doi: DOI HERE
Advance Access Publication Date: Day Month Year
Paper

PAPER

Graph-informed and FiLM-enhanced Multimodal Fusion for Myocardial Infarction Prediction

Xiantong Xiang,[1,*] Second Author,[2] Third Author,[3] Fourth Author[3] and Fifth Author[4]

[1]Department, Organization, Street, Postcode, State, Country, [2]Department, Organization, Street, Postcode, State, Country, [3]Department, Organization, Street, Postcode, State, Country and [4]Department, Organization, Street, Postcode, State, Country
[*]Corresponding author. email-id.com

FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

Abstract

Accurate and timely diagnosis of cardiovascular diseases, particularly myocardial infarction (MI), remains a critical clinical challenge. Existing electrocardiogram (ECG) analysis methods often rely solely on a single data modality, such as raw signals or waveform images, which limits their ability to capture the broader physiological context. To address this limitation, we propose GFM-MIP, a Graph-informed and FiLM-enhanced Multimodal Fusion framework for myocardial infarction prediction. GFM-MIP integrates 12-lead ECG time-series signals, ECG images, and laboratory test results through a unified architecture. Specifically, it employs a Graphormer encoder to model inter-lead dependencies in ECG signals and a Vision Transformer to extract morphological patterns from ECG images, both modulated by patient-specific laboratory features using Feature-wise Linear Modulation (FiLM). A Transformer-based fusion module captures cross-modal interactions, while a contrastive learning objective encourages alignment between signal and image modalities. Experimental results on a real-world clinical dataset and three public benchmarks demonstrate that GFM-MIP consistently outperforms state-of-the-art baselines across multiple evaluation metrics. Ablation studies further validate the contribution of each modality and architectural component. The proposed framework offers a clinically meaningful and scalable solution for robust, multimodal cardiovascular diagnosis.

Key words: ECG, Multimodal Learning, Cross-modal Fusion, Myocardial Infarction Diagnosis

Introduction

Myocardial infarction (MI), a severe manifestation of cardiovascular disease (CVD), remains a major cause of morbidity and mortality worldwide. Prompt and accurate diagnosis of MI is critical to improving patient survival rates, reducing complications, and guiding personalized treatment. Electrocardiography (ECG), particularly the 12-lead ECG, serves as a frontline diagnostic tool due to its low cost, noninvasiveness, and ease of acquisition in emergency settings [1–3]. However, interpreting ECG data is inherently challenging, as it involves complex morphological patterns, variable signal dynamics, and lead-specific dependencies that are not always obvious to automated systems or even experienced clinicians.

Conventional approaches for automated MI diagnosis [4–12] typically rely on a single data modality, such as ECG time-series [13] or derived ECG images [14], and employ either convolutional or recurrent neural networks for feature extraction. While these methods have shown promise, they face fundamental limitations. First, unimodal inputs fail to reflect the full physiological context necessary for robust decision-making, especially in complex clinical cases. Second, existing models often treat ECG leads as independent or flat sequences, disregarding their anatomical organization and inter-lead correlations. Moreover, most methods overlook laboratory biomarkers, such as cardiac enzymes or inflammatory markers, which carry vital biochemical information that complements ECG-based patterns [13].

前面部分感觉有点短，看是否能扩展一下

These limitations highlight the need for a more comprehensive and structured approach to MI diagnosis. In practice, cardiologists often consider multiple sources of evidence—including ECG morphology, temporal signal patterns, and laboratory test results—when making diagnostic decisions. Designing a computational model that mimics this multimodal reasoning process requires tackling several challenges: how to jointly model structured dependencies in temporal signals, how to incorporate patient-specific physiological information, and how to integrate heterogeneous modalities in a unified and coherent learning framework [15, 16].

In this work, we propose a unified framework to address these challenges by integrating multimodal physiological data through structured encoding and semantic alignment. We propose GFM-MIP (Graph-informed and FiLM-enhanced Multimodal Fusion for Myocardial Infarction Prediction), a novel deep learning framework that integrates three complementary modalities: ECG time-series, ECG images, and laboratory test results. Specifically, GFM-MIP employs a Graphormer-based encoder [17] to capture the relational structure among ECG leads, and a Vision Transformer to extract morphological features from ECG images. Patient-specific physiological context is injected into both branches via Feature-wise Linear Modulation (FiLM), enabling dynamic modulation of feature extraction. A Transformer-based fusion module aggregates the cross-modal representations, while a symmetric contrastive learning objective aligns time-series and image modalities in the shared latent space. The entire framework is jointly optimized using a hybrid loss that balances supervised classification and contrastive alignment.

Our main contributions are summarised as follows:

- We propose a unified multimodal framework for MI prediction that integrates ECG signals, ECG images, and laboratory biomarkers.
- We introduce a Graphormer-based temporal encoder to model the spatial and temporal dependencies among ECG leads.
- We leverage FiLM-based modulation to inject patient-specific physiological information into both temporal and spatial branches.
- We design a contrastive learning objective to align heterogeneous modalities and improve multimodal consistency.
- We conduct extensive experiments on real-world clinical datasets and public benchmarks, demonstrating that our method achieves state-of-the-art performance across multiple evaluation metrics.

Through these innovations, GFM-MIP addresses key limitations of traditional single- modality ECG approaches, providing a robust, interpretable, and clinically valuable solution for myocardial infarction diagnosis.

Related Work

ECG-Based Deep Learning Methods

Automated ECG interpretation has been a long-standing research focus in computational cardiology. Early studies primarily relied on handcrafted features and traditional classifiers. With the rise of deep learning, CNN-based and RNN-based architectures became dominant due to their ability to capture local patterns and sequential dynamics from raw ECG signals. Liu et al. [18] proposed a CNN model for detecting acute myocardial infarction (MI), showing improved diagnostic performance over conventional rule-based systems. Jafarian et al. [19] explored both shallow and deep neural networks to localize infarct regions based on signal morphology, while Lee et al. [20] utilized deep learning to detect low ejection fraction from ECG signals.

More recently, transformer-based models have been introduced for ECG analysis due to their superiority in modeling long-range dependencies. Nie et al. [21] demonstrated that temporal self-attention can enhance ECG forecasting performance in long sequences. However, most existing methods are unimodal, focusing exclusively on time-series signals or 2D ECG images. Such approaches overlook the complementary value of biochemical indicators and lack the ability to capture diverse physiological cues, limiting their diagnostic robustness in clinical settings [22, 23].

Graph Neural Networks and Transformers in Biomedical Applications

Graph-based learning has become increasingly prominent in medical AI due to its strength in modeling structured and relational data. In ECG applications, the 12-lead signal system can be naturally represented as a graph, where each node corresponds to a lead, and edges reflect anatomical or physiological connectivity. Backhaus et al. [24] employed graph models to analyze myocardial strain, emphasizing the importance of spatial dependencies. Similarly, Alkhodair et al. (2022) leveraged GNNs to detect cardiac abnormalities by modeling spatial relations between ECG leads, significantly improving classification interpretability.

In parallel, Vision Transformers (ViTs) have shown competitive results in medical image analysis, outperforming CNNs in capturing global morphological patterns. Kilimci et al. [14] demonstrated the effectiveness of ViTs in ECG image classification. ViT variants have also been successfully applied in other domains such as pathology and ophthalmology, highlighting their broad applicability in medical vision tasks.

Moreover, hybrid architectures that combine graph neural networks and transformer modules have been explored in multi-organ segmentation [25], cancer subtype classification [26], and neuroimaging analysis [27]. These works support the premise that combining relational inductive biases from GNNs with global attention from transformers can yield more expressive biomedical representations. Yet, to our knowledge, few studies have jointly applied GNNs and transformers to the modeling of ECG data across both temporal and morphological dimensions within a unified diagnostic framework.

Multimodal Fusion and Contrastive Learning in Clinical Modeling

Multimodal fusion has emerged as a powerful approach to enhance diagnostic accuracy by leveraging diverse patient information [22, 28–30]. Several works combine ECG with clinical metadata or laboratory values to improve prediction. Al-Zaiti et al. [31, 32] demonstrated that integrating ECG signals with structured clinical features significantly improves the identification of occlusion MI. Toprak et al. [13] used machine learning on high-sensitivity cardiac troponin (hs-cTn) to achieve more accurate triage decisions. Sun et al. [33] and Kalmady et al. [34] further highlighted the population-scale utility of ECG models augmented with lab data for mortality prediction and multi-condition screening.

Contrastive learning has also gained traction in biomedical representation learning [35], especially for improving modality alignment and generalization. Wei et al. [36] proposed a bimodal masked autoencoder for ECG that incorporates contrastive losses to encourage consistent representations across domains. More recent frameworks like TimeMAE [37] and TimesURL [38] introduced sophisticated augmentation and pretext strategies for time-series contrastive learning, though they are often limited to single-modality or intra-signal tasks.

Despite these advancements, the combination of structured fusion, patient-conditioned modulation, and cross-modal contrastive alignment remains underexplored in clinical modeling. This motivates the development of unified frameworks—such as our proposed GFM-MIP—that jointly address heterogeneity, personalization, and semantic coherence across ECG time-series, images, and lab features.

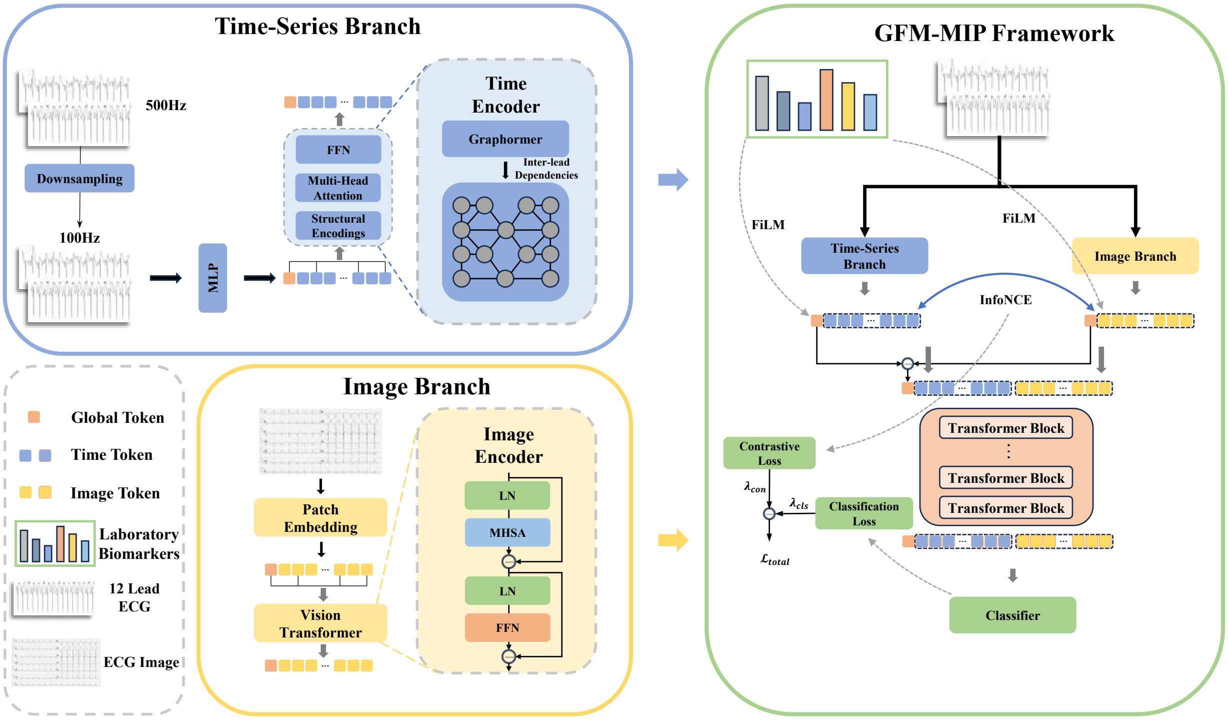

Fig. 1. Overview of the GFM-MIP framework. The proposed model performs multimodal myocardial infarction classification by jointly leveraging ECG time-series, ECG images, and laboratory biomarkers. A Graphormer encoder extracts inter-lead temporal features from ECG signals, while a Vision Transformer captures morphological patterns from ECG images. Both branches are dynamically modulated by laboratory features via FiLM layers to enable patient-specific encoding. The representations are then integrated through a Transformer-based fusion module that captures fine-grained cross-modal interactions. A contrastive learning objective further aligns the modalities in a shared latent space, enhancing robustness and diagnostic accuracy.

The Proposed Method

We propose GFM-MIP (Graph-informed and FiLM-enhanced Multimodal Fusion for Myocardial Infarction Prediction), a unified multimodal learning framework tailored for ECG-based cardiovascular disease classification. GFM-MIP is designed to effectively integrate three heterogeneous yet complementary modalities: 12-lead ECG time-series, ECG images, and laboratory biomarkers.

子标题有些口语化，
建议改一些能直接体现内容和创新的
Multimodality of xxx？这种类似的

Input Modalities and Notation

Traditional ECG classification models predominantly rely on a single modality, such as either time-series signals or waveform images, and typically overlook additional clinical insights from complementary sources like laboratory test results. GFM-MIP overcomes these limitations by integrating three distinct yet complementary modalities, creating a comprehensive representation of patient health:

$$\mathcal{D}_i = \{\mathbf{X}_{\mathrm{ECG},i}, \mathbf{X}_{\mathrm{img},i}, \mathbf{x}_{\mathrm{lab},i}, y_i\}, \tag{1}$$

where $\mathbf{X}_{\mathrm{ECG},i} \in \mathbb{R}^{12\times T}$ denotes the 12-lead ECG time series, $\mathbf{X}_{\mathrm{img},i} \in \mathbb{R}^{1\times 224\times 224}$ represents the ECG grayscale image, $\mathbf{x}_{\mathrm{lab},i} \in \mathbb{R}^{D_{\mathrm{lab}}}$ is the laboratory feature vector, and $y_i \in \{0,1\}$ is the binary diagnostic label.

The integration of these modalities is motivated by their inherent diagnostic complementarity and physiological interpretability in cardiovascular medicine:

ECG Time Series ($\mathbf{X}_{\mathrm{ECG}}$): Provide high-resolution temporal information reflecting cardiac electrophysiological activity. Each ECG lead offers distinct anatomical perspectives, enabling precise identification of abnormalities like ST-segment elevation, T-wave inversion, and arrhythmias. For example, characteristic ST-segment elevation in leads V2–V4 typically signals anterior myocardial infarction.

ECG Images ($\mathbf{X}_{\mathrm{img}}$): Capture a two-dimensional representation of ECG waveform morphology, preserving waveform shapes, amplitude relationships, and rhythm patterns in alignment with clinical visual interpretation practices. This visual modality enriches the model with structural and morphological insights not fully accessible through temporal signals alone.

Laboratory Features ($\mathbf{x}_{\mathrm{lab}}$): Offer essential systemic physiological context through biomarkers like troponin (cardiac injury indicator), D-dimer (marker of thrombosis), and white blood cell count (inflammation marker). These features provide critical supplementary clinical information, facilitating differential diagnosis even in scenarios presenting similar ECG patterns.

By integrating these three modalities, GFM-MIP significantly enhances representation robustness. Temporal ECG signals encapsulate detailed electrophysiological dynamics, images deliver morphological context, and laboratory data add systemic physiological dimensions. This multimodal strategy mitigates common challenges such as signal noise, incomplete data, and ambiguous waveform interpretation often encountered with unimodal methods.

A distinctive innovation of GFM-MIP is the explicit incorporation of laboratory features ($\mathbf{x}_{\mathrm{lab}}$), which are

内容好像有些混，讲模态的时候同时讲创新，还讲一些其他，
是不是可以在刚开始的时候来一个overview（正好图1好像没引用），
把这些创新/原理什么的都先讲一遍，后面就只讲方法步骤和细节了。

notably absent from most publicly available ECG datasets. Incorporating lab-derived features allows patient-specific modulation of ECG representations through FiLM layers, greatly enhancing the specificity and adaptability of the model. For instance, patients exhibiting similar ECG waveforms can be effectively differentiated by their distinct troponin levels, which GFM-MIP leverages to tailor representation encoding accordingly.

To operationalize this multimodal design, we developed a custom dataset featuring synchronized acquisition of all three modalities. This dataset facilitates precise alignment and integration of temporal, spatial, and biochemical information, marking both conceptual and practical advancements over existing unimodal and bimodal datasets.

Collectively, GFM-MIP synthesizes cardiac electrophysiological data ($\mathbf{X}_{\text{ECG}}$), waveform morphology ($\mathbf{X}_{\text{img}}$), and systemic physiological status ($\mathbf{x}_{\text{lab}}$) into a unified and clinically aligned diagnostic representation ($\mathcal{D}_i$). This comprehensive approach significantly enhances prediction accuracy, model robustness, and clinical interpretability, particularly in complex cardiovascular classification scenarios.

Temporal Graph Encoder with FiLM Modulation

To model the structured spatiotemporal nature of 12-lead ECG signals, GFM-MIP employs a Graphormer-based encoder augmented with Feature-wise Linear Modulation (FiLM). This design captures both inter-lead dependencies and patient-specific contextual variation.

Let $\mathbf{X}_{\text{ECG}} = [\mathbf{x}_1, \ldots, \mathbf{x}_{12}] \in \mathbb{R}^{12\times T}$, where each $\mathbf{x}_\ell \in \mathbb{R}^T$ represents the time series from lead ℓ. Each lead is projected to a latent representation using a learnable linear transformation:

$$\mathbf{h}_\ell^{(0)} = \mathbf{W}_{\text{proj}} \cdot \mathbf{x}_\ell + \mathbf{b}_{\text{proj}} \in \mathbb{R}^d, \quad \forall \ell \in \{1, \ldots, 12\}. \tag{2}$$

We then encode lead-wise spatial relationships via a Graphormer block, in which spatial biases are implicitly learned and incorporated into the attention mechanism. Specifically, the encoded node representations $\mathbf{H}^{(l)} \in \mathbb{R}^{12\times d}$ are updated layer-wise using a graph-based self-attention operator:

$$\text{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{QK}^\top + \mathbf{B}}{\sqrt{d}}\right)\mathbf{V}, \tag{3}$$

where $\mathbf{B} \in \mathbb{R}^{12\times 12}$ represents learnable spatial bias terms among the ECG leads.

To enable context-aware encoding, we apply FiLM modulation using laboratory features $\mathbf{x}_{\text{lab}}$. The FiLM generator maps lab vectors to scale and shift parameters:

$$\gamma = f_\gamma(\mathbf{x}_{\text{lab}}), \quad \beta = f_\beta(\mathbf{x}_{\text{lab}}), \tag{4}$$

which modulate intermediate features at each encoding layer:

$$\mathbf{H}_{\text{FiLM}}^{(l)} = \gamma \odot \mathbf{H}^{(l)} + \beta. \tag{5}$$

This mechanism injects personalized physiological information into the lead representations, enhancing patient-specific discrimination.

Finally, we extract a global token $\mathbf{z}_{\text{ts,cls}}$ by prepending a learnable [CLS] vector and applying an output projection:

$$\mathbf{z}_{\text{ts,cls}} = \text{TransformerEncoder}([\mathbf{h}_{\text{cls}}; \mathbf{H}^{(L)}]) \in \mathbb{R}^d. \tag{6}$$

This component outputs temporally and spatially enriched ECG embeddings, personalized by clinical laboratory context, for downstream fusion and classification.

Morphological Encoder with Visual FiLM

GFM-MIP employs a Vision Transformer (ViT) architecture augmented by Feature-wise Linear Modulation (FiLM) to effectively capture morphological information from ECG images, integrating patient-specific clinical context derived from laboratory biomarkers.

Given an ECG grayscale image $\mathbf{X}_{\text{img}} \in \mathbb{R}^{1\times H\times W}$, we partition it into N patches, each of size $P \times P$. Each patch is flattened and linearly projected into a latent embedding space:

$$\mathbf{z}_i^{(0)} = \mathbf{W}_{\text{patch}} \cdot \text{Flatten}(\mathbf{X}_i) + \mathbf{b}_{\text{patch}} \in \mathbb{R}^d, \quad i = 1, \ldots, N. \tag{7}$$

We prepend a learnable classification token $\mathbf{z}_{\text{cls}}$ to these embeddings and incorporate positional encodings $\mathbf{E}_{\text{pos}}$ to retain spatial context:

$$\mathbf{Z}^{(0)} = [\mathbf{z}_{\text{cls}}; \mathbf{z}_1^{(0)} + \mathbf{e}_1; \ldots; \mathbf{z}_N^{(0)} + \mathbf{e}_N] \in \mathbb{R}^{(N+1)\times d}. \tag{8}$$

To condition these representations on patient-specific clinical information, FiLM modulation parameters $\gamma^{(l)}$ and $\beta^{(l)}$ are computed from laboratory feature vectors $\mathbf{x}_{\text{lab}}$. These parameters modulate embeddings after each transformer encoder block:

$$\mathbf{Z}_{\text{FiLM}}^{(l)} = \gamma^{(l)} \odot \mathbf{Z}^{(l)} + \beta^{(l)}. \tag{9}$$

This modulation injects patient-specific physiological bias into the visual encoding of ECG morphology, facilitating personalized feature extraction.

The final image representation $\mathbf{z}_{\text{img,cls}}$ is obtained from the output [CLS] token:

$$\mathbf{z}_{\text{img,cls}} = \mathbf{Z}_0^{(L)} \in \mathbb{R}^d. \tag{10}$$

Consequently, this branch effectively captures spatially rich morphological features from ECG images, contextualized by laboratory-derived biomarkers, significantly enhancing the multimodal representational power of GFM-MIP.

Cross-modal Fusion Module

To effectively integrate the representations derived from the time-series and image branches, GFM-MIP introduces a dedicated Transformer encoder-based fusion module. This component is designed to capture both modality-specific and cross-modal interactions, encompassing global diagnostic patterns and fine-grained temporal-spatial dependencies.

Let the following notations define the input to the fusion module:

- $\mathbf{z}_{\text{ts,cls}} \in \mathbb{R}^d$: global classification token from the time-series branch;
- $\mathbf{z}_{\text{img,cls}} \in \mathbb{R}^d$: global classification token from the image branch;
- $\mathbf{H}_{\text{ts}} \in \mathbb{R}^{12\times d}$: lead-level representations from the time-series encoder;

- $\mathbf{H}_{\text{img}} \in \mathbb{R}^{N \times d}$: patch-level embeddings from the ViT image encoder.

The two global tokens $\mathbf{z}_{\text{ts,cls}}$ and $\mathbf{z}_{\text{img,cls}}$ are concatenated and projected through a learnable linear transformation to produce a unified multimodal token:

$$\mathbf{z}_{\text{global}} = \mathbf{W}_f[\mathbf{z}_{\text{ts,cls}}; \mathbf{z}_{\text{img,cls}}] + \mathbf{b}_f \in \mathbb{R}^d, \tag{11}$$

where $\mathbf{W}_f \in \mathbb{R}^{d \times 2d}$ and $\mathbf{b}_f \in \mathbb{R}^d$ are trainable parameters. This global token represents a synthesized summary of temporal and morphological information across modalities.

The fused token is then concatenated with the full sequence of lead- and patch-level embeddings:

$$\mathbf{Z}^{(0)}_{\text{fusion}} = [\mathbf{z}_{\text{global}}; \mathbf{H}_{\text{ts}}; \mathbf{H}_{\text{img}}] \in \mathbb{R}^{(1+12+N) \times d}. \tag{12}$$

This input sequence is fed into a stack of Transformer encoder layers that jointly model hierarchical dependencies and semantic interactions across modalities. These layers are responsible for learning attention-driven relationships that may span time-series leads, image patches, or combinations thereof:

$$\mathbf{Z}^{(L)}_{\text{fusion}} = \text{TransformerEncoder}(\mathbf{Z}^{(0)}_{\text{fusion}}). \tag{13}$$

The output at the first position of the final layer—corresponding to the fused global token—is extracted as the comprehensive multimodal representation:

$$\mathbf{z}_{\text{fusion,cls}} = \mathbf{Z}^{(L)}_{\text{fusion}}[0] \in \mathbb{R}^d. \tag{14}$$

This fused embedding encapsulates the full temporal, spatial, and clinical context, serving as the primary input for subsequent classification and contrastive learning modules.

Contrastive Alignment Objective

To promote consistent and semantically aligned representations across modalities, GFM-MIP incorporates a contrastive learning mechanism tailored for the multimodal ECG setting. This strategy enhances the coherence of the learned feature space by bringing together embeddings from different modalities that correspond to the same patient, while simultaneously pushing apart those from different patients.

Let $\mathbf{z}_{\text{ts},i}$ and $\mathbf{z}_{\text{img},i}$ denote the global classification token embeddings extracted from the time-series and image branches, respectively, for the i-th patient in a mini-batch of size B. The alignment objective is based on a symmetric InfoNCE loss, formulated as:

$$\mathcal{L}_{\text{con}} = -\frac{1}{2B} \sum_{i=1}^{B} \left(\ell_i^{\text{ts}\to\text{img}} + \ell_i^{\text{img}\to\text{ts}} \right) \tag{15}$$

where

$$\ell_i^{\text{ts}\to\text{img}} = \log \frac{\exp\left(\text{sim}(\mathbf{z}_{\text{ts},i}, \mathbf{z}_{\text{img},i})/\tau\right)}{\sum_{j=1}^{B} \exp\left(\text{sim}(\mathbf{z}_{\text{ts},i}, \mathbf{z}_{\text{img},j})/\tau\right)} \tag{16}$$

$$\ell_i^{\text{img}\to\text{ts}} = \log \frac{\exp\left(\text{sim}(\mathbf{z}_{\text{img},i}, \mathbf{z}_{\text{ts},i})/\tau\right)}{\sum_{j=1}^{B} \exp\left(\text{sim}(\mathbf{z}_{\text{img},i}, \mathbf{z}_{\text{ts},j})/\tau\right)} \tag{17}$$

where $\text{sim}(\cdot,\cdot)$ denotes cosine similarity, and τ is a temperature scaling factor that controls the sharpness of the distribution.

This bidirectional formulation ensures mutual agreement between the representations from both modalities. By minimizing this loss, the model learns to align ECG signal and image representations at the patient level, reinforcing semantic correspondence across heterogeneous views. This contrastive signal acts as a regularization force, complementing the supervised classification objective and contributing to more robust multimodal feature learning under clinical supervision.

Joint Optimization

Final classification in GFM-MIP is performed using the fused multimodal representation $\mathbf{z}_{\text{fusion,cls}}$, which captures global information integrated across time-series, image, and laboratory modalities. This embedding is passed through a linear classifier to produce class logits:

$$\hat{\mathbf{y}} = \text{softmax}(\mathbf{W}_{\text{cls}} \cdot \mathbf{z}_{\text{fusion,cls}} + \mathbf{b}_{\text{cls}}), \tag{18}$$

where $\mathbf{W}_{\text{cls}}$ and $\mathbf{b}_{\text{cls}}$ are learnable parameters. The classification loss $\mathcal{L}_{\text{cls}}$ is computed using the cross-entropy criterion, with optional weighting schemes to address potential class imbalance.

In addition to classification, GFM-MIP incorporates the contrastive loss $\mathcal{L}_{\text{con}}$ previously described in Section D, which aligns modality-specific embeddings in the latent space. This dual-objective training strategy allows the model to simultaneously optimize for predictive accuracy and representational coherence.

The overall training objective combines both components as follows:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{cls}} \cdot \mathcal{L}_{\text{cls}} + \lambda_{\text{con}} \cdot \mathcal{L}_{\text{con}}, \tag{19}$$

where λ_{cls} and λ_{con} are hyperparameters that balance the classification and contrastive learning contributions. This composite loss ensures that GFM-MIP develops both discriminative and well-aligned multimodal representations tailored for robust clinical prediction.

移到前面overview的时候？
整个方法设计的很好，逻辑很强，
但是现在的版本方法部分逻辑不强

Summary

GFM-MIP offers a unified multimodal solution for cardiovascular disease prediction by integrating ECG time-series, ECG images, and laboratory test features. Through the combination of modality-specific encoders, FiLM-based patient-level modulation, and a Transformer-based fusion mechanism, the architecture is capable of capturing complex temporal, morphological, and physiological patterns. By incorporating contrastive learning into the optimization process, GFM-MIP ensures semantic consistency across modalities, encouraging robust and interpretable feature representations. This integrated design not only enhances classification performance but also aligns well with clinical reasoning, thereby advancing the development of explainable and personalized diagnostic models.

建议还是拆开两个subsection，缺点就是Evaluation metrics的subsection会稍微短了一点，但会更清晰一点。

Results

Datasets and Metrics

To assess the effectiveness of our proposed model, we conducted comprehensive experiments using three publicly available datasets: the Ningbo dataset [39] from Ningbo First Hospital, the PTB-XL dataset [40] provided by Physikalisch-Technische Bundesanstalt, and the Chapman dataset [41] jointly developed by Chapman University and Shaoxing People's Hospital. These datasets consist of 34,905, 21,837, and 10,247 clinical records, respectively, each comprising 12-lead ECG signals spanning 10 seconds. The raw ECG data were sourced from the PhysioNet Computing in Cardiology (CinC) 2021 Challenge database [42], in which all recordings were uniformly sampled at 500 Hz.

Table 1. Statistics of ECG Datasets

Dataset	#Samples	Time-series	Lab Tests	Labels	Sampling frequency
Ningbo	34,905	12 leads, 10 s	✗	25 classes	500Hz
PTB-XL	21,837	12 leads, 10 s	✗	21 classes	500Hz
Chapman	10,247	12 leads, 10 s	✗	18 classes	500Hz
SDU-SH	2,123	12 leads, 10 s	✓	MI / Non-MI	500Hz

In the CinC 2021 database, each ECG record is annotated with one or multiple labels corresponding to various cardiac rhythm categories, mapped to SNOMED-CT codes. Therefore, the experiments conducted on these public datasets involved multi-label classification tasks.

Moreover, to further validate our model's utility in clinical practice, we collaborated with the Second Hospital of Shandong University to construct a new, proprietary dataset comprising 2,123 clinical records. Each record in this dataset includes a 12-lead ECG and corresponding laboratory test results, annotated with a single binary label. Consequently, this particular dataset was employed for binary classification tasks.

We utilized four widely adopted performance metrics to rigorously evaluate classification performance: accuracy (ACC), sample-level F1 score (F1), area under the receiver operating characteristic curve (AUROC), and area under the precision-recall curve (AUPRC).

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{20}$$

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{21}$$

$$Precision = \frac{TP}{TP + FP} \tag{22}$$

$$Recall = \frac{TP}{TP + FN} \tag{23}$$

where TP, TN, FP, and FN denote the number of true positives, true negatives, false positives, and false negatives, respectively.

Data Preprocessing

To ensure robust feature extraction and reliable classification performance, we performed a systematic preprocessing pipeline on all ECG datasets, addressing common artifacts and standardizing the data format.

Denoising. Raw ECG recordings typically contain various types of interference, such as noise, baseline wander, and motion artifacts, which adversely affect classification accuracy [43]. To mitigate these disturbances, we applied a Butterworth bandpass filter with cutoff frequencies of 0.05 and 75 Hz [44], preserving essential physiological information while reducing unwanted noise.

Downsampling. All ECG signals were downsampled from the original 500 Hz sampling rate to 100 Hz. This reduction significantly decreases computational complexity without substantial loss of diagnostic information, a practice consistent with previous studies in ECG-based modeling [45, 46].

Normalization. To alleviate potential distribution shift effects, instance normalization [47] was applied independently to each lead of every ECG record. This step ensures consistency in amplitude scales across different recordings, thereby improving model generalization [21].

Label Reconstruction. Original SNOMED-CT codes assigned to each ECG record were converted into discrete categorical labels. After this mapping process, the Ningbo, PTB-XL, and Chapman datasets contained 25, 22, and 19 distinct classes, respectively. Notably, all datasets exhibited varying degrees of class imbalance, characterized by substantial disparities in the distribution of positive versus negative samples within certain categories, as well as significant variation in the number of samples across different classes.

Through this preprocessing pipeline, we standardized input quality and format, laying a solid foundation for subsequent feature extraction and classification tasks.

Compared Methods

To comprehensively evaluate the effectiveness of our proposed GFM-MIP framework, we compare it against several representative self-supervised and contrastive learning baselines for time-series modeling.

1. **TF-C**[48]: A contrastive learning framework designed to extract temporally discriminative features from univariate and multivariate time series by contrasting representations of temporally shifted segments. It emphasizes capturing both global and local temporal dynamics.
2. **TS-TCC**[49]: This method constructs multiple augmented views of the same time series and applies contrastive objectives across different temporal perspectives. It enhances the model's robustness by explicitly learning temporal invariances.
3. **CPC**[50]: A predictive coding approach that maximizes mutual information between the current context and future latent representations. It encourages the encoder to retain predictive structure from the sequence, thus learning high-quality temporal representations.
4. **TimesURL**[38]: A unified pretraining framework that integrates multiple self-supervised objectives, such as context prediction and sequence reordering. It aims to extract generalizable representations applicable to diverse downstream tasks.
5. **SimMTM**[51]: It learns temporal dependencies by randomly masking subsequences and reconstructing them, offering simplicity and effectiveness in pretraining.
6. **PatchTST**[21]: A pure Transformer-based forecasting model that partitions time series into non-overlapping patches. By utilizing global self-attention across patches, it achieves strong performance in capturing long-range temporal dependencies.
7. **TimeMAE**[37]: A masked autoencoding framework tailored for time-series data. It reconstructs randomly masked segments using encoder-decoder architecture and has shown competitive performance in transfer learning settings.

These methods serve as strong baselines for benchmarking multimodal and unimodal ECG representation learning.

放methods的最后

Implementation Details

Our model is implemented using PyTorch and trained on a single RTX 4090 GPU. During training, we set the batch size to 32 and employ 4 worker threads for data loading. The total number of training epochs is 50, with the initial learning rate set to 3.26×10^{-5}. A linear warm-up strategy is applied over approximately 13% of the total training steps , followed by cosine annealing for learning rate decay.

We optimize the model using the AdamW optimizer with default parameters and apply a loss function composed of two terms: a supervised classification loss and a contrastive alignment loss. The weighting coefficients for the classification and contrastive objectives are set to $\lambda_{\text{cls}} = 1.0$ and $\lambda_{\text{contrast}} = 0.445$, respectively. The temperature parameter in the contrastive loss is fixed at 0.1.The fusion module adopts a Transformer encoder structure with 3 layers, each consisting of 4 attention heads . The shared fusion representation has a dimensionality of 256.

To address the issue of label imbalance present in the clinical dataset, we apply class weighting during the training process to stabilize optimization and improve the reliability of evaluation metrics.All baseline methods are trained using the recommended hyperparameter settings provided in their original implementations, ensuring a fair comparison with our model.

这两个子标题是否也能改到能体现方法的优势。这里是两个数据集不同，现在的写法就是两个普通的实验结果展示。

Experimental Results

Experimental Results on the SDU-SH Dataset

To comprehensively evaluate the diagnostic performance of GFM-MIP in a practical clinical context, we conducted extensive experiments on the SDU-SH dataset, which comprises paired 12-lead ECG recordings and corresponding laboratory test results. For a thorough comparison, we selected various strong baseline methods, including supervised transformer variants (TF-C), temporal contrastive learning models (TS-TCC, CPC), self-supervised learning frameworks (TimesURL, SimMTM), and recent time-series transformer models (PatchTST, TimeMAE).

As shown in Table 2, GFM-MIP demonstrated consistently superior performance across all four evaluation metrics—accuracy (ACC), F1 score, AUROC, and AUPRC—surpassing the baseline models. GFM-MIP outperformed all baselines on SDU-SH by +2.15% in AUROC and +2.33% in AUPRC, particularly excelling in high-sensitivity recall scenarios. Notably, GFM-MIP achieved significant improvements in the F1 score and AUPRC, indicating robust predictive performance and strong capabilities for addressing class imbalance. Compared with top-performing baselines such as TS-TCC, GFM-MIP exhibited clear advantages in AUROC and AUPRC, reflecting its superior ability to differentiate positive and negative clinical cases effectively.

While certain baselines, including TS-TCC and CPC, exhibited strong individual performances, they lacked explicit incorporation of patient-specific physiological information and multi-modal fusion mechanisms. By contrast, GFM-MIP leverages an early-stage FiLM modulation strategy that integrates laboratory biomarkers, modality-specific encoders, and cross-modal alignment through contrastive learning. This sophisticated design enables GFM-MIP to effectively model patient heterogeneity and subtle diagnostic signals embedded within multimodal ECG data.

These findings highlight that GFM-MIP not only competes strongly against existing state-of-the-art methods but also demonstrates unique strengths when applied to rich, clinically annotated multimodal datasets. Its design facilitates precise and personalized predictions, emphasizing its practical clinical value.

Table 2. Comparison among ours model and baseline methods over the SDU-SH Dataset.The best results are bold and the second best are underlined.

Models	SDU-SH			
	ACC	F1	AUROC	AUPRC
TF-C$_1$	93.73	89.42	95.93	92.26
TF-C$_R$	95.96	89.35	95.16	91.22
TS-TCC	97.80	94.51	96.72	95.59
CPC	95.44	90.70	96.45	96.47
TimesURL	96.43	93.18	95.75	94.48
SimMTM$_D$	92.20	88.60	94.35	85.34
SimMTM$_Q$	92.29	86.91	95.15	90.63
PatchTST	94.09	89.70	96.22	92.37
TimeMAE	95.22	92.71	96.53	95.61
GFM-MIP (Ours)	**98.82**	**96.36**	**98.87**	**98.80**

Experimental Results on Public Datasets

To further validate GFM-MIP's generalizability under varied real-world conditions, we evaluated its performance on three publicly available ECG datasets: Ningbo, PTB-XL, and Chapman. Unlike the SDU-SH dataset, these public datasets do not include laboratory test features, necessitating a reduced variant of GFM-MIP without laboratory-based modulation. This scenario allowed us to investigate the framework's resilience when operating with incomplete modalities.

Even in the absence of laboratory data, Table 3 shows that GFM-MIP delivered strong and competitive results across all datasets. On the Ningbo dataset, GFM-MIP achieved the highest accuracy and F1 scores among the evaluated models, maintaining comparable AUROC scores relative to top-performing alternatives. Similarly, on the PTB-XL dataset, despite the lack of laboratory features, GFM-MIP exhibited consistently robust accuracy and balanced performance across metrics. On the Chapman dataset, GFM-MIP also achieved superior accuracy and F1 scores, demonstrating its reliable predictive capacity and generalization capability across datasets with varying distributions and label complexities.

Notably, several baseline models benefited from sophisticated self-supervised pretraining or specialized contrastive learning objectives. Nevertheless, even without complete multimodal input, GFM-MIP maintained strong performance through its fundamental architectural strengths, including modality-specific encoders and structured cross-modal interaction. These features allow GFM-MIP to extract robust and transferable representations from the available ECG modalities.

Overall, these experiments underscore the flexibility and robustness of GFM-MIP, highlighting its capability to maintain strong performance even when faced with modality limitations. This further supports its practical applicability and adaptability for diverse real-world ECG classification scenarios.

Ablation Experiments

To comprehensively evaluate the specific contributions of each component within the GFM-MIP framework, we systematically

Table 3. Comparison among ours model and baseline methods on Public Datasets. The best results are bold and the second best are underlined.

Models	Ningbo				PTB-XL				Chapman			
	ACC	F1	AUROC	AUPRC	ACC	F1	AUROC	AUPRC	ACC	F1	AUROC	AUPRC
TF-C_I	55.87	75.30	92.83	49.09	51.08	78.84	91.06	53.58	59.59	77.10	91.89	51.84
TF-C_R	57.58	76.92	92.27	50.25	50.05	77.61	88.24	49.70	60.57	76.74	89.59	**52.27**
TS-TCC	61.07	81.79	92.61	53.72	55.75	**81.23**	88.42	50.60	61.69	78.29	89.18	46.28
CPC	59.26	79.32	**95.25**	**56.08**	55.24	80.61	**93.14**	**55.03**	59.25	75.97	**91.97**	46.33
TimesURL	59.43	81.20	94.50	53.49	53.85	80.04	91.23	50.99	55.44	75.96	89.99	48.13
SimMTM$_D$	54.85	75.67	90.79	44.68	52.98	78.84	87.17	45.03	45.88	61.12	86.43	38.50
SimMTM$_R$	55.02	74.10	91.48	49.36	54.67	80.43	85.80	47.24	59.44	75.74	88.25	47.03
PatchTST	55.28	77.05	93.47	50.07	54.90	80.40	90.64	50.05	53.34	73.93	91.21	49.28
TimeMAE	57.76	79.97	93.55	54.30	54.05	80.32	90.49	49.91	57.20	74.58	86.82	46.81
GFM-MIP (Ours)	**61.53**	**82.71**	91.91	52.81	**56.09**	79.40	88.58	51.36	**64.73**	**78.92**	89.84	48.78

conducted ablation experiments shown in Table 4 addressing both modality integration and structural design. These experiments provided detailed insights into how each element impacts the model's performance and robustness.

The model variants with specific modalities ablated are denoted as "w/o Laboratory," "w/o IMG," and "w/o TS," corresponding to the exclusion of laboratory test features, ECG images, and ECG time-series data, respectively. The complete version of our model is referred to as the "Full Model." In addition to modality ablations, we also investigate several architectural variations. The variant without contrastive learning is denoted as "w/o Contrast," while the model without the cross-modal fusion module is denoted as "w/o Fusion." Moreover, we examine two alternative fusion strategies: incorporating laboratory features only at the fusion stage, referred to as "Late-Concat," and employing a fully shared transformer fusion structure across all modalities, denoted as "All-Shared." These ablation settings are designed to isolate the contributions of each modality and architectural component to the overall performance of the GFM-MIP framework.

Initially, we investigated the role of incorporating multiple data modalities by individually removing them from the complete model. Eliminating the laboratory test features—which serve as personalized physiological context through FiLM modulation—resulted in noticeable performance degradation. This finding underscores the significance of integrating systemic biomarkers to enhance patient-specific modeling and improve classification accuracy. Similarly, when either ECG time series or ECG images were used exclusively, model performance significantly deteriorated compared to the multimodal baseline. Particularly, the absence of ECG time-series data markedly reduced performance, reflecting the critical role of electrophysiological dynamics, while the exclusion of ECG images confirmed the importance of morphological features. Thus, both temporal and morphological modalities offer unique and complementary diagnostic information that jointly enhances diagnostic effectiveness.

Furthermore, we examined the influence of key architectural elements by selectively removing or modifying them. Removing the contrastive learning objective negatively affected representation consistency and alignment across modalities, leading to less coherent multimodal representations and reduced classification capability. Omitting contrastive learning led to a 2.18% AUPRC drop, confirming its role in enforcing cross-modal alignment, as further illustrated in the latent space visualization (Fig. 2). Similarly, substituting the Transformer-based deep fusion module with a simpler concatenation strategy resulted in decreased model performance, highlighting the necessity of complex cross-modal interaction mechanisms for effectively capturing intricate modality interdependencies. Additionally, delaying the integration of laboratory features until after the fusion stage, rather than embedding them through early FiLM modulation, significantly compromised the generalization and diagnostic power of the model. This emphasizes the importance of early-stage personalization to effectively influence subsequent feature representation. Lastly, replacing modality-specific encoders with a shared generic encoder structure severely degraded performance, clearly indicating the necessity of distinct architectural designs tailored specifically for each data modality.

Table 4. Performance analysis of our model under different ablation settings.

Variants	**ACC**	**F1 Score**	**AUC**	**AUPRC**
Full Model(Ours)	**98.82**	**96.36**	**98.87**	**98.80**
w/o Laboratory	94.58	89.59	96.99	95.35
w/o IMG	67.53	62.32	64.63	39.48
w/o TS	90.57	82.30	97.43	94.08
w/o Contrast	96.46	92.15	97.41	96.62
w/o Fusion	94.59	90.82	96.82	95.40
Late-Concat	85.14	76.23	92.62	89.42
All-Shared	50.94	48.51	76.02	54.77

Collectively, these ablation studies shown in Table 4 reveal that GFM-MIP's high diagnostic performance and robustness do not arise from any single isolated component. Instead, the model achieves its superior results through the cohesive integration of multimodal input data, personalized feature modulation via FiLM, sophisticated structured fusion through transformer-based modules, and representation alignment through contrastive learning. Each of these components contributes distinctly yet synergistically to the overall effectiveness, clinical interpretability, and generalization capability of the framework.

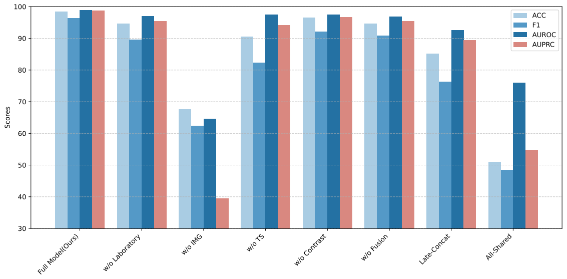

Fig. 2. Impact of modality and architectural ablations on classification metrics (SDU-SH dataset)

Parameter Analysis

Temperature Parameter and Contrastive Loss Weight

To investigate the influence of critical hyperparameters on model performance, we conduct a focused sensitivity analysis on two key components of our training objective: the temperature parameter τ in the contrastive loss and the weighting factor λ_{con} that balances the classification and contrastive objectives.

Temperature Parameter τ: As shown in Fig. 3, the model exhibits considerable sensitivity to the temperature scaling in the contrastive loss. Lower values (e.g., $\tau = 0.05$) tend to produce sharper probability distributions, potentially leading to overfitting in representation alignment and performance degradation. Conversely, overly large values (e.g., $\tau = 1.0$) flatten the similarity scores, reducing the discriminative power of contrastive pairs. The model achieves optimal F1 performance when τ lies in the intermediate range of approximately 0.1 to 0.2, suggesting a balanced trade-off between contrast sharpness and generalization.

Contrastive Loss Weight λ_{con}: Varying the contribution of the contrastive loss reveals its importance in representation learning. With no contrastive component ($\lambda_{\text{con}} = 0$), the model relies solely on supervised classification, which leads to underutilization of cross-modal alignment. As λ_{con} increases, the model benefits from enhanced modality interaction, yielding improved performance up to a point. However, excessively high values (e.g., $\lambda_{\text{con}} = 1.0$) begin to suppress the supervised signal, causing a slight decline in performance. These observations validate the necessity of balancing contrastive and classification signals, with the best performance achieved when λ_{con} is set between 0.3 and 0.5.

This analysis highlights the critical role of careful hyperparameter tuning in contrastive learning-based multimodal frameworks. Optimal values of τ and λ_{con} not only improve classification performance but also enhance the robustness of modality alignment, ultimately leading to more generalizable and reliable clinical predictions.

Fusion Layer Depth

To evaluate how the number of fusion layers affects model performance, we conducted an ablation analysis by varying the number of Transformer layers in the cross-modal fusion module from 2 to 8. As shown in Fig. 4, performance trends are consistent across all four datasets: SDU-SH, Ningbo, PTB-XL, and Chapman. Specifically, model accuracy improves with increasing fusion depth up to 6 layers, after which the performance either plateaus or slightly declines.

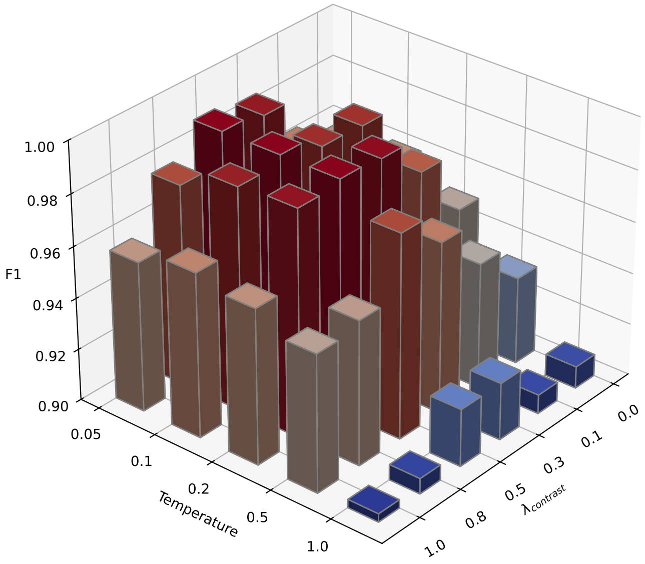

Fig. 3. The model performance with different parameter settings over SDU-SH Dataset

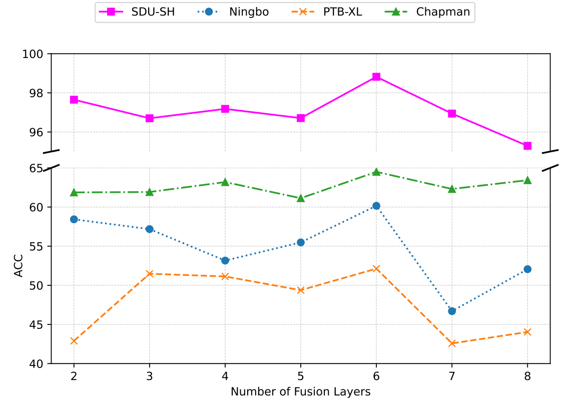

Fig. 4. Impact of fusion depth on classification accuracy across datasets

This suggests that a moderate depth offers the best trade-off between representational richness and overfitting risk. A shallow fusion module may fail to capture complex cross-modal interactions, while excessive depth can introduce noise or lead to optimization difficulties. These findings highlight the importance of tuning the fusion depth to balance expressiveness and generalization in multimodal architectures like GFM-MIP.

Visualization Analysis

To further understand the impact of contrastive learning on cross-modal representation alignment, we conduct a qualitative visualization of the learned feature embeddings from both the ECG time-series and image branches. Specifically, we project the high-dimensional modality-specific embeddings into a two-dimensional space using t-SNE. For each sample, we plot its ECG time-series embedding ($\mathbf{z}_{\text{ts}}$) and its corresponding image

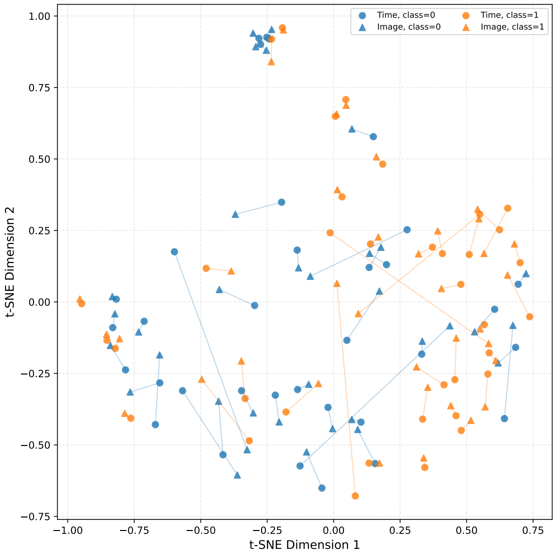

Fig. 5. Latent Space Visualization of Modality Alignment via Contrastive Learning. Connected pairs represent same patients.

embedding ($\mathbf{z}_{\text{img}}$) as paired points, connected by a dashed line to represent the same patient.

As illustrated in Figure 5, GFM-MIP effectively aligns representations across modalities. The embeddings of the same sample from different modalities are positioned closely in the latent space, demonstrating successful cross-modal feature consistency. This alignment is a direct consequence of the bidirectional contrastive loss, which encourages paired samples to share semantic similarity despite modality heterogeneity.

In addition to pairwise proximity, the scatter plot also reveals clear clustering patterns based on diagnostic labels. Samples from different classes tend to form distinct clusters, indicating that the shared embedding space preserves discriminative information for classification. This separation highlights the model's ability to encode label-relevant structure while maintaining cross-modal coherence.

These observations confirm the dual benefit of our contrastive learning strategy: it not only improves intra-sample alignment across modalities but also enhances inter-class separability, contributing to both robust fusion and accurate prediction in the GFM-MIP framework.

Conclusion

In this study, we introduced GFM-MIP, a comprehensive multimodal framework designed for effective and interpretable ECG-based cardiovascular disease prediction. By combining ECG time-series data, ECG images, and laboratory test results, GFM-MIP addresses the limitations inherent in single-modality approaches, providing a more holistic and clinically meaningful representation of patient health.

GFM-MIP integrates several innovative design elements. Modality-specific encoders were employed to accurately capture the distinct features of time-series signals and images. Additionally, we introduced Feature-wise Linear Modulation (FiLM) to embed patient-specific physiological information derived from laboratory tests into the encoding process, facilitating personalized representation learning. A Transformer-based fusion module was also incorporated to effectively model the rich interactions between different modalities, while contrastive learning further strengthened the semantic alignment across these diverse data sources.

Extensive experimental evaluations clearly demonstrate that GFM-MIP achieves superior performance compared to various baseline models and ablation variants across multiple metrics. These results underline the importance of multimodal integration, early-stage physiological conditioning, and structured cross-modal fusion in improving diagnostic accuracy and enhancing model robustness.

Future research could explore several promising directions. Incorporating additional modalities, such as clinical notes, echocardiographic imaging, or genomic data, could further enrich the patient representations. Extending the framework to multi-label or multi-task scenarios would broaden its clinical applicability, potentially supporting more complex diagnostic tasks. Investigations into the interpretability of learned representations and real-world clinical implementation will also be valuable for enhancing clinical adoption.

Overall, this study demonstrates that structured fusion of ECG modalities and laboratory features—guided by alignment loss and patient context—can meaningfully enhance diagnostic precision in real-world cardiovascular applications.

References

1. Shenda Hong *et al.* Opportunities and challenges of deep learning methods for electrocardiogram data: A systematic review. *Computers in Biology and Medicine*, 122:103801, July 2020. ISSN 00104825. doi: 10.1016/j.compbiomed.2020.103801.
2. S. Sahoo *et al.* Machine learning approach to detect cardiac arrhythmias in ECG signals: A survey. *Irbm*, 41(4):185–194, August 2020. ISSN 19590318. doi: 10.1016/j.irbm.2019.12.001.
3. Andre Esteva *et al.* A guide to deep learning in healthcare. *Nature Medicine*, 25(1):24–29, January 2019. ISSN 1078-8956, 1546-170X. doi: 10.1038/s41591-018-0316-z.
4. Omneya Attallah and Dina A. Ragab. Auto-MyIn: Automatic diagnosis of myocardial infarction via multiple GLCMs, CNNs, and SVMs. *Biomedical Signal Processing and Control*, 80:104273, February 2023. ISSN 17468094. doi: 10.1016/j.bspc.2022.104273.
5. Javad Hassannataj Joloudari *et al.* Application of artificial intelligence techniques for automated detection of myocardial infarction: A review. *Physiological Measurement*, 43(8):8TR01, August 2022. ISSN 0967-3334, 1361-6579. doi: 10.1088/1361-6579/ac7fd9.
6. Mahboobeh Jafari *et al.* Automated diagnosis of cardiovascular diseases from cardiac magnetic resonance imaging using deep learning models: A review. *Computers in Biology and Medicine*, 160:106998, June 2023. ISSN 00104825. doi: 10.1016/j.compbiomed.2023.106998.
7. V. Jahmunah *et al.* Automated detection of coronary artery disease, myocardial infarction and congestive heart failure using GaborCNN model with ECG signals. *Computers in Biology and Medicine*, 134:104457, July 2021. ISSN 00104825. doi: 10.1016/j.compbiomed.2021.104457.
8. Johannes Tobias Neumann *et al.* Personalized diagnosis in suspected myocardial infarction. *Clinical Research in Cardiology*, 112(9):1288–1301, September 2023. ISSN 1861-0684, 1861-0692. doi: 10.1007/s00392-023-02206-3.

9. Andreas Schuster *et al.* Fully automated cardiac assessment for diagnostic and prognostic stratification following myocardial infarction. *Journal of the American Heart Association*, 9(18):e016612, September 2020. ISSN 2047-9980. doi: 10.1161/JAHA.120.016612.
10. Manish Sharma *et al.* A novel automated diagnostic system for classification of myocardial infarction ECG signals using an optimal biorthogonal filter bank. *Computers in Biology and Medicine*, 102:341–356, November 2018. ISSN 00104825. doi: 10.1016/j.compbiomed.2018.07.005.
11. M.G. Tsipouras *et al.* Automated diagnosis of coronary artery disease based on data mining and fuzzy modeling. *IEEE Transactions on Information Technology in Biomedicine*, 12(4):447–458, July 2008. ISSN 1089-7771. doi: 10.1109/TITB.2007.907985.
12. Jieshuo Zhang *et al.* Automated detection and localization of myocardial infarction with staked sparse autoencoder and TreeBagger. *IEEE Access*, 7:70634–70642, 2019. ISSN 2169-3536. doi: 10.1109/ACCESS.2019.2919068.
13. Betül Toprak *et al.* Diagnostic accuracy of a machine learning algorithm using point-of-care high-sensitivity cardiac troponin I for rapid rule-out of myocardial infarction: A retrospective study. *The Lancet Digital Health*, 6(10):e729–e738, October 2024. ISSN 25897500. doi: 10.1016/S2589-7500(24)00191-2.
14. Zeynep Hilal Kilimci *et al.* Heart disease detection using vision-based transformer models from ECG images.
15. Jialu Tang *et al.* Electrocardiogram-language model for few-shot question answering with meta learning, October 2024.
16. Lauri Holmstrom *et al.* Deep learning-based electrocardiographic screening for chronic kidney disease. *Communications Medicine*, 3(1):73, May 2023. ISSN 2730-664X. doi: 10.1038/s43856-023-00278-w.
17. Cunjun Yu *et al.* Spatio-temporal graph transformer networks for pedestrian trajectory prediction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 507–523. Springer, 2020.
18. Wen-Cheng Liu *et al.* A deep learning algorithm for detecting acute myocardial infarction. *Eurointervention*, 17(9):765–773, October 2021. ISSN 1774-024X. doi: 10.4244/EIJ-D-20-01155.
19. Kamal Jafarian *et al.* Automating detection and localization of myocardial infarction using shallow and end-to-end deep neural networks. *Applied Soft Computing*, 93:106383, August 2020. ISSN 15684946. doi: 10.1016/j.asoc.2020.106383.
20. Chun-Ho Lee *et al.* Artificial intelligence-enabled electrocardiogram screens low left ventricular ejection fraction with a degree of confidence. *DIGITAL HEALTH*, 8:205520762211432, January 2022. ISSN 2055-2076, 2055-2076. doi: 10.1177/20552076221143249.
21. Yuqi Nie *et al.* A time series is worth 64 words: Long-term forecasting with transformers, March 2023.
22. Shaan Khurshid *et al.* ECG-based deep learning and clinical risk factors to predict atrial fibrillation. *Circulation*, 145(2):122–133, January 2022. ISSN 0009-7322, 1524-4539. doi: 10.1161/CIRCULATIONAHA.121.057480.
23. Neal Yuan *et al.* Deep learning of electrocardiograms in sinus rhythm from US veterans to predict atrial fibrillation. *JAMA Cardiology*, 8(12):1131, December 2023. ISSN 2380-6583. doi: 10.1001/jamacardio.2023.3701.
24. Sören J. Backhaus *et al.* Artificial intelligence fully automated myocardial strain quantification for risk stratification following acute myocardial infarction. *Scientific Reports*, 12(1):12220, July 2022. ISSN 2045-2322. doi: 10.1038/s41598-022-16228-w.
25. Ju-Hyeon Nam *et al.* Transgunet: Transformer meets graph-based skip connection for medical image segmentation. *arXiv preprint arXiv:2502.09931*, 2025.
26. Bingjun Li and Sheida Nabavi. A multimodal graph neural network framework for cancer molecular subtype classification. *BMC bioinformatics*, 25(1):27, 2024.
27. Xiaoxiao Li *et al.* Braingnn: Interpretable brain graph neural network for fmri analysis. *Medical Image Analysis*, 74:102233, 2021.
28. Joshua Lampert *et al.* A novel ECG-based deep learning algorithm to predict cardiomyopathy in patients with premature ventricular complexes. *JACC: Clinical Electrophysiology*, 9(8):1437–1451, August 2023. ISSN 2405500X. doi: 10.1016/j.jacep.2023.05.025.
29. Jasper Boeddinghaus *et al.* Machine learning for myocardial infarction compared with guideline-recommended diagnostic pathways. *Circulation*, 149(14):1090–1101, April 2024. ISSN 0009-7322, 1524-4539. doi: 10.1161/CIRCULATIONAHA.123.066917.
30. Dimitrios Doudesis *et al.* Machine learning for diagnosis of myocardial infarction using cardiac troponin concentrations. *Nature Medicine*, 29(5):1201–1210, May 2023. ISSN 1078-8956, 1546-170X. doi: 10.1038/s41591-023-02325-4.
31. Salah Al-Zaiti *et al.* Machine learning for the ECG diagnosis and risk stratification of occlusion myocardial infarction at first medical contact, January 2023.
32. Salah S. Al-Zaiti *et al.* Machine learning for ECG diagnosis and risk stratification of occlusion myocardial infarction. *Nature Medicine*, 29(7):1804–1813, July 2023. ISSN 1546-170X. doi: 10.1038/s41591-023-02396-3.
33. Weijie Sun *et al.* Towards artificial intelligence-based learning health system for population-level mortality prediction using electrocardiograms. *npj Digital Medicine*, 6(1):21, February 2023. ISSN 2398-6352. doi: 10.1038/s41746-023-00765-3.
34. Sunil Vasu Kalmady *et al.* Development and validation of machine learning algorithms based on electrocardiograms for cardiovascular diagnoses at the population level. *npj Digital Medicine*, 7(1):133, May 2024. ISSN 2398-6352. doi: 10.1038/s41746-024-01130-8.
35. Gurukripa N. Kowlgi *et al.* Deep learning for premature ventricular contraction-cardiomyopathy. *JACC: Clinical Electrophysiology*, 9(8):1452–1454, August 2023. ISSN 2405500X. doi: 10.1016/j.jacep.2023.07.003.
36. Yufeng Wei *et al.* Bimodal Masked Autoencoders with internal representation connections for electrocardiogram classification. *Pattern Recognition*, 161:111311, May 2025. ISSN 00313203. doi: 10.1016/j.patcog.2024.111311.
37. Mingyue Cheng *et al.* TimeMAE: Self-supervised representations of time series with decoupled masked autoencoders, March 2023.
38. Jiexi Liu and Songcan Chen. TimesURL: Self-supervised contrastive learning for universal time series representation learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(12):13918–13926, March 2024. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v38i12.29299.

39. Jianwei Zheng *et al.* Optimal multi-stage arrhythmia classification approach. *Scientific Reports*, 10(1):2898, February 2020. ISSN 2045-2322. doi: 10.1038/s41598-020-59821-7.
40. Patrick Wagner *et al.* PTB-XL, a large publicly available electrocardiography dataset. *Scientific Data*, 7(1):154, May 2020. ISSN 2052-4463. doi: 10.1038/s41597-020-0495-6.
41. Jianwei Zheng *et al.* A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients. *Scientific Data*, 7(1):48, February 2020. ISSN 2052-4463. doi: 10.1038/s41597-020-0386-x.
42. Matthew A Reyna *et al.* Will two do? Varying dimensions in electrocardiography: The PhysioNet/computing in cardiology challenge 2021. In *2021 Computing in Cardiology (Cinc)*, pages 1–4, Brno, Czech Republic, September 2021. IEEE. ISBN 978-1-6654-7916-5. doi: 10.23919/CinC53138.2021.9662687.
43. Xiaoyun Xie *et al.* A multi-stage denoising framework for ambulatory ecg signal based on domain knowledge and motion artifact detection. *Future Generation Computer Systems*, 116:103–116, 2021.
44. Vessela Krasteva *et al.* Real-time arrhythmia detection with supplementary ecg quality and pulse wave monitoring for the reduction of false alarms in icus. *Physiological measurement*, 37(8):1273, 2016.
45. Huaicheng Zhang *et al.* MaeFE: Masked autoencoders family of electrocardiogram for self-supervised pretraining and transfer learning. *IEEE Transactions on Instrumentation and Measurement*, 72:1–15, 2023. ISSN 0018-9456, 1557-9662. doi: 10.1109/TIM.2022.3228267.
46. Harold Martin *et al.* Real-time frequency-independent single-lead and single-beat myocardial infarction detection. *Artificial intelligence in medicine*, 121:102179, 2021.
47. Dmitry Ulyanov *et al.* Instance normalization: The missing ingredient for fast stylization, November 2017.
48. Xiang Zhang *et al.* Self-supervised contrastive pre-training for time series via time-frequency consistency.
49. Emadeldeen Eldele *et al.* Time-series representation learning via temporal and contextual contrasting, June 2021.
50. Temesgen Mehari and Nils Strodthoff. Self-supervised representation learning from 12-lead ECG data. *Computers in Biology and Medicine*, 141:105114, February 2022. ISSN 00104825. doi: 10.1016/j.compbiomed.2021.105114.
51. Jiaxiang Dong *et al.* SimMTM: A simple pre-training framework for masked time-series modeling.

Author Name. This is sample author biography text. The values provided in the optional argument are meant for sample purposes. There is no need to include the width and height of an image in the optional argument for live articles. This is sample author biography text this is sample author biography text this is sample author biography text this is sample author biography text this is sample author biography text this is sample author biography text this is sample author biography text this is sample author biography text.

Author Name. This is sample author biography text this is sample author biography text this is sample author biography text this is sample author biography text this is sample author biography text this is sample author biography text this is sample author biography text this is sample author biography text.