HGDC-Fuse: Clinical Multi-modal Fusion with Heterogeneous Graph and Disease Correlation Learning for Multi-Disease Prediction

Yueheng Jiang, Peng Zhang*
Zhejiang University, Hangzhou, China
{yuehengj, pengz}@zju.edu.cn

***Abstract*—Accurate multi-disease diagnosis using multi-modal data such as electronic health records and medical imaging is a critical clinical challenge. Although existing deep learning methods have achieved initial success in this area, a significant gap persists for their real-world application. This gap arises because existing methods often overlook unavoidable practical challenges, such as modality missingness, noise, temporal asynchrony, and evidentiary inconsistency across modalities for different diseases. To overcome these limitations, we propose HGDC-Fuse, a novel framework that constructs a patient-centric multi-modal heterogeneous graph to robustly integrate asynchronous and incomplete multi-modal data. Moreover, we design a heterogeneous graph learning module to aggregate multi-source information, featuring a disease correlation-guided attention layer that resolves the modality inconsistency issue by learning disease-specific modality weights based on disease correlations. On the large-scale MIMIC-IV and MIMIC-CXR datasets, HGDC-Fuse significantly outperforms state-of-the-art methods.**

***Index Terms*—multi-modal fusion, multi-disease prediction, heterogeneous graph, disease correlation learning**

I. Introduction

Multi-disease diagnosis is a core task in clinical decision-making, where clinicians integrate complementary evidence from heterogeneous data sources such as electronic health records (EHR) and medical imaging. Recent deep learning methods [1]–[4] have shown promising results in multi-modal disease prediction, yet a substantial gap remains between research performance and real-world clinical deployment. This gap mainly stems from several practical challenges that are insufficiently addressed by existing studies:

Challenge I: Modality Missingness and Noise. In real practice, some modalities are inevitably missing or noisy due to clinical workflows and data collection constraints. For example, chest radiographs (CXR) may be unavailable, while EHR variables are often sparse or irregularly recorded [5]. Late fusion can tolerate missing modalities by combining independent predictions, but it captures limited cross-modal interactions [6]. Recent studies attempt to address this issue through modality generation [7], [8], shared and modality-specific disentanglement [4], [9], [10], or sequence modeling [3]. However, these methods may amplify noise, discard useful signals, or become difficult to extend reliably to more modalities.

Challenge II: Modality Temporal Asynchrony. Clinical modalities are collected on different temporal schedules. EHR is typically recorded densely, whereas CXR is acquired irregularly according to clinical need [11]. Longitudinal CXR sequences may contain important information about disease progression and treatment response [12]. However, many existing multi-modal methods [1]–[4] either concatenate unimodal representations or only use the latest available CXR, without explicitly modeling temporal dependencies across modalities. Consequently, clinically meaningful cross-modal temporal interactions are overlooked.

Challenge III: Modality Inconsistency in Multi-label Prediction. For different diseases, EHR and CXR may provide inconsistent or even conflicting evidence. Estimating disease-specific modality importance is therefore crucial for accurate multi-label prediction. Although prior work [4] applies modality-level attention, it does not explicitly model disease relationships such as co-occurrence and dependency, which are highly relevant in multi-disease diagnosis. In real clinical populations, comorbidities are common. For example, hypertension often co-occurs with diabetes and cardiovascular diseases because of shared metabolic risk factors [13].

To address these challenges, we propose **HGDC-Fuse**: Clinical Multi-modal **Fus**ion with **H**eterogeneous **G**raph and **D**isease **C**orrelation Learning for Multi-Disease Prediction. Our core idea is to leverage graph structures to enable flexible multi-modal fusion and to explicitly model inter-disease dependencies. First, we construct a patient-centric multi-modal heterogeneous graph with two complementary edge types: time-aware cross-modal edges to model temporal asynchrony, and similar-patient edges to enhance representation learning. This graph design is robust to modality missingness and can naturally generalize to other multimodal clinical inputs beyond EHR and CXR. Next, we introduce a type-specific aggregation strategy to preserve source-specific semantics. Building upon the fused patient representations, we further construct a label graph to explicitly capture disease correlations and learn semantic label embeddings. These label embeddings are then used to guide disease-aware cross-modal attention, enabling the model to generate disease-specific modality weights.

*Corresponding author.

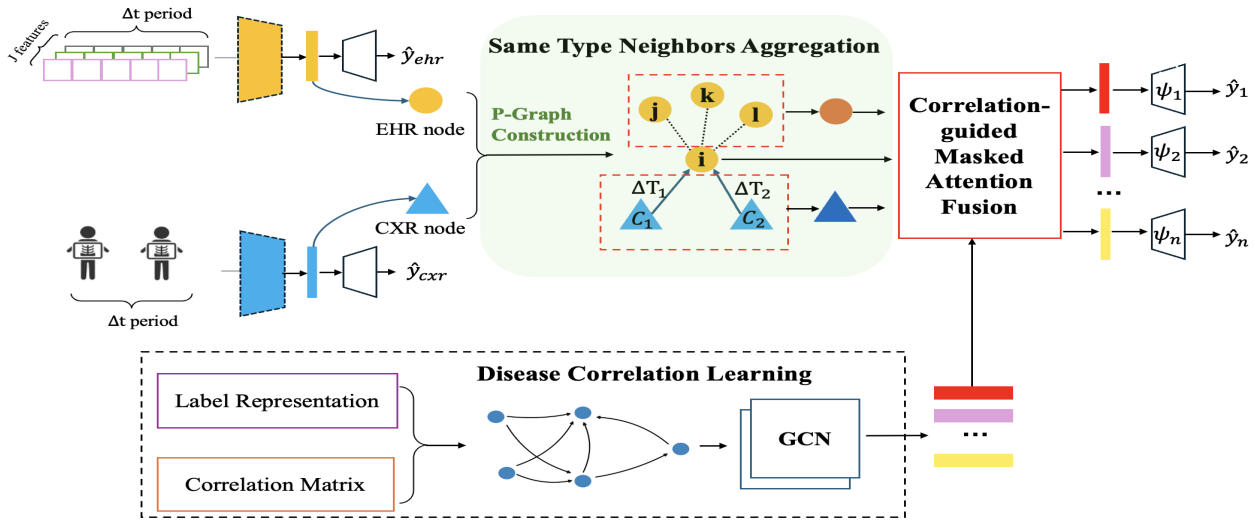

Fig. 1. Overall architecture of our model HGDC-Fuse.

Unlike conventional self-attention mechanisms, our attention explicitly incorporates rich disease relationship priors. As a result, HGDC-Fuse produces more meaningful and discriminative disease-specific representations, thereby addressing modality inconsistency in multi-label disease prediction.

In summary, our main contributions are as follows:

- We propose HGDC-Fuse, a multi-modal heterogeneous graph learning framework that addresses modality temporal asynchrony, missingness, and noise.
- HGDC-Fuse resolves modality inconsistency by leveraging disease correlations to learn disease-specific modality significance.
- We conduct experiments on large-scale real-world datasets and our results demonstrate the superior performance of HGDC-Fuse over state-of-the-art baselines for multi-disease prediction.

II. Preliminary

In this paper, we consider multi-disease prediction based on two modalities: time-series electronic health records (EHR), and chest X-ray images (CXR). Each patient s is associated with a single EHR sequence, while the availability of CXR varies across patients, ranging from none to multiple images. Let $\mathbf{E}_s \in \mathbb{R}^{T_s \times J}$ denote the EHR data of patient s, where T_s is the length of the time series and J is the number of clinical features. Let C_s^k denote the k-th CXR image of patient s, and let $\mathbf{C}_s = \{C_s^1, \ldots, C_s^K\}$ be the set of all CXRs associated with patient s, where K is the number of available images.

III. Method

A. Overview

The architecture of HGDC-Fuse is presented in Figure 1. The framework consists of three main components: (1) *Heterogeneous Patient Graph Construction*, which encodes modality-specific features using a Transformer for EHR and ResNet-50 for CXR sequences, and integrates them into a patient-centric graph using time-aware and similarity-based edges to handle asynchrony and missingness; (2) *Same Type Neighbors Aggregation*, which extracts messages from similar patient neighbors and longitudinal image history to compensate for noise; and (3) *Disease Correlation-guided Attention Fusion*, which resolves modality inconsistency by utilizing a GCN to model disease co-occurrence patterns. These learned semantic prototypes guide a cross-modal attention mechanism to dynamically assign disease-specific weights to EHR and CXR features. Finally, the resulting representations are fed into a multi-label classification layer for joint optimization.

B. Heterogeneous Patient Graph Structure Construction

Modality-specific Encoders. Given that EHR and CXR represent highly heterogeneous modalities, we employ specialized encoders to process each raw data source independently. Specifically, we use a Transformer [14] encoder to obtain the EHR representation $\mathbf{h}_s^{ehr}$, and ResNet-50 [15] to extract visual features $\mathbf{h}_s^{cxr,k}$ from each CXR image C_s^k.

Graph Topology and Edge Definitions. For each patient s, we construct a patient-centric heterogeneous graph $\mathcal{G}_s = (\mathcal{V}_s, \mathcal{E}_s)$, where the node features are the representations $\mathbf{h}_s^{ehr}$ and $\mathbf{h}_s^{cxr,k}$ directly obtained from the modality-specific encoders. Specifically, the node set $\mathcal{V}_s$ is dynamically defined based on modality availability: it contains a target EHR node n_s^E, some similar EHR nodes $n_{s'}^E$ and a set of intra-patient CXR nodes $\mathcal{V}_s^{cxr} = \{n_s^{C,1}, \ldots, n_s^{C,K}\}$ only when imaging data is present. When the CXR modality is missing, $\mathcal{V}_s^{cxr} = \emptyset$.

In the P-Graph, the EHR node n_s^E serves as the target node, connecting to two types of neighbors:

- Intra-patient CXR neighbors (n_s^C): As depicted in Fig. 2, these nodes only exist when the CXR modality is available. These nodes are connected to n_s^E via directed cross-modal edges. The edge weights encode the relative acquisition times $\Delta T(n_s^C)$, defined as the duration from admission to the CXR scan. A larger ΔT indicates a more

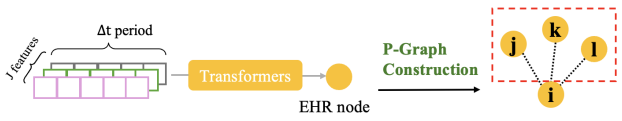

Fig. 2. Heterogeneous patient graph construction when the CXR modality is missing.

recent observation. The cross-modal CXR → EHR edges are formally defined as:

$$\mathcal{E}_s^{cxr\to ehr} = \left\{(n_i, v_i, \Delta T(n_i)) \mid n_i \in \mathcal{V}s^{cxr}, v_i = n_s^E\right\} \tag{1}$$

- Inter-patient EHR neighbors ($n_{s'}^E$): These are constructed by selecting similar but distinct patients s' within the same batch, based on the cosine similarity between their EHR embeddings $\mathbf{h}^{ehr}$. The undirected and unweighted inter-patient EHR–EHR edges are defined as:

$$\mathcal{E}_s^{ehr\text{-}ehr} = \left\{(n_s^E, n_{s'}^E) \mid \cos(\mathbf{h}_s^{ehr}, \mathbf{h}_{s'}^{ehr}) > \delta\right\} \tag{2}$$

where δ is the similarity threshold for edge creation.

C. Same Type Neighbors Aggregation

Within the P-Graph, each EHR node n_s^E aggregates information from two heterogeneous neighbor sets: inter-patient EHR nodes and intra-patient CXR nodes. To preserve domain-specific semantics, we devise a type-specific aggregation strategy to obtain two distinct message vectors: $\mathbf{m}_s^{E\leftarrow E}$ and $\mathbf{m}_s^{E\leftarrow C}$.

The message $\mathbf{m}_s^{E\leftarrow E}$ is computed by a multi-head attention mechanism over all neighbor EHR nodes $\mathcal{N}^E(n_s^E)$:

$$\mathbf{m}_s^{E\leftarrow E} = \Big\|_{i=1}^{H} \sum_{n_{s'}^E \in \mathcal{N}^E(s)} \alpha_{ss'}^{(i)} \mathbf{W}_E^{(i)} \mathbf{h}_{s'}^{ehr} \tag{3}$$

The attention weights $\alpha_{ss'}^{(i)}$ are obtained by:

$$e_{ss'}^{(i)} = \text{LeakyReLU}\left(\mathbf{a}^{(i)\top}\left[\mathbf{W}_E^{(i)}\mathbf{h}_s^{ehr} \,\|\, \mathbf{W}_E^{(i)}\mathbf{h}_{s'}^{ehr}\right]\right) \tag{4}$$

$$\alpha_{ss'}^{(i)} = \frac{\exp(e_{ss'}^{(i)})}{\sum_{n_{s''}^E \in \mathcal{N}^E(s)} \exp(e_{ss''}^{(i)})} \tag{5}$$

Here, $\mathbf{a}^{(i)}$ is a learnable attention vector, and $\mathbf{W}_E^{(i)}$ is a trainable linear projection matrix for the i-th attention head.

For intra-patient CXR nodes n_s^C, we implement a time-aware attention mechanism to resolve temporal asynchrony. The normalized time weight for the j-th CXR of patient s is computed as:

$$w_j^{(s)} = \frac{\exp(\Delta T(n_s^{C,j}))}{\sum_{k=1}^{K} \exp(\Delta T(n_s^{C,k}))} \tag{6}$$

Then, the cross-modal message $\mathbf{m}_s^{E\leftarrow C}$ is derived as a time-weighted sum:

$$\mathbf{m}_s^{E\leftarrow C} = \sum_{j=1}^{K} w_j^{(s)} \cdot \mathbf{W}_C \mathbf{h}_s^{cxr,j} \tag{7}$$

where $\mathbf{W}_C$ is a learnable transformation matrix. By assigning higher weights to images with larger ΔT, this mechanism prioritizes recent clinical evidence by incorporating temporal proximity into the aggregation process, effectively capturing the clinical evolution reflected in the CXR sequence.

D. Disease Correlation Learning

Inspired by the fact that physicians often consider disease co-occurrence patterns in multi-disease diagnoses [16], we aim to model the complex inter-relationships among diseases to improve clinical prediction. Drawing inspiration from ML-GCN [17], we construct a disease correlation graph where each node corresponds to a disease class.

Each label is represented by a one-hot word embedding vector, forming an initial matrix $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_N]^\top \in \mathbb{R}^{N\times d}$, where N is the number of disease labels and d is the embedding dimension. We compute the correlation matrix $\mathbf{A} \in \mathbb{R}^{N\times N}$ using conditional co-occurrence statistics extracted from the training set. Specifically, the element A_{ij} is defined as the conditional probability that label j occurs given label i:

$$A_{ij} = \frac{\text{co-occur}(i,j)}{\text{count}(i)}, \quad i \neq j \tag{8}$$

where co-occur(i,j) denotes the number of samples where both labels i and j appear, and count(i) is the number of samples annotated with label i. To remove noisy correlations, a threshold τ is applied:

$$A_{ij} = \begin{cases} 1, & \text{if } A_{ij} >= \tau \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

We apply a two-layer Graph Convolutional Network (GCN) [18] to update the label embeddings:

$$\tilde{\mathbf{Z}} = \text{GCN}(\widehat{\boldsymbol{A}}, \mathbf{Z}) \tag{10}$$

where $\widehat{\boldsymbol{A}}$ is the normalized version of correlation matrix $\mathbf{A}$. The output $\tilde{\mathbf{Z}} \in \mathbb{R}^{N\times d'}$ serves as disease-aware prototypes that encode higher-order co-occurrence semantics and will later guide disease-specific feature fusion.

E. Disease Correlation-guided Attention Fusion

We propose a Disease Correlation-guided Attention Fusion (CGA) module to adaptively fuse multi-source information for each disease.

Each target node n_s^E has three types of latent features: $\mathbf{h}_s^{\text{ehr}} \in \mathbb{R}^d$, $\mathbf{m}_s^{E\leftarrow E} \in \mathbb{R}^d$, and $\mathbf{m}_s^{E\leftarrow C} \in \mathbb{R}^d$. These are stacked as:

$$\mathbf{T}_s = [\mathbf{h}_s^{\text{ehr}},\ \mathbf{m}_s^{E\leftarrow E},\ \mathbf{m}_s^{E\leftarrow C}] \in \mathbb{R}^{3\times d} \tag{11}$$

We aim to obtain a label-specific fused representation for each disease k by using the disease label embedding $\mathbf{z}_n \in \mathbb{R}^d$ as the query vector. Specifically, we project the features into the key and value spaces:

$$\mathbf{q_n} = \mathbf{W}_q \mathbf{z}_n, \quad \mathbf{K} = \mathbf{W}_k \mathbf{T}_s, \quad \mathbf{V} = \mathbf{W}_v \mathbf{T}_s \tag{12}$$

where $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v \in \mathbb{R}^{d\times d}$ are learnable projection matrices. The attention weights are computed via scaled dot-product:

$$\boldsymbol{\alpha}_n = \mathrm{softmax}\left(\frac{\mathbf{K}\mathbf{q}_n^\top}{\sqrt{d}} + \mathbf{M}_s\right). \tag{13}$$

where mask $\mathbf{M}_s \in \{0, -\infty\}^3$, $M_{s,3} = -\infty$ when CXR is missing. The final disease-specific representation $\tilde{\mathbf{h}}_n$ is given as:

$$\tilde{\mathbf{h}}_n = \boldsymbol{\alpha}_n^\top \mathbf{V} \in \mathbb{R}^d. \tag{14}$$

F. Prediction and Optimization

After obtaining the final representations $\tilde{\mathbf{h}}_n$, the final prediction for the n^{th} disease can be obtained using a feedforward layer: $\hat{y}_n = \psi_n(\tilde{\mathbf{h}}_n)$. We optimize HGDC-Fuse using a cross-entropy loss as:

$$\mathcal{L} = \sum_{n=1}^{N} y_n \log(\hat{y}_n) + (1 - y_n)\log(1 - \hat{y}_n). \tag{15}$$

where N is the number of prediction classes.

IV. Experiments

A. Datasets and preprocessing

We utilize two large-scale real-world datasets, MIMIC-IV [19] and MIMIC-CXR [20], to evaluate the performance of `HGDC-Fuse`. MIMIC-IV contains de-identified data of adult patients admitted to the intensive care units (ICU) of Beth Israel Deaconess Medical Center (BIDMC), while MIMIC-CXR provides the corresponding chest radiographs. Following prior work [21], we extract 17 routinely monitored clinical variables from the EHR, comprising five categorical and twelve continuous variables. Our primary task is multi-disease prediction, where 25 disease phenotype labels are generated based on diagnosis codes.

To align with clinical needs for early prediction, we utilize data within the first 48 hours of ICU admission. While existing methods retrieve only the last available CXR within this observation window, `HGDC-Fuse` incorporates all available CXRs captured during the same period to better capture disease evolution through sequential imaging. In total, we identified 59,344 ICU stays with valid EHR records, of which 10,630 stays are associated with at least one CXR image, averaging 1.89 images per stay.

We construct two experimental settings: a full dataset including all patient stays regardless of CXR availability, and a matched subset containing only stays with both EHR and CXR data. Both are split into training, validation, and testing sets using a 7:1:2 ratio. Notably, following [4], when training with the matched subset, we randomly remove the CXR modality for 30% of the samples within each mini-batch as an additional data augmentation to improve model robustness.

B. Experimental Implementation and Evaluation Metrics

The proposed framework was implemented in PyTorch 2.5.1 and trained on an NVIDIA GeForce RTX 4090 GPU. We utilized a grid search strategy to optimize hyperparameters based on validation set performance, subsequently reporting final results on the test set. Specifically, the search spaces for the similarity threshold δ and the correlation threshold τ were defined as $\{0, 0.1, 0.2, \ldots, 0.9, 1\}$. Following this exhaustive search, the optimal values were determined to be $\delta = 0.6$ and $\tau = 0.4$. Given the highly imbalanced distribution of disease labels in the MIMIC datasets, we employed the Area Under the Precision-Recall Curve (PRAUC) as the primary evaluation metric to assess the performance of `HGDC-Fuse` and all baseline models.

C. Baselines

We compare the following baselines: Transformer [14] is a uni-modal baseline utilizing only EHR data; MMTM [1] and DAFT [2] are frameworks assuming full modality availability, where we compensate for missing CXR modalities with all-zero tensors. Specifically, MMTM uses squeeze-and-excitation blocks for feature calibration, while DAFT employs dynamic affine transformations to integrate EHR features into the imaging branch; MedFuse [3] is an LSTM-based clinical fusion model designed for asynchronous EHR and CXR data, and MedFuse-II is its variant using ResNet-50 and Transformer encoders to ensure a fair comparison; finally, DrFuse [4] is a **state-of-the-art** framework that addresses missing modalities and modality inconsistency by learning disentangled shared and specific representations. DrFuse utilizes the same ResNet-50 and Transformer backbones as our proposed model to maintain consistency in feature extraction capabilities.

D. Overall Performance of Multi-Disease Prediction

Performance on Matched Subset. When trained and tested on the matched subset, `HGDC-Fuse` achieves a relative improvement of 4.4% over DrFuse, and a significant gain of 15.2% compared to the uni-modal Transformer. These results demonstrate that our graph-based approach effectively fuses modalities by capturing cross-modal correlations and temporal dependencies within the CXR sequence.

Robustness under Modality Missingness. In the more challenging full dataset training setting, where the majority of samples lack CXR data, `HGDC-Fuse` exhibits superior

TABLE I
OVERALL PERFORMANCE MEASURED BY THE MACRO AVERAGE OF PRAUC OVER ALL 25 DISEASE LABELS. NUMBERS IN **BOLD** INDICATE THE BEST PERFORMANCE IN EACH COLUMN.

Model	Trained with the *matched* subset		Trained with the *full* dataset	
	test on matched	*test on full*	*test on matched*	*test on full*
Transformer	0.408	0.374	0.435	0.418
MMTM	0.416	0.359	0.422	0.407
DAFT	0.417	0.348	0.430	0.409
MedFuse	0.427	0.329	0.434	0.405
MedFuse-II	0.418	0.329	0.427	0.412
DrFuse	0.450	0.384	0.470	0.419
HGDC-Fuse	**0.470**	**0.386**	**0.489**	**0.434**

TABLE II

PER-DISEASE PRAUC PERFORMANCE AND COMPARISON. MODELS ARE EVALUATED ON THE MATCHED SUBSET. PERCENTAGES IN PARENTHESES DENOTE RELATIVE IMPROVEMENT OVER THE BEST UNI-MODAL BASELINE. IMPROVEMENTS $> 5\%$ ARE **BOLDED**, WHILE DECREASES $> 5\%$ ARE UNDERLINED.

Disease Label	CXR (ResNet50)	EHR (Transformer)	DrFuse	HGDC-Fuse
Acute and unspecified renal failure	0.4854	0.5129	**0.5533 (+7.9%)**	**0.5533 (+7.9%)**
Acute cerebrovascular disease	0.1486	0.3976	**0.4532 (+14.0%)**	**0.4905 (+23.4%)**
Acute myocardial infarction	0.1610	0.1438	**0.1756 (+9.1%)**	**0.1972 (+22.5%)**
Cardiac dysrhythmias	0.5777	0.4601	0.5222 (-9.6%)	0.5966 (+3.3%)
Chronic kidney disease	0.4191	0.4485	0.4106 (-8.5%)	**0.4850 (+8.1%)**
Chronic obstructive pulmonary disease	0.3786	0.2166	0.2709 (-28.4%)	**0.3979 (+5.1%)**
Complications of surgical/medical care	0.3189	0.3679	**0.4110 (+11.7%)**	**0.4033 (+9.6%)**
Conduction disorders	0.6116	0.1836	0.2178 (-64.4%)	0.6390 (+4.5%)
Congestive heart failure	0.6045	0.4984	0.5420 (-10.3%)	**0.6648 (+10.0%)**
Coronary atherosclerosis	0.6471	0.5617	0.5606 (-13.4%)	0.6709 (+3.7%)
Diabetes mellitus with complications	0.1823	0.5054	0.5202 (+2.9%)	0.5259 (+4.1%)
Diabetes mellitus without complication	0.2987	0.3542	**0.3767 (+6.4%)**	**0.4006 (+13.1%)**
Disorders of lipid metabolism	0.5946	0.6097	0.5862 (-3.9%)	0.5974 (-2.0%)
Essential hypertension	0.5510	0.5734	0.5790 (+1.0%)	0.5983 (+4.3%)
Fluid and electrolyte disorders	0.5950	0.6602	0.6638 (+0.5%)	0.6867 (+4.0%)
Gastrointestinal hemorrhage	0.1453	0.1069	**0.1817 (+25.0%)**	**0.2241 (+54.2%)**
Hypertension with complications	0.3992	0.4264	0.3750 (-12.1%)	**0.4631 (+8.6%)**
Other liver diseases	0.3601	0.2445	0.2979 (-17.3%)	**0.4050 (+12.5%)**
Other lower respiratory disease	0.1759	0.1579	0.1719 (-2.3%)	0.1807 (+2.7%)
Other upper respiratory disease	0.0998	0.1115	**0.1346 (+20.7%)**	**0.1732 (+55.3%)**
Pleurisy; pneumothorax	0.1826	0.1241	0.1698 (-7.0%)	**0.2286 (+25.2%)**
Pneumonia	0.3741	0.3671	**0.4092 (+9.4%)**	**0.4457 (+19.1%)**
Respiratory failure; arrest (adult)	0.5213	0.5855	0.5964 (+1.9%)	0.6146 (+5.0%)
Septicemia (except in labor)	0.3974	0.5077	0.5314 (+4.7%)	**0.5464 (+7.6%)**
Shock	0.3875	0.5334	0.5524 (+3.6%)	0.5570 (+4.4%)
Average Rank	3.48	3.12	2.36	**1.04**

robustness. HGDC-Fuse outperformed DrFuse by 3.6% on the full test set and 4.0% on the matched test set. Notably, all multi-modal baselines underperform the uni-modal Transformer when tested on the full dataset, except for DrFuse, which only marginally exceeds Transformer. In contrast, HGDC-Fuse maintains a substantial margin of 3.8% over the Transformer. This resilience is attributed to our model's adaptive graph topology and population-level priors, which effectively leverage partially missing training information.

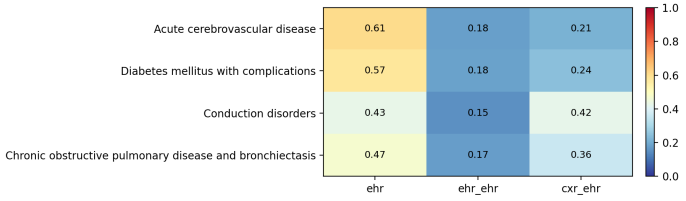

Fig. 3. Visualization of disease-specific modal significance. The heatmap displays the mean attention weights assigned by the CGA module across positive samples for selected disease labels.

E. Disease-Wise Prediction Performance

Table II details the disease-wise PRAUC scores for uni-modal methods, DrFuse, and HGDC-Fuse, where values in parentheses denote the relative difference against the best uni-modal baseline. The results underscore the prevalence of modality inconsistency, where EHR and CXR data contribute disparately across different phenotypes. Specifically, *Acute cerebrovascular disease* and *Diabetes mellitus with complications* are predominantly driven by EHR data, while *Conduction disorders* and *Chronic obstructive pulmonary disease and bronchiectasis* exhibit a significantly higher reliance on CXR imaging.

While DrFuse often suffers from performance degradation—dropping over 5% on 10 out of 25 labels compared to the best uni-modal baseline—HGDC-Fuse consistently achieves superior or comparable results across nearly all categories. On the EHR-dominant *Acute cerebrovascular disease*, our model outperforms DrFuse by 8.2%, while on the CXR-reliant *Conduction disorders*, where DrFuse drops significantly (-64.4%), HGDC-Fuse maintains a 4.5% gain over the best uni-modal baseline. Notably, for *other upper respiratory disease* and *gastrointestinal hemorrhage*, our framework yields substantial improvements of 55.3% and 54.2%, respectively.

To further investigate how HGDC-Fuse adaptively resolves modality inconsistency, we visualize the mean attention weights of the Correlation-Guided Attention (CGA) module for positive samples across these four representative diseases (Fig. 3). The attention mechanism successfully aligns with clinical intuition: for EHR-dominant diseases, the model as-

TABLE III
RESULTS OF THE ABLATION STUDY

Model	PRAUC @matched subset	PRAUC @full dataset
w/o EHR-EHR	0.4500	0.3812
w/o multi-cxr	0.4506	0.3698
w/o CGA	0.4548	0.3760
HGDC-Fuse	0.4698	0.3860

signs significantly higher weights to the EHR modality (up to 0.61) with minimal cross-modal involvement. In contrast, for CXR-dependent diseases such as *Conduction disorders*, the cross-modal attention weight (CXR $\rightarrow$ EHR) increases substantially to 0.42, matching the EHR weight. These results demonstrate that `HGDC-Fuse` successfully leverages label correlations to infer disease-specific modal significance, effectively mitigating the negative impact of conflicting modalities through dynamic weight redistribution.

F. Ablation Study

To evaluate the contribution of each component within `HGDC-Fuse`, we conducted an ablation study using the matched subset training setting, with results summarized in Table III. We designed three variants:

- w/o EHR-EHR, which removes inter-patient neighbors to assess the impact of population-level priors.
- w/o multi-CXR, which utilizes only the last available radiograph, thereby discarding the temporal evolution of imaging.
- w/o CGA, which replaces the correlation-guided attention with a standard self-attention mechanism.

The performance decline observed in the first two variants confirms that our heterogeneous graph effectively handles modality temporal asynchrony, missingness, and noise. Furthermore, the superiority of the full model over the w/o CGA variant indicates that the label correlation-guided mechanism is crucial, as it introduces clinical prior knowledge to resolve modal conflicts more effectively than general self-attention.

V. CONCLUSION

In this paper, we proposed HGDC-Fuse, a novel framework that utilizes heterogeneous graph learning to address critical real-world challenges of modality missingness and temporal asynchrony, while explicitly capturing disease correlations to tackle modality inconsistency in clinical multi-disease prediction. Extensive experiments demonstrate that HGDC-Fuse significantly outperforms state-of-the-art methods. Our work highlights the potential of heterogeneous graphs and disease correlation modeling for developing more robust and reliable multi-modal diagnostic systems for clinical application.

REFERENCES

[1] H. R. V. Joze, A. Shaban, M. L. Iuzzolino, and K. Koishida, "Mmtm: Multimodal transfer module for cnn fusion," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 13 289–13 299.

[2] S. Pölsterl, T. N. Wolf, and C. Wachinger, "Combining 3d image and tabular data via the dynamic affine feature map transform," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2021, pp. 688–698.

[3] N. Hayat, K. J. Geras, and F. E. Shamout, "Medfuse: Multi-modal fusion with clinical time-series data and chest x-ray images," in *Machine Learning for Healthcare Conference*. PMLR, 2022, pp. 479–503.

[4] W. Yao, K. Yin, W. K. Cheung, J. Liu, and J. 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*, vol. 38, no. 15, 2024, pp. 16 416–16 424.

[5] A. R. Miller and C. Tucker, "Privacy protection and technology diffusion: The case of electronic medical records," *Management science*, vol. 55, no. 7, pp. 1077–1093, 2009.

[6] S.-C. Huang, A. Pareek, S. Seyyedi, I. Banerjee, and M. P. Lungren, "Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines," *NPJ digital medicine*, vol. 3, no. 1, p. 136, 2020.

[7] M. Ma, J. Ren, L. Zhao, S. Tulyakov, C. Wu, and X. Peng, "Smil: Multimodal learning with severely missing modality," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 3, 2021, pp. 2302–2310.

[8] A. Sharma and G. Hamarneh, "Missing mri pulse sequence synthesis using multi-modal generative adversarial network," *IEEE transactions on medical imaging*, vol. 39, no. 4, pp. 1170–1183, 2019.

[9] H. Wang, Y. Chen, C. Ma, J. Avery, L. Hull, and G. 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*, 2023, pp. 15 878–15 887.

[10] C. Chen, Q. Dou, Y. Jin, H. Chen, J. Qin, and P.-A. Heng, "Robust multimodal brain tumor segmentation via feature disentanglement and gated fusion," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2019, pp. 447–456.

[11] A. Al Shahrani and K. Al-Surimi, "Daily routine versus on-demand chest radiograph policy and practice in adult icu patients-clinicians' perspective," *BMC Medical Imaging*, vol. 18, no. 1, p. 4, 2018.

[12] A. Borghesi and R. Maroldi, "Covid-19 outbreak in italy: experimental chest x-ray scoring system for quantifying and monitoring disease progression," *La radiologia medica*, vol. 125, no. 5, pp. 509–513, 2020.

[13] A. N. Long and S. Dagogo-Jack, "Comorbidities of diabetes and hypertension: mechanisms and approach to target organ protection," *The journal of clinical hypertension*, vol. 13, no. 4, pp. 244–251, 2011.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[16] K. Barnett, S. W. Mercer, M. Norbury, G. Watt, S. Wyke, and B. Guthrie, "Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study," *The Lancet*, vol. 380, no. 9836, pp. 37–43, 2012.

[17] Z.-M. Chen, X.-S. Wei, P. Wang, and Y. Guo, "Multi-label image recognition with graph convolutional networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5177–5186.

[18] A. Mazari and H. Sahbi, "Mlgcn: Multi-laplacian graph convolutional networks for human action recognition," in *The British machine vision conference (BMVC)*, 2019.

[19] A. E. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, and S. Horng, "Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports," *Scientific data*, vol. 6, no. 1, p. 317, 2019.

[20] A. E. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T. J. Pollard, S. Hao, B. Moody, B. Gow *et al.*, "Mimic-iv, a freely accessible electronic health record dataset," *Scientific data*, vol. 10, no. 1, p. 1, 2023.

[21] H. Harutyunyan, H. Khachatrian, D. C. Kale, G. Ver Steeg, and A. Galstyan, "Multitask learning and benchmarking with clinical time series data," *Scientific data*, vol. 6, no. 1, p. 96, 2019.