** CXR and EHR data via local and global fusion.



Final static fusion and prediction

Local EHR Encoding:

Cross-attention: Interval time embeddings (Query)

& Global EHR features (Key and Value)

$$\mathbf{E}_i^{\text{local}} = \text{softmax}\left(\frac{\mathbf{QK}^\top}{\sqrt{d}} + \boxed{\text{AttnMask}}\right) \cdot \mathbf{V}$$

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

Local CXR-EHR Fusion:

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

Global Hierarchical Fusion:

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

Experiment Results: Disease Progression Identification

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**

- DiPro excels in modeling **disease progression in sequential CXRs**.

  Disentangled temporal features → clearer disease dynamics

  Progression-aware Enhancement → emphasizes progression semantics

- **Adding EHR boosts unimodal DiPro**

  Confirms effective use of complementary EHR features

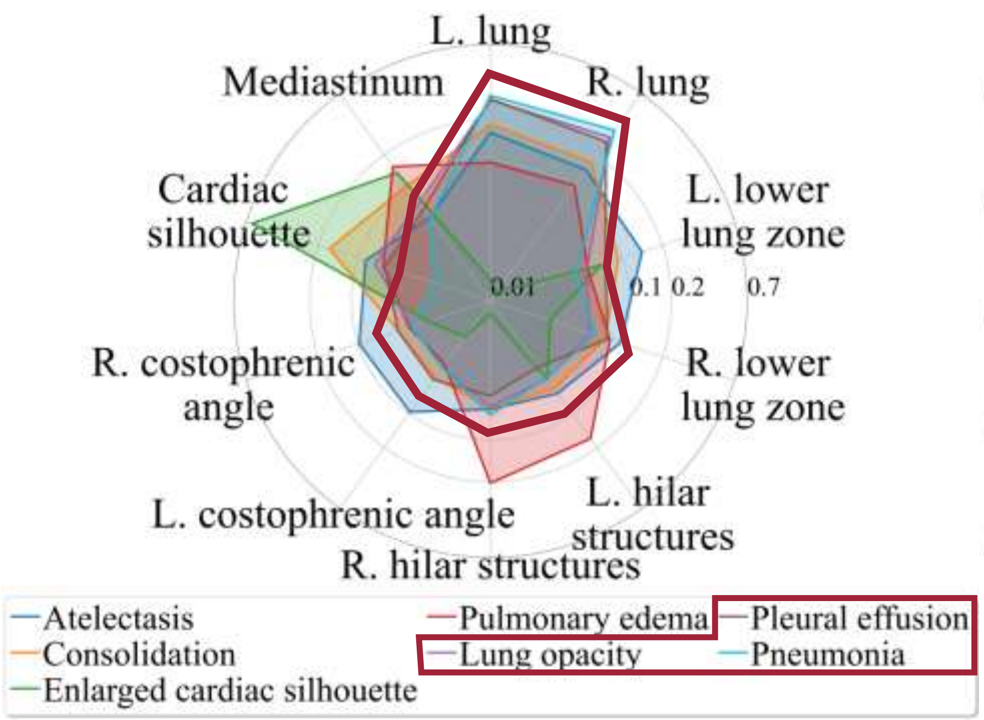
Experiment Results: General ICU Prediction

Method	CXR Used: Last	CXR Used: Long.	Mortality: AUPRC	Mortality: AUROC	Length of Stay: Kappa	Length of Stay: 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**

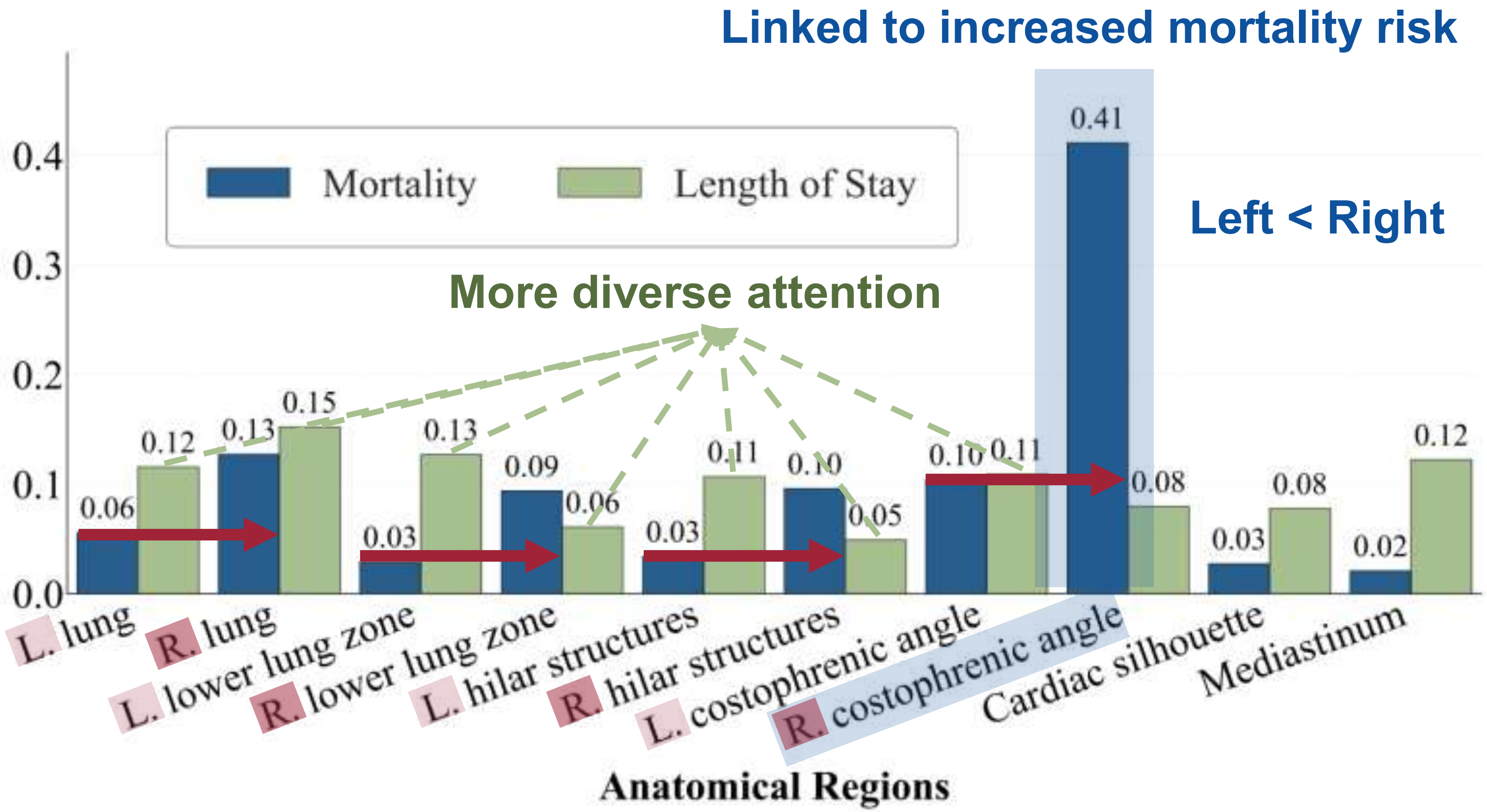
- Existing models experience **performance drop** with **longitudinal CXRs**.
- DiPro **alleviates redundancy and misalignment** in longitudinal CXRs and EHR.

Experiment Results: General ICU Prediction

Averaged attention weights of CXR regions in different tasks:



(a) Disease Progression Identification



(b) General ICU Prediction

Shared pathological regions ➡ **DiPro echoes with clinical knowledge**

Conclusion: Key Takeaways

- **Disentangle** Dynamic from Static Representations:
  - ➡ Mitigate redundancy & improve temporal feature fidelity.
- Incorporate **Progression-Direction Awareness**:
  - ➡ Enhances the model's sensitivity of disease evolution patterns.
- **Multiscale** Fusion of Longitudinal Multimodal Data:
  - ➡ Achieves comprehensive integration across modalities.

Thank you!

Code

Paper

Poster session: Thu 4 Dec 4:30 p.m. — 7:30 p.m.