AMF-MedIT: An Efficient Align-Modulation-Fusion Framework for Medical Image-Tabular Data

Congjing Yu[a,b], Jing Ye[a,b], Yang Liu[a,b], Xiaodong Zhang[c] and Zhiyong Zhang[a,b,*]

[a]School of Electronics and Communication Engineering, Sun Yat-sen University, Shenzhen, 518107, Guangdong, China
[b]Guangdong Provincial Key Laboratory of Advanced IntelliSense Technology, Sun Yat-sen University, Shenzhen, 518107, Guangdong, China
[c]Department of Radiology, the Third Affiliated Hospital of Southern Medical University, Guangzhou, 510515, Guangdong, China

ARTICLE INFO

Keywords:
Multimodal learning
Image-tabular fusion
Contrastive learning
Medical tabular data
Mamba
Interpretable deep learning

ABSTRACT

Multimodal medical analysis combining image and tabular data has gained increasing attention. However, effective fusion remains challenging due to cross-modal discrepancies in feature dimensions and modality contributions, as well as the noise from high-dimensional tabular inputs. To address these problems, we present AMF-MedIT, an efficient Align-Modulation-Fusion framework for medical image and tabular data integration, particularly under data-scarce conditions. Built upon a self-supervised learning strategy, we introduce the Adaptive Modulation and Fusion (AMF) module, a novel, streamlined fusion paradigm that harmonizes dimension discrepancies and dynamically balances modality contributions. It integrates prior knowledge to guide the allocation of modality contributions in the fusion and employs feature masks together with magnitude and leakage losses to adjust the dimensionality and magnitude of unimodal features. Additionally, we develop FT-Mamba, a powerful tabular encoder leveraging a selective mechanism to handle noisy medical tabular data efficiently. Extensive experiments, including simulations of clinical noise, demonstrate that AMF-MedIT achieves superior accuracy, robustness, and data efficiency across multimodal classification tasks. Interpretability analyses further reveal how FT-Mamba shapes multimodal pretraining and enhances the image encoder's attention, highlighting the practical value of our framework for reliable and efficient clinical artificial intelligence applications.

arXiv:2506.19439v2 [cs.CV] 16 Sep 2025

1. Introduction

Rrecently, multimodal learning that integrates medical images and tabular data has emerged as a critical research frontier in precision medicine. While computer-aided diagnosis based on images has been extensively investigated, the complementary role of tabular data in enhancing clinical decision-making is gaining attention. Such tabular data—typically derived from electronic health records (EHRs), laboratory tests, and standardized questionnaires—provide rich clinical context regarding the current symptoms and past medical history, which visual patterns alone cannot capture [1]. Studies have consistently shown that integrating this clinical context improves image interpretation and supports more accurate diagnoses [2].

However, real-world applications of medical images and tabular data pose two challenges. First, significant disparities in feature dimension and confidence levels across modalities complicate the fusion process. **(1) Dimension discrepancy:** Image encoders require high-dimensional representations to retain anatomical details, whereas tabular encoders generate more compact features. For example, a ResNet-50 encoder extracts 2,048-dimensional global features, whereas a typical feature vector dimension of TabNet is 128 [3]. This imbalance in feature dimensions can lead to uneven parameter allocation during fusion, where higher-dimensional features tend to dominate in parameter space. **(2) Confidence-level discrepancy:** The reliability of each modality in multimodal fusion is influenced by the downstream task and data quality. When images provide definitive pathological signatures but tabular data suffer from missing values, an optimal model is expected to prioritize image evidence. Conversely, in predicting Ischemic Heart Disease (IHD), laboratory and lifestyle factors in tabular data often serve as more reliable prognostic indicators, with images playing a supplementary role [4]. This disparity in confidence necessitates fusion strategies with adaptive capabilities to appropriately balance multimodal contributions. Second, medical tabular data inherently contain substantial noise, which hinders robust and accurate fusion [5]. These data consist of numerous features, many

*Corresponding author.
yucj@mail2.sysu.edu.cn (C. Yu); yejing26@mail2.sysu.edu.cn (J. Ye); liuy2529@mail2.sysu.edu.cn (Y. Liu); dong78@smu.edu.cn (X. Zhang); zhangzhy99@mail.sysu.edu.cn (Z. Zhang)

of which can be irrelevant to the target disease. For instance, among the selected 117 data fields for cardiac prediction in [6], over 60% have importance scores below 0.005 as measured by integrated gradients. Additionally, the prevalent problem of missing data further amplifies the noise of medical tabular data [7].

These challenges, coupled with the persistent data scarcity in the medical domain, place greater demands on both fusion strategies and tabular representation learning, requiring a delicate balance between model capacity and data efficiency. Although the recent adaptation of Self-Supervised Learning (SSL) to the medical field has alleviated data scarcity in encoder training, core challenges in medical image-tabular fusion and tabular feature extraction remain unresolved. Naive concatenation fusion ignores both dimension and confidence disparities, whereas complex cross-attention mechanisms are data-hungry and computationally inefficient. In parallel, while architectures like Mamba [8] have advanced long-sequence modeling in natural language processing, medical tabular encoding still relies on either simplistic Multi-Layer Perceptrons (MLPs) or transformer-based models that require substantial training data to perform effectively.

In this work, we aim to propose an efficient fusion paradigm that resolves cross-modal disparities through a streamlined architecture, while enhancing the tabular encoder under limited data conditions. Our method is based on three insights. First, inspired by contrastive learning, constructing a unified cross-modal representation space can mitigate representation discrepancies in a straightforward manner. However, unlike matching tasks, medical inference requires preserving modality-specific distinctions [9]. in which only the dimensionality and magnitude of unimodal features are aligned, while allowing the contributions and essential information specific to each modality to remain distinct. Second, given the scarcity of medical multimodal data, leveraging prior knowledge from preliminary experiments enables more efficient and interpretable fusion than directly learning cross-modal interactions from limited data. Therefore, we incorporate modality confidence priors into fusion to guide adaptive contribution allocation. Third, medical tabular data can be viewed as long feature sequences containing both relevant and irrelevant elements. Mamba's selective and sequential modeling capabilities make it well-suited for processing such complex inputs [8]. This motivates us to explore Mamba for medical tabular representation.

Building on these insights, we present AMF-MedIT, an efficient Align–Modulation–Fusion framework for integrating medical images and tabular data. The framework first aligns the two modalities to obtain preliminary unimodal representations, then employs feature modulation to generate fusion-compatible unimodal representations, and finally fuses the modulated representations for downstream tasks. At its core, AMF-MedIT leverages the Adaptive Modulation and Fusion (AMF) module to ensure effective fusion. The modulation aims to reduce heterogeneity among unimodal features while adaptively adjusting the contributions of each modality, enabling robust and accurate integration across diverse downstream scenarios. Specifically, we first define a modality confidence ratio r_{conf} to incorporate prior information regarding modality confidence, and derive two modulation objectives. For implementation, we design a streamlined, easy-to-train, and adaptive module architecture. The modulation objectives are achieved through feature masks that truncate features, complemented by magnitude and leakage losses to guide training.

Additionally, for multimodal representation learning and alignment, we introduce FT-Mamba, an adaptation of Mamba tailored for tabular feature extraction and pretrained using contrastive learning. Leveraging the selective State-Space Mechanisms (SSMs), FT-Mamba effectively captures discriminative features from noisy medical records without relying on large-scale training data. Meanwhile, the self-supervised contrastive learning strategy enables initial multimodal alignment in an unlabeled manner, allowing FT-Mamba to provide effective supervision for the image encoder.

We conducted extensive comparative experiments on both clean and noisy datasets. Results demonstrate that AMF-MedIT achieves a favorable balance between data efficiency and multimodal performance, showing strong adaptability to noise in both image and tabular modalities. Furthermore, ablation studies and interpretability analyses highlight the respective advantages of the AMF module in fusion and FT-Mamba in tabular feature extraction.

The main contributions are summarized as follows.

- We develop AMF-MedIT, an efficient framework for integrating medical images and tabular data. At its core, the AMF module is a novel modulation-based paradigm that mitigates dimension disparities and adaptively balances modality contributions by leveraging prior knowledge.

- We introduce FT-Mamba, an adaptation of Mamba for medical tabular data that enhances feature extraction from noisy records, while employing self-supervised contrastive learning to enable unsupervised pretraining of image–table encoders.

- We conduct extensive experiments under both clean and clinically noisy conditions, showing that AMF-MedIT achieves superior accuracy, robustness, and data efficiency compared with recent methods.
- We provide in-depth ablation and interpretability analyses, revealing how the choice of tabular encoder influences multimodal contrastive pretraining.

2. Related research

2.1. Self-Supervised Multimodal Learning

Self-supervised learning is proposed to extract meaningful representations from unlabeled data. In multimodal learning, training models still often require expensive human annotation [10]. To mitigate this challenge, SSL has been applied to multimodal settings with various pretext tasks. CLIP utilizes image-text contrastive learning on massive web data, achieving remarkable zero-shot performance [11]. Subsequent works also combine tasks such as Masked Language Modeling (MLM) and Image-Text Matching (ITM) to enhance cross-modal understanding [12]. Recently, SSL has been explored for image-tabular fusion [6, 13, 14, 9], but existing studies have not addressed specific challenges in the medical domain. Building on contrastive learning, our approach explicitly tackles the challenges in a medical context.

2.2. Multimodal Medical Image-Tabular Learning

Multimodal image-tabular learning exploits tabular data to facilitate visual task learning, which is widely used in the medical field [1]. Several studies have employed late fusion, aggregating predictions from separate unimodal models for final decisions [4, 15]. To facilitate deeper and earlier interactions between modalities, joint fusion has become more prevalent. Typically, features from different modalities are concatenated and then processed using fully connected networks [16, 17, 33], transformer-based [18, 19], or traditional machine learning models [20] for fusion. To enhance modality interaction, some studies employ cross-attention modules [21] to joint unimodal features. However, concatenation-based methods cannot reconcile the discrepancies between unimodal features, resulting in suboptimal fusion performance. Meanwhile, cross-attention methods rely entirely on full pairwise attention without leveraging prior knowledge, leading to high computational overhead, and making it difficult to achieve effective learning on small-scale medical datasets.

While SSL has advanced image-tabular learning, fusion strategies remain unchanged. Most studies still rely on feature concatenation [6, 14, 22] or cross-attention mechanisms [23, 13, 9, 24]. To address this limitation, we propose a novel fusion paradigm that flexibly regulates unimodal features within a streamlined structure, optimizing the trade-off between fusion performance and data efficiency.

2.3. Deep Learning for Tabular Data

Tabular data present unique challenges for deep learning (DL) models, requiring them to handle various data-related issues such as noise, imprecision, missing values, diverse attribute types, and varying value ranges [5]. Existing research on deep learning for tabular data provides diverse novel architectures. Gorishniy et al. [25] compared the main deep architectures, including MLP, ResNet, and FT-Transformer, finding ResNet to be a simple yet effective baseline, whereas FT-Transformer serves as a universal architecture. Recently, Mamba has gained attention for leveraging selective SSMs to achieve linear-time complexity and memory efficiency [8]. Thielmann et al. explored its application in the tabular domain, demonstrating Mamba's feasibility for tabular data [26].

Despite this progress, tabular data processing in the medical domain remains relatively underdeveloped. Some studies use raw tabular data as input [20, 27], while others employ MLP-like architectures for simple processing [16, 19, 17, 6]. However, such simplistic models often fall short of capturing meaningful representations from noisy and high-dimensional medical tabular data. Recent works have adopted transformer-based models to enhance tabular feature extraction [18, 23, 13, 22, 9, 24]. Nevertheless, these models require large datasets for effective training and often encounter optimization difficulties on limited medical data. Inspired by FT-Transformer and Mambular, we propose a streamlined adaptation of the Mamba architecture for tabular data, specifically designed to handle the noisy nature of medical tabular data.

3. Methods

3.1. Overall Framework

The overall framework of AMF-MedIT is shown in Fig. 1. The training pipeline comprises two stages: (a) Imaging-tabular pretraining and (b) Fine-tuning for adaptive modulation and fusion. In each stage, the image-tabular pairs, ($\boldsymbol{x}_{img} \in \mathbb{R}^{H\times W\times D}$, $\boldsymbol{x}_{tab} \in \mathbb{R}^{N}$) are fed into the corresponding unimodal encoders, $\boldsymbol{E}_{img}$ and $\boldsymbol{E}_{tab}$. Here, H, W, and D denote the height, width, and number of channels of the input image, respectively, and N represents the number of tabular features. We use ResNet-50 [28] as the image encoder and FT-Mamba as the tabular encoder, which will be introduced in Section 3.2. The encoders then generate the original unimodal embeddings, $\boldsymbol{z}_{img} \in \mathbb{R}^{D_{img}}$ and $\boldsymbol{z}_{tab} \in \mathbb{R}^{D_{tab}}$, where D_{img} and D_{tab} denote the dimensions of unimodal features.

In the pretraining stage, the projection heads $\boldsymbol{\Phi}_{img}$ and $\boldsymbol{\Phi}_{tab}$ map $\boldsymbol{z}_{img}$ and $\boldsymbol{z}_{tab}$ into a shared latent space as $\hat{\boldsymbol{z}}_{img} \in \mathbb{R}^{D_\Phi}$ and $\hat{\boldsymbol{z}}_{tab} \in \mathbb{R}^{D_\Phi}$ separately. D_Φ denotes the dimensions of projections. On the supervise of the Image-Tabular Contrastive (ITC) loss, the projections are pulled and pushed to maximize the similarity between matching image-tabular pairs while minimizing the similarity of non-matching pairs.

In the fine-tuning stage, we propose a simple AMF module for multimodal fusion. Guided by the proposed magnitude and leakage losses, this module transforms the original unimodal features, $\boldsymbol{z}_{img}$ and $\boldsymbol{z}_{tab}$, into modulated features, $\tilde{\boldsymbol{z}}_{img}$ and $\tilde{\boldsymbol{z}}_{tab}$, which are then combined to form a cohesive fused feature, $\boldsymbol{z}_{fused}$. Finally, $\boldsymbol{z}_{fused}$ is fed to a Fully-Connected (FC) layer to generate the classification probability.

3.2. FT-Mamba

Inspired by FT-Transformer [25], we adapt the Mamba architecture [8] for tabular feature extraction and name it as FT-Mamba (Feature Tokenizer + Mamba). As shown in Fig. 1 (a), we first separately tokenize the categorical and numerical tabular features to embeddings and add a `[CLS]` token at the sequence end. Then a stack of mamba blocks is used to extract the high-dimensional features from the embeddings, and the final representation of the `[CLS]` token is used as the original tabular embedding, $\boldsymbol{z}_{tab}$.

3.2.1. Feature Tokenizer

Our feature tokenizer module follows the design of [25]. It transfers the tabular input $\boldsymbol{x}_{tab}$ to embeddings $\boldsymbol{t}_{tab} \in \mathbb{R}^{N\times D_{tab}}$ as shown in Fig. 2 (a). For continuous features, the feature value is mapped by a linear function with D_{tab}-dimensional weights and bias. For categorical features, each category is mapped to a vector through a lookup table and is added with the corresponding bias.

3.2.2. Mamba Blocks

As shown in Fig. 2 (b) and (c), Mamba maps input token $\boldsymbol{t}_{tab}$ to the representation $\boldsymbol{z}_{tab}$ through a hidden state $\boldsymbol{h}_{tab}$, which stores and updates long-term information without explicit self-attention. Moreover, the module employs adaptive parameters $\mathbf{B}$, $\mathbf{C}$ as functions of the input, allowing it to selectively focus on relevant information while filtering out irrelevant features. Unlike transformers, which compute all pairwise interactions, Mamba reduces unnecessary computations, facilitating efficient optimization in data-limited scenarios.

3.3. Adaptive Modulation and Fusion Module

3.3.1. Problem Formulation

Modality feature heterogeneity and imbalanced modality contributions can undermine the effectiveness of multimodal fusion. To address this, feature modulation aims to make the original unimodal features more compatible for fusion by mitigating heterogeneity in feature magnitude and dimensionality across modalities, while leveraging prior knowledge regarding modality confidence to adjust their contributions.

We first employ the ℓ_1-norm to quantify the contribution of a unimodal feature in naive multimodal concatenation. The ℓ_1-norm of a feature element directly reflects its magnitude, and the sum over all elements provides a simple and interpretable measure of the overall contribution C of a unimodal feature, as in (1a). Intuitively, modalities with larger overall contributions dominate in concatenated representations.

$$C(\boldsymbol{z}_{img}) \overset{\text{def}}{=\!=} \left\|\boldsymbol{z}_{img}\right\|_1 = \sum_{i=0}^{D_{img}} \left|z_{img\,i}\right| \tag{1a}$$

$$\bar{C}(\boldsymbol{z}_{img}) \overset{\text{def}}{=\!=} \frac{\left\|\boldsymbol{z}_{img}\right\|_1}{D_{img}} = \frac{\sum_{i=0}^{D_{img}} \left|z_{img_i}\right|}{D_{img}} \tag{1b}$$

To eliminate the influence of dimensionality, we define the average contribution $\bar{C}$ as in (1b), which reflects the average magnitude per feature element. Equations (1a) and (1b) use $\boldsymbol{z}_{img}$ as an example, and the same concept applies to the tabular modality.

To further incorporate prior knowledge regarding modality confidence into fusion, we introduce the modality confidence ratio r_{conf}. This ratio represents the confidence level of image features relative to tabular features or a given downstream task, and the confidence levels can be derived from prior knowledge, such as preliminary experiments or expert experience. In our experiments, r_{conf} is calculated based on unimodal performance metrics obtained from pretrained encoders on downstream tasks, denoted as **Metrics**:

$$r_{\text{conf}} = \frac{\textbf{Metrics}(\boldsymbol{z}_{img})}{\textbf{Metrics}(\boldsymbol{z}_{tab})}. \tag{2}$$

Based on the above definitions, we propose two objectives for modulation:

1. **Magnitude alignment.** To avoid the large-magnitude features overwhelming the smaller ones, the average feature magnitude across modulated unimodal features should be comparable:

$$\left|\bar{C}(\tilde{\boldsymbol{z}}_{img}) - \bar{C}(\tilde{\boldsymbol{z}}_{tab})\right| \to 0. \tag{3}$$

2. **Confidence and dimensionality adjustment.** To ensure that more reliable modalities contribute proportionally more to downstream tasks, the overall contribution of each modality in the fused representation should be proportional to its confidence level. Given the constraint of comparable feature magnitudes as required in (3), the overall contribution of a modality is mainly determined by its feature dimensionality according to (1b). Hence, the relative contributions can be modulated by scaling the feature lengths, which can be readily implemented within neural networks. This objective can be expressed as:

$$\frac{C(\boldsymbol{z}_{img})}{C(\boldsymbol{z}_{tab})} = r_{\text{conf}} \Rightarrow \frac{L_{img}}{L_{tab}} = r_{\text{conf}}, \tag{4}$$

   where L_{img} and L_{tab} denote the dimension of $\tilde{\boldsymbol{z}}_{img}$ and $\tilde{\boldsymbol{z}}_{tab}$, respectively.

To achieve the two objectives in (3) and (4), we propose the AMF module, which modulates the original unimodal features using a straightforward architecture and simple training tasks, enabling effective and interpretable multimodal fusion.

3.3.2. Implementation

The structure of the AMF module is shown in the Fig. 3. First, two projector heads, $\boldsymbol{\Psi}_{img}$ and $\boldsymbol{\Psi}_{tab}$, map the original features $\boldsymbol{z}_{img}$ and $\boldsymbol{z}_{tab}$ to the transitional features $\boldsymbol{z}'_{img}, \boldsymbol{z}'_{tab} \in \mathbb{R}^{D_{out}}$, where D_{out} corresponds to the dimension of the fused feature. Then, the adjustment of feature dimensions according to r_{conf} is not realized by changing the structure of the projection heads. Instead, it is realized by learning appropriate parameter values during training, guided by the subsequent feature mask and leakage loss. This design enhances the stability and generalizability of the module: when r_{conf} varies due to changes in downstream data quality, there is no need to modify the projection head structure, as the AMF module can adaptively modulate the features. Finally, the modulated features, $\tilde{\boldsymbol{z}}_{img} \in \mathbb{R}^{L_{img}}$ and $\tilde{\boldsymbol{z}}_{tab} \in \mathbb{R}^{L_{tab}}$, are concatenated to obtain the fused feature $\boldsymbol{z}_{fused}$ with a dimension of $D_{out} = L_{img} + L_{tab}$, which is subsequently fed into the decoder for downstream tasks.

Feature mask. In practical implementation, we fix D_{out} as a hyperparameter to ensure a consistent interface to downstream decoders. This design decouples the adjustment of modality contributions from the architecture of the fusion and decoding modules, thereby improving the generality of the AMF module and avoiding repeated reconfiguration of decoder input layers. Given a fixed D_{out} and combined with (4), the allocation of dimensions between

modalities is adaptively determined by the modality confidence ratio r_{conf} as follows:

$$D_{out} \text{ is fixed,} \quad \begin{cases} L_{img}(r_{\text{conf}}) &= D_{out} - \left\lfloor \frac{D_{out}}{1+r_{\text{conf}}} \right\rfloor, \\ L_{tab}(r_{\text{conf}}) &= \left\lfloor \frac{D_{out}}{1+r_{\text{conf}}} \right\rfloor. \end{cases} \tag{5}$$

To explicitly adjust the dimensionalities of unimodal features, we define a feature mask for each modality to truncate the transitional features. For the imaging modality, the mask is a binary vector of size D_{out}, where the first L_{img} entries are set to 1 and the remaining L_{tab} entries are set to 0. Conversely, the tabular feature mask adopts the opposite configuration:

$$\begin{cases} \boldsymbol{m}_{img} &= [\underbrace{1, 1, \dots, 1}_{L_{img}}, \underbrace{0, 0, \dots, 0}_{L_{tab}}] \\ \boldsymbol{m}_{tab} &= [\underbrace{0, 0, \dots, 0}_{L_{img}}, \underbrace{1, 1, \dots, 1}_{L_{tab}}]. \end{cases} \tag{6}$$

By applying the feature masks to the transitional features, the features are divided into a retained region, $\tilde{\boldsymbol{z}}_*$, and a discarded region, $\tilde{\boldsymbol{z}}^c_*$, as formulated in (7). Consequently, the retained region of each modality satisfies the dimensionality objective specified in (4).

$$\begin{aligned} \tilde{\boldsymbol{z}}_{img} &= \left[\boldsymbol{m}_{img} \odot \boldsymbol{z}'_{img}\right]_{0:L_{img}}, \qquad \tilde{\boldsymbol{z}}^c_{img} = \left[\boldsymbol{m}_{tab} \odot \boldsymbol{z}'_{img}\right]_{L_{img}:D_{out}}; \\ \tilde{\boldsymbol{z}}_{tab} &= \left[\boldsymbol{m}_{tab} \odot \boldsymbol{z}'_{tab}\right]_{L_{img}:D_{out}}, \qquad \tilde{\boldsymbol{z}}^c_{img} = \left[\boldsymbol{m}_{img} \odot \boldsymbol{z}'_{tab}\right]_{0:L_{img}} \end{aligned} \tag{7}$$

Here, $\odot$ denotes the element-wise product between two vectors. During training, the projections are optimized under the guidance of the leakage and magnitude losses, which encourage modality-specific information to be concentrated in the retained region while maintaining comparable feature magnitudes. The retained regions are then used as the modulated features, which are concatenated to form the final fused representation as follows:

$$\boldsymbol{z}_{fused} = \boldsymbol{m}_{img} \odot \boldsymbol{z}'_{img} + \boldsymbol{m}_{tab} \odot \boldsymbol{z}'_{tab}. \tag{8}$$

Leakage loss. A potential issue with masking is that meaningful information may "leak" into the discarded region, thereby reducing the effectiveness of the retained features. To address this, we introduce a leakage loss to measure the residual information magnitude within the masked-out regions:

$$\mathcal{L}_{\text{leakage}} = \frac{1}{2\mathcal{N}} \sum_{j \in \mathcal{N}} \left(\bar{C}\left(\tilde{\boldsymbol{z}}^c_{img,j}\right) + \bar{C}\left(\tilde{\boldsymbol{z}}^c_{tab,j}\right) \right) = \frac{1}{2\mathcal{N}} \sum_{j \in \mathcal{N}} \left(\frac{\left\| \left[\boldsymbol{z}'_{img,j}\right]_{L_{img}:D_{out}} \right\|_1}{L_{tab}} + \frac{\left\| \left[\boldsymbol{z}'_{img,j}\right]_{0:L_{img}} \right\|_1}{L_{img}} \right). \tag{9}$$

where $\mathcal{N}$ is the batch size. By minimizing this loss jointly with the cross-entropy (CE) loss, the projection heads $\boldsymbol{\Psi}_{img}$ and $\boldsymbol{\Psi}_{tab}$ are encouraged to compress discriminative information into the retained region, ensuring that truncation by feature masks does not compromise the expressiveness of the fused features.

Magnitude loss. To satisfy the objectives in (3), we define a magnitude loss that measures the discrepancy in information magnitude between the retained feature regions of the two modalities:

$$\mathcal{L}_{\text{mag}} = \frac{1}{\mathcal{N}} \sum_{j \in \mathcal{N}} \left(\bar{C}(\tilde{\boldsymbol{z}}_{img,j}) - \bar{C}(\tilde{\boldsymbol{z}}_{tab,j}) \right) = \frac{1}{\mathcal{N}} \sum_{j \in \mathcal{N}} \left| \frac{\left\| \left[\boldsymbol{z}'_{img,j}\right]_{0:L_{img}} \right\|_1}{L_{img}} - \frac{\left\| \left[\boldsymbol{z}'_{tab,j}\right]_{L_{img}:D_{out}} \right\|_1}{L_{tab}} \right|. \tag{10}$$

By minimizing this loss, the module is guided to reduce the magnitude gap across modalities, while ensuring the feasibility of adjusting the overall modality contribution through dimensional modulation.

3.4. Optimization

In the pretraining stage, image–tabular pairs are used for contrastive learning and the ITC loss is applied [6, 11]. Intuitively, the goal of ITC is to align paired image and tabular representations while pushing apart unpaired ones. The image feature of a given sample is encouraged to be close to its corresponding tabular feature, and simultaneously distant from tabular features of other samples in the batch. Formally, the image-to-tabular contrastive loss is defined in (11), where $\mathcal{N}$ is the batch size and τ is the temperature hyperparameter:

$$\mathcal{L}_{\text{i2t}} = -\frac{1}{\mathcal{N}} \sum_{j \in \mathcal{N}} log \frac{\exp(\cos(\hat{\boldsymbol{z}}_{img}{}^{j}, \hat{\boldsymbol{z}}_{tab}{}^{j})/\tau)}{\sum_{k \in \mathcal{N}, k \neq j} \exp(\cos(\hat{\boldsymbol{z}}_{img}{}^{j}, \hat{\boldsymbol{z}}_{tab}{}^{k})/\tau)} \tag{11}$$

The tabular-to-image contrastive loss adopts a symmetric formulation, and the overall ITC loss is computed as their average as: $\mathcal{L}_{\text{ITC}} = \left(\mathcal{L}_{\text{i2t}} + \mathcal{L}_{\text{t2i}}\right)/2$.

In the fine-tuning stage, the loss function consists of three components. Alongside the leakage loss in (9) and magnitude loss in (10) for multimodal fusion, we employ CE loss for classification. For the OL3I dataset, which suffers from severe class imbalance, we apply a weighted CE loss as defined in (12), where p is the model output, y is the classification label, and α is a hyperparameter for mitigate class imbalance.

$$\mathcal{L}_{\text{weighted ce}} = -\frac{1}{\mathcal{N}} \sum_{j \in \mathcal{N}} \left(\alpha y^j log(p^j) - (1-\alpha)(1-y^j) log(1-p^j)\right) \tag{12}$$

For the DVM dataset, standard CE loss is used as formulated in (13), where C denotes the number of classes.

$$\mathcal{L}_{\text{ce}} = -\frac{1}{\mathcal{N}} \sum_{j \in \mathcal{N}} \sum_{c \in C} y_c^j log(p_c^j) \tag{13}$$

The overall loss function is a weighted sum of these three losses, with λ_1, λ_2 as hyperparameters to balance leakage and magnitude loss as follows:

$$\mathcal{L}_{\text{OL3I}} = \mathcal{L}_{\text{weighted ce}} + \lambda_1 \mathcal{L}_{\text{leakage}} + \lambda_2 \mathcal{L}_{\text{magnitude}} \tag{14a}$$

$$\mathcal{L}_{\text{DVM}} = \mathcal{L}_{\text{ce}} + \lambda_1 \mathcal{L}_{\text{leakage}} + \lambda_2 \mathcal{L}_{\text{magnitude}}. \tag{14b}$$

4. Experiments

4.1. Datasets

We conducted experiments on two multimodal datasets: OL3I (Opportunistic L3 computed tomography slices for Ischemic Heart Disease risk assessment) [29] and DVM (Data Visual Marketing) [30]. OL3I is a medical dataset that combines grayscale computed tomography (CT) images with high-dimensional clinical tabular features, while DVM is a large-scale natural dataset for automotive applications, offering a complementary benchmark for general-purpose scenarios.

The OL3I dataset contains 8,139 axial CT slices at the third lumbar vertebra (L3) level, each paired with structured medical record data. Every image is a 512×512 grayscale slice, and the tabular modality includes 423 features covering patient demographics, laboratory results, and aggregated clinical codes. In our study, we adopted the 1-year IHD prediction task as the downstream objective. Due to the severe class imbalance, we used the Area Under the receiver operating Characteristic Curve (AUC) as the evaluation metric. All experiments followed the dataset's predefined training, validation, and test splits.

The DVM dataset comprises 35,719 image-tabular pairs, each consisting of a 300×300 front-view RGB image of a car and corresponding tabular data on its sales and technical data. To avoid label leakage in the car model classification task, we excluded features such as brand, model year, and vehicle size, resulting in 13 tabular input features. Following the protocol in [6], we considered 101 target classes, each with at least 100 samples. Top-1 accuracy (ACC) was used as the evaluation metric, and the data was randomly split into training, validation, and test sets with a ratio of 72%, 18%, and 10%, respectively. Additional preprocessing details are provided in Appendix A.1.

4.2. Experimental Setup

In our experiments, the image encoder is a ResNet-50, which generates image features of dimension 2048. The tabular encoder is implemented as FT-Mamba, consisting of two Mamba blocks, and produces tabular features of dimension 64. All projection heads in the network are MLPs with a single hidden layer. During pretraining, the projection heads map features to a 128-dimensional space [6]. In the AMF module, the projection heads project features to a 2048-dimensional space. A fully connected layer is used for the final classification.

We first pretrain the image and tabular encoders on the OL3I and DVM datasets, respectively, following the contrastive learning strategy described in [6]. Prior to fine-tuning, each pretrained encoder is frozen and paired with an FC classification layer to form a unimodal model. The unimodal performance on the downstream task is then used to compute the modality confidence ratio r_{conf} according to (2). For downstream tasks, fine-tuning is performed under two strategies: in the frozen strategy, the parameters of pretrained encoders remain fixed, whereas in the trainable strategy, all parameters are updated during training. All results are reported as the mean and standard deviation across five different random seeds, and more implementation details are provided in Appendix A.

4.3. Comparison with Recent Methods on Downstream Classification

To evaluate the effectiveness of our proposed AMF-MedIT framework, we compare it with both supervised baselines and recent SSL approaches on the two downstream classification tasks. The supervised baselines include a ResNet-50 for the image modality [28] , an FT-Transformer for the tabular modality [25], and a Concat Fuse method for medical image and tabular fusion [33]. For SSL-based multimodal methods, we include MMCL [6], TIP [13], and AMF-MedIT. MMCL and AMF-MedIT follow the same pretraining strategy, while TIP adopts the pretraining procedure described in its original paper.

Table 1
Classification performance of AMF-MedIT and recent methods for image and tabular data.

Model	OL3I AUC (%)		DVM ACC (%)	
	❄[1]	🔥[2]	❄	🔥
Supervised Unimodal and Multimodal Methods				
ResNet-50 [28]	58.29 ± 4.16		93.00 ± 1.05	
FT-Transformer [25]	74.66 ± 1.25		94.30 ± 0.29	
Concat Fuse [33]	76.49 ± 0.61		90.37 ± 7.54	
SSL Multimodal Pre-training Methods				
MMCL [6]	76.31 ± 0.45	77.18 ± 0.82	93.51 ± 1.06	97.45 ± 0.19
TIP [13]	73.02 ± 0.40	75.87 ± 0.44	98.84 ± 0.05	97.91 ± 0.11
AMF-MedIT (ours)	76.97 ± 0.75	**78.36 ± 0.12**	94.76 ± 0.45	**98.89 ± 0.11**

[1] Frozen strategy.
[2] Trainable strategy.

The results are summarized in Table 1. For the OL3I task, AMF-MedIT achieves the best performance under the trainable strategy, with an AUC of 78.36%, outperforming both MMCL (77.18%) and TIP (75.87%). For the DVM task, which contains a substantially larger number of training samples, AMF-MedIT also achieves the highest accuracy of 98.89%, marginally higher than TIP (98.84%). While TIP can fully exploit its attention-intensive architecture and achieve competitive performance when ample data are available, its reliance on large datasets limits its generalization to smaller medical cohorts. In contrast, AMF-MedIT shows consistent advantages across both large- and small-scale settings, benefiting from its straightforward yet adaptive fusion design. Furthermore, compared with supervised unimodal baselines, the improvements are substantial, underscoring the advantages of multimodal SSL pretraining.

4.4. Robustness to Noisy Data

To evaluate the robustness of the models under noisy conditions, we simulated two types of noise commonly encountered in clinical practice based on the OL3I dataset: tabular data missingness and additive Gaussian noise in images.

- Tabular data missingness: We adopted the random feature missingness (RFM) strategy, where a random subset of feature columns is removed and the missing values are imputed with the mean of the corresponding feature. The missing rates were set to $[0, 0.25, 0.5, 0.75]$.
- Image Gaussian noise: We injected additive Gaussian noise into the images, with noise levels controlled by the standard deviation σ, set to $[0, 0.1, 0.15, 0.25]$.

We designed two experimental scenarios: (i) noisy tabular data with clean images, and (ii) noisy images with clean tabular data. In each scenario, the noise was applied to the training, validation, and test splits of the downstream tasks, and we fine-tuned the pretrained models of three image-tabular SSL frameworks (MMCL [6], TIP [13], and ours) on the noisy datasets and compared their downstream performance. Fig. 4 presents the multimodal performance on noisy data.

For tabular missingness, TIP achieves relatively low absolute AUC but shows the smallest performance drop, indicating strong stability. This advantage comes from its dedicated missing-feature processing mechanism, which masks absent tabular features and enables consistent adaptation under both frozen and trainable settings. Although AMF-MedIT demonstrates strong performance and good adaptability under the frozen setting, its performance degrades notably under the trainable setting when the missing rate exceeds 0.5. This limitation may arise because r_{conf} is configured based on the frozen-encoder performance in our experiments; at high missing rates, the confidence scores of frozen and trainable encoders diverge, leading to suboptimal adjustment. In future work, the settings of r_{conf} could be refined by leveraging unimodal performance under the trainable strategy to achieve a more accurate estimation, and even further adjusted by incorporating domain expertise from medical professionals.

For image noise, AMF-MedIT delivers the best overall performance with minimal degradation, demonstrating strong adaptability. TIP, however, is less effective in this case. Since its mechanism is tailored only for handling tabular missingness, it cannot accommodate degraded image quality, resulting in weaker robustness compared with AMF-MedIT and MMCL.

Overall, AMF-MedIT exhibits the strongest robustness across both noise scenarios. Rather than depending on a modality-specific mechanism, the r_{conf} adjustment in our model operates in a modality-agnostic manner, providing broader adaptability.

4.5. Ablation Studies

4.5.1. Ablation Studies on Tabular Encoder and Model Interpretation

To evaluate the effectiveness of the proposed FT-Mamba tabular encoder within AMF-MedIT, we replace it with three representative tabular encoders: MLP [6, 33], ResNet [25], and FT-Transformer [25]. Each encoder is trained under three settings: (i) unimodal supervised training, (ii) multimodal contrastive pretraining, and (iii) multimodal fine-tuning on downstream tasks. To further interpret how different tabular encoders operate and influence contrastive learning, we additionally perform Integrated Gradients (IG) analysis [31] on unimodal tabular models and Grad-CAM analysis [32] on multimodal encoders.

Performance. The results of tabular encoders on the OL3I and DVM datasets are summarized in Table 2. On the OL3I dataset, FT-Mamba achieves the best AUC both in unimodal (77.36%) and multimodal (78.36%) settings. On the DVM dataset, FT-Transformer achieves the best overall accuracy in both unimodal supervised training (94.30%) and multimodal fine-tuning (99.23%), with FT-Mamba ranking closely behind (92.71% and 98.89%, respectively). In contrast, MLP and ResNet performed much worse. Overall, these results suggest that FT-Mamba offers better adaptability across datasets of different scales: it shows clear advantages on the smaller OL3I dataset, while remaining competitive with FT-Transformer on the larger DVM dataset.

Integrated gradients analysis of tabular encoders. Fig. 5 shows the overall distribution of feature importance estimated by IG across all tabular features on the OL3I dataset, while Fig. 6(a) groups the 423 features into three clinical categories for clarity. More detailed IG results are provided in Appendix B. The results reveal that MLP and ResNet show weak feature discrimination: MLP distributes attention almost uniformly across features, while ResNet exhibits a single peak but still spreads over many irrelevant variables. Quantitatively, their most important features account for only 18% (ResNet) and even less for MLP of the total importance, far below FT-Mamba (45%) and FT-Transformer (36%) according to Table B.1. As a result, over 60% of their attention remains dispersed across low-importance features, reflecting a limited ability to extract clinically meaningful signals.

Table 2
Ablation studies on different tabular encoders.

Tabular Encoder	OL3I AUC (%)			DVM ACC (%)		
	Unimodal	Multimodal ❄	Multimodal 🔥	Unimodal	Multimodal ❄	Multimodal 🔥
MLP [6]	74.67 ± 2.06	76.63 ± 0.22	76.86 ± 0.35	85.50 ± 0.50	94.92 ± 0.43	97.88 ± 0.42
ResNet [25]	74.43 ± 1.08	76.50 ± 0.83	75.46 ± 0.92	85.19 ± 0.37	95.09 ± 0.54	97.28 ± 0.70
FT-Transformer [25]	74.66 ± 1.25	75.62 ± 0.49	76.46 ± 1.57	**94.30 ± 0.29**	93.83 ± 0.90	**99.23 ± 0.35**
FT-Mamba (ours)	**77.36 ± 1.58**	76.97 ± 0.75	**78.36 ± 0.12**	92.71 ± 0.28	94.76 ± 0.45	98.89 ± 0.11

In contrast, FT-Mamba and FT-Transformer concentrate on a small subset of features, with more than 60% of their attention massed on clinically salient variables. However, the patterns differ in quality. Fig. 6(b) and 6(c) show the specific high-importance features for FT-Mamba and FT-Transformer. FT-Mamba's high-importance features are strongly aligned with well-established cardiovascular risk factors for IHD, such as age, sex, Charlson Comorbidity Index (CCI), and records related to the cardiovascular system, which are consistent with the prior findings [4]. By comparison, FT-Transformer assigns substantial attention to features with little medical relevance (e.g., A12, A06, J05), suggesting it may capture noise patterns when training data are limited.

Grad-CAM analysis of multimodal encoders. Fig. 7 visualizes Grad-CAMs from different image–tabular encoder pairs, alongside a unimodal supervised ResNet-50. The supervised ResNet-50 shows no distinct focal regions, confirming its limited capacity to capture meaningful visual cues under scarce supervision. Models with MLP or ResNet tabular encoders produce narrow attention localized around the lumbar spine, reflecting weak and overly singular guidance to the image encoder. In contrast, FT-Mamba and FT-Transformer enable broader, multi-point attention patterns, consistent with their stronger long-range modeling ability. However, FT-Transformer often activates irrelevant or blank areas, suggesting its noisier tabular features misguide the image encoder during contrastive training. By comparison, FT-Mamba directs attention toward clinically relevant structures—including musculature, subcutaneous and visceral fat, and cardiovascular regions—among which visceral fat proportion is a key biomarker in predicting IHD [4].

Taken together, these results highlight the flexible and efficient capability of FT-Mamba in modeling tabular data. It can effectively extract highly relevant tabular features without being constrained by limited data or abundant irrelevant variables. Furthermore, it demonstrates superior capacity to provide medically meaningful supervision for the image modality. This simple yet effective working mechanism enables AMF-MedIT to achieve advantages in learning both tabular and image feature representations.

4.5.2. Ablation Study on Multimodal Fusion Module

To evaluate the effectiveness of the proposed AMF module, we replaced it with two mainstream multimodal fusion methods: concatenation and cross-attention. In the concatenation approach [6, 33], the unimodal feature vectors generated by the pretrained encoders are concatenated and fed into a fully connected layer for classification. For cross-attention [13], the feature map from the final ResNet-50 layer is flattened into image tokens, which are then combined with the tabular token generated by FT-Transformer or FT-Mamba and passed through a 2-layer multimodal cross-attention module with an embedding dimension of 512. The resulting CLS token is finally used for prediction. Table 3 summarizes both classification performance and computational costs. All measurements were conducted with a batch size of 1. The inference latency was measured on an NVIDIA RTX 4090 GPU, averaged over 100 runs.

On the OL3I dataset, AMF achieves the highest AUC (78.36%) under the trainable strategy, outperforming concatenation (76.49%) and cross-attention (73.48%). On the DVM dataset, AMF remains competitive under a trainable strategy (98.89%), demonstrating robust performance across datasets of different scales. These results confirm that simple fusion methods lack sufficient representational capacity, while highly complex methods like cross-attention are less robust under low-data regimes.

From an efficiency perspective, the cross-attention module incurs substantially higher parameters, FLOPs, and inference latency compared to AMF, reflecting its greater computational complexity. Notably, its cost grows quadratically with the spatial resolution of input images, due to token-level pairwise attention. In contrast, AMF performs interaction based on modality-level confidence scores rather than token-level pairwise computation, making

Table 3
Ablation studies on different multimodal fusion modules.

Fusion Module	#Param (M)	OL3I AUC (%)		OL3I (Cost)		DVM ACC (%)		DVM (Cost)	
		❄	🔥	FLOPs	Lat.	❄	🔥	FLOPs	Lat.
Concact [6]	-	75.80 ± 0.24	76.49 ± 0.60	-	-	89.08 ± 1.06	97.41 ± 0.36	-	-
Cross-attn. [25]	9.49	73.36 ± 0.30	73.48 ± 0.98	3.67	1.54	96.96 ± 0.52	98.15 ± 0.31	3.17	1.21
AMF (ours)	8.53	76.97 ± 0.75	**78.36 ± 0.12**	0.0085	0.12	94.76 ± 0.45	**98.89 ± 0.11**	0.0085	0.12

Note: FLOPs = floating-point operations (in GMac, 10^9 multiply–accumulate operations).
Lat. = latency (in ms).

its complexity effectively independent of image size. This design ensures that AMF is both scalable and practical, providing a strong balance between performance and efficiency for real-world multimodal applications.

4.5.3. Ablation Study on Loss Functions

To evaluate the contribution of the proposed auxiliary loss functions—magnitude loss and leakage loss, we fixed $r_{\text{conf}} = 0.95$ and compared four different loss configurations during fine-tuning: (1) $\mathcal{L}_{\text{fine-tuning}} = \mathcal{L}_{\text{ce}}$; (2) $\mathcal{L}_{\text{fine-tuning}} = \mathcal{L}_{\text{ce}} + \lambda_1 \mathcal{L}_{\text{leakage}}$; (3) $\mathcal{L}_{\text{fine-tuning}} = \mathcal{L}_{\text{ce}} + \lambda_2 \mathcal{L}_{\text{magnitude}}$; (4) $\mathcal{L}_{\text{fine-tuning}} = \mathcal{L}_{\text{ce}} + \lambda_1 \mathcal{L}_{\text{leakage}} + \lambda_2 \mathcal{L}_{\text{magnitude}}$ (ours).

Table 4
AUCs of AMF-MedIT with different loss configurations on the OL3I dataset.

$\mathcal{L}_{\text{ce}}$	$\mathcal{L}_{\text{leakage}}$	$\mathcal{L}_{\text{magnitude}}$	OL3I AUC (%)		DVM AUC (%)	
			❄	🔥	❄	🔥
✓	×	×	73.11 ± 3.99	78.07 ± 0.64	88.31 ± 1.07	97.30 ± 0.36
✓	✓	×	75.46 ± 1.71	78.05 ± 0.36	91.01 ± 0.55	96.79 ± 0.70
✓	×	✓	74.11 ± 3.49	78.08 ± 0.35	91.20 ± 0.57	97.08 ± 0.30
✓	✓	✓	**76.97 ± 0.75**	**78.36 ± 0.12**	**94.76 ± 0.45**	**98.89 ± 0.11**

The AUCs under different loss configurations are presented in Table 4. The best performance is achieved when all three losses are jointly applied. Removing the leakage loss $\mathcal{L}_{\text{leakage}}$ leads to a substantial drop in performance, especially under the frozen strategy (-2.86% AUC), indicating that the absence of information flow regulation across modalities causes significant semantic leakage and degraded fusion quality. Similarly, omitting the magnitude loss also reduces performance and introduces instability, as it fails to constrain the feature magnitude alignment between the image and tabular modalities, which can impede effective fusion.

The improvement brought by these two auxiliary losses is more evident in the frozen training setting, where the fusion module receives no gradient from the backbone. In such cases, relying solely on the classification loss $\mathcal{L}_{\text{ce}}$ is insufficient for guiding the training of the AMF module, leading to higher performance variance and lower overall accuracy. These results collectively validate the necessity of both magnitude and leakage losses for effective multimodal training, as they enhance representation compatibility, stabilize optimization, and ultimately improve classification performance.

5. Discussion

In this work, we proposed AMF-MedIT, an efficient framework for integrating medical image and tabular data. Our approach was motivated by the substantial heterogeneity between modalities, particularly in terms of dimensionality and confidence levels, as well as the challenges posed by small and noisy clinical datasets. To address these issues, we designed a twofold strategy: (1) introducing prior knowledge through the AMF fusion module to achieve robust and adaptive integration across modalities, and (2) leveraging self-supervised contrastive pretraining together with an efficient tabular encoder, FT-Mamba, to enhance multimodal representation learning under limited data.

Compared with existing methods, AMF-MedIT demonstrates consistent advantages in both small-scale medical datasets and large-scale natural datasets. In clinical practice, multimodal learning often faces the dual challenges of

data scarcity and noise. Medical image–tabular multimodal datasets remain limited in availability due to the high costs of data acquisition and the complexities of data management; moreover, missing clinical records or measurement errors can lead to unreliable tabular features, while artifacts and acquisition noise may compromise image quality. These challenges highlight that practical medical artificial intelligence (AI) systems require not only powerful models but also strategies that enhance efficiency and adaptability to limited data. Our experiments show that AMF-MedIT strikes an effective balance between performance and data efficiency, benefiting from its simple yet adaptive design. Specifically, instead of relying on cross-attention-based modules, we adopt a modulation-based fusion strategy that incorporates prior knowledge, thereby alleviating the training burden and avoiding the data inefficiency often caused by overly complex fusion modules. Furthermore, thanks to its modality-agnostic confidence adjustment mechanism, AMF-MedIT exhibits robust tolerance to noise in both image and tabular modalities. In addition, the introduction of feature masks and leakage loss enables adaptive adjustment to changing downstream data conditions without requiring structural modifications to the network, thereby providing broader clinical applicability compared with methods tailored to specific noise types.

Ablation studies further highlight the key contributions of both the FT-Mamba encoder and the AMF module. First, the tabular encoder plays a central role in maintaining both high accuracy and data efficiency. Interpretability analyses reveal that FT-Mamba captures fine-grained, discriminative tabular features more effectively than MLP and ResNet encoders, which are constrained by local attention. Moreover, by employing a selective mechanism, it achieves efficient feature extraction that remains robust under limited data. Compared with transformer-based models that rely on full attention, it is capable of filtering out irrelevant features while preserving medically meaningful signals, without overfitting to spurious noise patterns introduced by model complexity. This capability also enhances contrastive pretraining and guides the image encoder toward learning precise and clinically relevant attention patterns. Second, the AMF module achieves strong performance with remarkably low computational complexity, in contrast to concatenation- or attention-based approaches. Importantly, its complexity is independent of image resolution, which is advantageous for small datasets and satisfies the real-time requirements of clinical applications. Finally, loss function ablations confirm that leakage loss and density loss contribute to training stability and overall performance.

Despite the observed benefits, our study has several limitations. First, while FT-Mamba provides a good trade-off under limited data, it still underperforms FT-Transformer on large-scale datasets, suggesting that FT-Transformer may be preferable as larger medical datasets become available. Second, the configuration of the r_{conf} parameter in the AMF module remains sensitive. Our experiments on missing tabular features indicate that estimates derived from frozen strategies may not generalize well to trainable settings. More precise tuning strategies, potentially incorporating expert knowledge, are needed to improve robustness. Third, our validation was restricted to two multimodal datasets, and further evaluation on a wider range of clinical data is necessary. Lastly, our framework was applied only to classification tasks; its potential for other medical applications, such as segmentation or automated report generation, remains unexplored.

Looking forward, several directions can be pursued. Future work could focus on developing automated procedures to optimize r_{conf} based on dataset-specific characteristics, while integrating domain expertise to guide its estimation. Extending the evaluation to more diverse modalities, such as MRI and ultrasound, will further establish the generalizability of our framework. Moreover, expanding AMF-MedIT beyond classification to support segmentation and report generation tasks could broaden its impact on multimodal medical AI.

6. Conclusion

In conclusion, we proposed AMF-MedIT for integrating medical image and tabular data. At its core, the AMF module leverages streamlined modulation and prior knowledge to coordinate cross-modal disparities, providing an accurate and adaptable fusion strategy across diverse downstream conditions. Additionally, FT-Mamba serves as an efficient yet powerful tabular encoder, achieving a favorable balance between accuracy and data efficiency. Our interpretability analyses further reveal that encoder selection critically shapes multimodal contrastive pretraining, highlighting the essential role of FT-Mamba in multimodal feature representations. Beyond architectural contributions, AMF-MedIT improves robustness against noise and reduces reliance on large annotated datasets, offering a practical and scalable solution to the realities of clinical multimodal learning. We hope this framework will inspire advancements in medical multimodal models and their applications in healthcare.

CRediT Authorship Contribution Statement

Congjing Yu: Conceptualization, Data curation, Methodology, Validation, Writing - original draft, Writing - review & editing. **Jing Ye:** Validation, Writing - review & editing. **Yang Liu:** Validation, Writing - review & editing. **Xiaodong Zhang:** Supervision. **Zhiyong Zhang:** Funding acquisition, Supervision, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The work was supported by the Science and Technology Planning Project of Key Laboratory of Advanced IntelliSense Technology, Guangdong Science and Technology Department (2023B1212060024).

Appendix A Implementation Details

A.1 Data Preprocessing

For CT images in the OL3I dataset, the extreme grayscale values (greater than 275 and less than -125) were removed to enhance the image contrast. Then, the pixel values of images from both datasets were normalized to the range of 0 to 1.

For missing tabular data in the OL3I dataset, we applied median imputation. For tabular data in the DVM dataset, 13 features including 9 continuous features (Adv_year, Adv_month, Reg_year, Runned_Miles, Price, Seat_num, Door_num, Entry_price, Engine_size) and 4 categorical features (Color, Bodytype, Gearbox, Fuel_type) were remained for input. The continuous tabular features in both datasets were normalized using z-score normalization before being fed into the model.

A.2 Multimodal Pretraining

The pretraining procedures and augmentations followed those described in [6], where image augmentations were applied with a probability of 0.95 and tabular features were corrupted at a rate of 0.3. The optimizer used was Adam with a weight decay of $1e^{-4}$. The learning rate was selected based on validation performance, with possible values of $\{3e^{-3}, 3e^{-4}, 3e^{-4}\}$. The model was trained for 500 epochs using a cosine annealing scheduler with a warmup period of 10 epochs.

A.3 Downstream training

Considering the severe class imbalance in the OL3I dataset, where class 0 contains 7,784 samples and class 1 only 355 samples, the weight factor α in the weighted CE loss was set to 0.1 for class 0 and 2 for class 1. Additionally, a weighted sampler strategy was applied to the OL3I dataset, where class 0 in the training set was downsampled by a factor of 0.5, and class 1 was oversampled by a factor of 5.

During the fine-tuning of the AMF module, λ_1 and λ_2 were empirically set to 5.0 to ensure that the magnitude and leakage losses have a comparable magnitude and regularization strength to the cross-entropy loss.

To determine the embedding dimension D_{tab} for tabular encoders, we evaluated multiple candidate dimensions during supervised training on the OL3I dataset. For MLP and ResNet, the values tested were $\{128, 256, 512, 1024\}$, and for FT-Transformer and FT-Mamba, $\{32, 64, 96, 128, 256, 384\}$ were considered. Based on hyperparameter search, the tabular embedding dimension D_{tab} was set to 256 for MLP and ResNet, and 64 for FT-Transformer and FT-Mamba. The number of encoder blocks was fixed to 2 across all tabular models, following the configurations in [6, 25].

To optimize the training hyperparameters, we tested three learning rates $\{1e^{-2}, 1e^{-3}, 1e^{-4}\}$ using a fixed seed of 2022 for all supervised and fine-tuning experiments. In the fine-tuning of the cross-attention module, we observed severe overfitting and thus tested weight decay values of $\{1e^{-1}, 1e^{-2}, 1e^{-3}, 1e^{-4}\}$ to regularize training. In all other cases, the weight decay factor was fixed to 0. The optimal values were selected based on the validation performance.

All experiments utilized the Adam optimizer. Early stopping was applied based on the validation metric, triggered if no improvement greater than 0.0002 was observed over 10 consecutive epochs. After determining the optimal hyperparameters, each model was retrained using five random seeds (2022–2026) to mitigate randomness.

A.4 Quantifying the Modality Confidence Ratio

In our experiments, we assessed the confidence levels of unimodal feature vectors based on their respective unimodal performance. Specifically, after multimodal pretraining, we constructed unimodal models by utilizing the pretrained encoders and fine-tuned them on downstream tasks with a frozen strategy. The modality confidence ratio r_{conf} was then calculated as the ratio of the image-to-tabular unimodal model metrics on the testing set. The unimodal performance and the corresponding modality confidence ratio r_{conf} are presented in Table A.1 and A.2.

Table A.1
Unimodal encoder performance and corresponding modality confidence ratios with different tabular encoders.

Tabular Encoder	OL3I			DVM		
	Image AUC (%)	Tabular AUC (%)	r_{conf}	Image ACC (%)	Tabular ACC (%)	r_{conf}
MLP	74.99	77.41	0.97	66.11	79.65	0.83
ResNet	74.99	75.64	1.00	45.23	77.95	0.58
FT-Transformer	75.62	74.62	1.01	34.20	71.23	0.48
FT-Mamba	72.92	76.21	0.95	60.20	66.78	0.90

Table A.2
Unimodal performance of pretrained encoders and modality confidence ratios under noisy conditions.

Tabular Missing Rate	OL3I AUC (%)			Image σ	OL3I AUC (%)		
	Image	Tabular	r_{conf}		Image	Tabular	r_{conf}
0.25	72.92	76.14	0.96	0.10	69.37	76.21	0.91
0.50	72.92	66.93	1.09	0.15	64.16	76.21	0.96
0.75	72.92	64.07	1.14	0.25	56.89	76.21	0.89

A.5 Model interpretation

IG estimates the contribution of each input feature by integrating gradient values along a straight-line path from a baseline input to the target sample [31]. In our setting, the baseline input was defined as the mean of all training tabular samples. IG scores were computed for all training samples with respect to the disease class, and the final feature importance values were averaged to mitigate randomness. For consistent comparison, each feature's importance was normalized as a percentage of the total importance within the model, highlighting its relative influence during inference. We also applied Grad-CAM [32] to generate class activation maps (CAMs) from the pretrained multimodal models. For a consistent comparison, all models used a concatenation-based fusion framework to avoid confounding effects from fusion modules.

Appendix B Detailed Results of Integrated Gradient Analysis

Table B.1 presents the top 10 most important tabular features and their corresponding importance scores as identified by the IG analysis for the four unimodal tabular models. While MLP, ResNet, and FT-Transformer tend to focus on several variables that are less relevant to IHD, FT-Mamba demonstrates a strong emphasis on features closely associated with the cardiovascular system.

References

[1] S.-C. Huang, A. Pareek, S. Seyyedi, I. Banerjee, M. P. Lungren, Fusion of medical imaging and electronic health records using deep learning: A systematic review and implementation guidelines, NPJ DIGITAL MEDICINE 3 (1) (2020) 136. doi:10.1038/s41746-020-00341-z.

[2] J. N. Acosta, G. J. Falcone, P. Rajpurkar, E. J. Topol, Multimodal biomedical ai, NATURE MEDICINE 28 (9) (2022) 1773–1784. doi:10.1038/s41591-022-01981-2.

[3] S. Ö. Arik, T. Pfister, TabNet: Attentive interpretable tabular learning, Proceedings of the AAAI Conference on Artificial Intelligence 35 (8) (2021) 6679–6687. doi:10.1609/aaai.v35i8.16826.

Table B.1
Top 10 most important tabular features for MLP, ResNet, FT-Transformer and FT-Mamba unimodal tabular models.

Tabular Model	Feature	Description	Importance score
MLP	gender	Whether the patient is male	0.056
	age_at_scan	Age at time of abdominal scan in years	0.027
	Evaluation_14	Number of times a CPT code was documented in the group: Preventive Medicine Services	0.022
	V04	Number of prescriptions that match ATC code classification: Diagnostic Agents	0.021
	smoker	Whether it was documented that the patient is a smoker in the year prior to abdominal scan	0.021
	exp_wt_lab_diastolic	Exponentially weighted average of measurements diastolic blood pressure during the year prior to abdominal scan [mm Hg]	0.018
	S02	Number of prescriptions that match ATC code classification: Otologicals	0.018
	D10	Number of prescriptions that match ATC code classification: Anti-Acne Preparations	0.017
	B01	Number of prescriptions that match ATC code classification: Antithrombotic Agents	0.017
	A04	Number of prescriptions that match ATC code classification: Antiemetics And Antinauseants	0.016
ResNet	age_at_scan	Age at time of abdominal scan in years	0.180
	CCI	Charlson Comorbidity Index	0.034
	B01	Number of prescriptions that match ATC code classification: Antithrombotic Agents	0.021
	J02	Number of prescriptions that match ATC code classification: Antimycotics For Systemic Use	0.016
	Chapter_II_4	Number of times an ICD was documented in the year prior to imaging: Neoplasms of uncertain or unknown behaviour	0.014
	C09	Number of prescriptions that match ATC code classification: Agents Acting On The Renin-Angiotensin System	0.014
	Medical_9	Number of times a CPT code was documented in the group: Special Otorhinolaryngologic Services and Procedures	0.013
	Evaluation_14	Number of times a Current Procedural Terminology code was documented in the group: Preventive Medicine Services	0.013
	M02	Number of prescriptions that match ATC code classification: Topical Products For Joint And Muscular Pain	0.013
	exp_wt_lab_systolic	Exponentially weighted average of measurements of systolic blood pressure during the year prior to abdominal scan [mm Hg]	0.012
FT-Transformer	age_at_scan	Age at time of abdominal scan in years	0.362
	CCI	Charlson Comorbidity Index	0.322
	C08	Number of prescriptions that match ATC code classification: Calcium Channel Blockers	0.098
	A12	Number of prescriptions that match ATC code classification: Mineral Supplements	0.057
	J05	Number of prescriptions that match ATC code classification: Antivirals For Systemic Use	0.027
	gender	Whether the patient is male	0.026
	A06	Number of prescriptions that match ATC code classification: Drugs For Constipation	0.011
	Surgery_1	Number of times a CPT code was documented in the group: General Surgical Procedures	0.010
	C10	Number of prescriptions that match ATC code classification: Lipid Modifying Agents	0.010
	Evaluation_14	Number of times a CPT code was documented in the group: Preventive Medicine Services	0.010
FT-Mamba	age_at_scan	Age at time of abdominal scan in years	0.452
	CCI	Charlson Comorbidity Index	0.085
	gender	Whether the patient is male	0.063
	C03	Number of prescriptions that match ATC code classification: Diuretics	0.032
	Evaluation_14	Number of times a CPT code was documented in the group: Preventive Medicine Services	0.013
	C09	Number of prescriptions that match ATC code classification: Agents Acting On The Renin-Angiotensin System	0.012
	B05	Number of prescriptions that match ATC code classification: Blood Substitutes And Perfusion Solutions	0.011
	exp_wt_lab_systolic	Exponentially weighted average of measurements of systolic blood pressure during the year prior to abdominal scan [mm Hg]	0.010
	V06	Number of prescriptions that match ATC code classification: General Nutrients	0.009
	Chapter_IX_3	Number of times an ICD was documented in the year prior to imaging: Hypertensive diseases	0.008

[4] J. M. Zambrano Chaves, A. L. Wentland, A. D. Desai, I. Banerjee, G. Kaur, R. Correa, R. D. Boutin, D. J. Maron, F. Rodriguez, A. T. Sandhu, D. Rubin, A. S. Chaudhari, B. N. Patel, Opportunistic assessment of ischemic heart disease risk using abdominopelvic computed tomography and medical record data: A multimodal explainable artificial intelligence approach, Scientific Reports 13 (1) (2023) 21034. doi:10.1038/s41598-023-47895-y.

[5] V. Borisov, T. Leemann, K. Seßler, J. Haug, M. Pawelczyk, G. Kasneci, Deep neural networks and tabular data: A survey, IEEE Transactions on Neural Networks and Learning Systems 35 (6) (2024) 7499–7519. doi:10.1109/TNNLS.2022.3229161.

[6] P. Hager, M. J. Menten, D. Rueckert, Best of both worlds: Multimodal contrastive learning with tabular and imaging data, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23924–23935.

[7] X. Miao, Y. Wu, L. Chen, Y. Gao, J. Yin, An experimental survey of missing data imputation algorithms, IEEE Transactions on Knowledge and Data Engineering 35 (7) (2023) 6630–6650. doi:10.1109/TKDE.2022.3186498.

[8] A. Gu, T. Dao, Mamba: Linear-time sequence modeling with selective state spaces (Dec. 2023). arXiv:2312.00752, doi:10.48550/arXiv.2312.00752.

[9] Q. Chen, Y. Hong, Alifuse: Aligning and fusing multi-modal medical data for computer-aided diagnosis (Jan. 2024).

[10] R. Krishnan, P. Rajpurkar, E. J. Topol, Self-supervised learning in medicine and healthcare, NATURE BIOMEDICAL ENGINEERING 6 (12) (2022) 1346–1352. doi:10.1038/s41551-022-00914-1.

[11] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision (Feb. 2021). arXiv:2103.00020, doi:10.48550/arXiv.2103.00020.

[12] J. Li, R. R. Selvaraju, A. D. Gotmare, S. Joty, C. Xiong, S. Hoi, Align before fuse: Vision and language representation learning with momentum distillation (Oct. 2021). arXiv:2107.07651, doi:10.48550/arXiv.2107.07651.

[13] S. Du, S. Zheng, Y. Wang, W. Bai, D. P. O'Regan, C. Qin, TIP: Tabular-image pre-training for multimodal classification with incomplete data (Jul. 2024). arXiv:2407.07582, doi:10.48550/arXiv.2407.07582.

[14] W. Huang, Multimodal contrastive learning and tabular attention for automated alzheimer's disease prediction (Aug. 2023).

[15] K. Zhang, X. Liu, J. Xu, J. Yuan, W. Cai, T. Chen, K. Wang, Y. Gao, S. Nie, X. Xu, X. Qin, Y. Su, W. Xu, A. Olvera, K. Xue, Z. Li, M. Zhang, X. Zeng, C. L. Zhang, O. Li, E. E. Zhang, J. Zhu, Y. Xu, D. Kermany, K. Zhou, Y. Pan, S. Li, I. F. Lai, Y. Chi, C. Wang, M. Pei, G. Zang, Q. Zhang, J. Lau, D. Lam, X. Zou, A. Wumaier, J. Wang, Y. Shen, F. F. Hou, P. Zhang, T. Xu, Y. Zhou, G. Wang, Deep-learning models for the detection and incidence prediction of chronic kidney disease and type 2 diabetes from retinal fundus images, Nature Biomedical Engineering 5 (6) (2021) 533–545. doi:10.1038/s41551-021-00745-6.

[16] Y. C. Lee, J. Cha, I. Shim, W.-Y. Park, S. W. Kang, D. H. Lim, H.-H. Won, Multimodal deep learning of fundus abnormalities and traditional risk factors for cardiovascular risk prediction, npj Digital Medicine 6 (1) (2023) 1–10. doi:10.1038/s41746-023-00748-4.

[17] L. A. Vale-Silva, K. Rohr, Long-term cancer survival prediction using multimodal deep learning, Scientific Reports 11 (1) (2021) 13505. doi:10.1038/s41598-021-92799-4.
[18] T. Zheng, S. Sone, Y. Ushiku, Y. Oba, J. Ma, TNF: Tri-branch neural fusion for multimodal medical data classification (Mar. 2024).
[19] H. Zheng, Z. Lin, Q. Zhou, X. Peng, J. Xiao, C. Zu, Z. Jiao, Y. Wang, Multi-transSP: Multimodal transformer for survival prediction of nasopharyngeal carcinoma patients, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2022, Springer Nature Switzerland, Cham, 2022, pp. 234–243. doi:10.1007/978-3-031-16449-1_23.
[20] X. Zheng, Z. Yao, Y. Huang, Y. Yu, Y. Wang, Y. Liu, R. Mao, F. Li, Y. Xiao, Y. Wang, Y. Hu, J. Yu, J. Zhou, Deep learning radiomics can predict axillary lymph node status in early-stage breast cancer, Nature Communications 11 (1) (2020) 1236. doi:10.1038/s41467-020-15027-z.
[21] N. Jia, T. Jia, L. Zhao, B. Ma, Z. Zhu, Multi-modal global- and local- feature interaction with attention-based mechanism for diagnosis of alzheimer's disease, Biomedical Signal Processing and Control 95 (2024) 106404. doi:10.1016/j.bspc.2024.106404.
[22] N. Painchaud, P.-Y. Courand, P.-M. Jodoin, N. Duchateau, O. Bernard, Fusing echocardiography images and medical records for continuous patient stratification (Jan. 2024).
[23] H.-Y. Zhou, Y. Yu, C. Wang, S. Zhang, Y. Gao, J. Pan, J. Shao, G. Lu, K. Zhang, W. Li, A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics, Nature Biomedical Engineering 7 (6) (2023) 743–755. doi:10.1038/s41551-023-01045-x.
[24] Y. Chen, S. Zhu, Z. Fang, C. Liu, B. Zou, Y. Wang, S. Chang, F. Jia, F. Qin, J. Fan, Y. Peng, C. Wang, Toward robust early detection of alzheimer's disease via an integrated multimodal learning approach (Aug. 2024).
[25] Y. Gorishniy, I. Rubachev, V. Khrulkov, A. Babenko, Revisiting deep learning models for tabular data, in: Proceedings of the 35th International Conference on Neural Information Processing Systems, NIPS '21, Curran Associates Inc., Red Hook, NY, USA, 2021.
[26] A. F. Thielmann, M. Kumar, C. Weisser, A. Reuter, B. Säfken, S. Samiee, Mambular: A sequential model for tabular deep learning (Aug. 2024).
[27] S. Pölsterl, T. N. Wolf, C. Wachinger, Combining 3D image and tabular data via the dynamic affine feature map transform, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2021, Springer International Publishing, Cham, 2021, pp. 688–698. doi:10.1007/978-3-030-87240-3_66.
[28] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778. doi:10.1109/CVPR.2016.90.
[29] S. AIMI, OL3I (opportunistic l3 computed tomography slices for ischemic heart disease risk assessment) (2023).
URL https://stanfordaimi.azurewebsites.net/datasets/3263e34a-252e-460f-8f63-d585a9bfecfc
[30] J. Huang, B. Chen, L. Luo, S. Yue, I. Ounis, DVM-CAR: A large-scale automotive dataset for visual marketing research and applications, CoRR abs/2109.00881 (2021). arXiv:2109.00881.
URL https://arxiv.org/abs/2109.00881
[31] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML'17, JMLR.org, 2017, p. 3319–3328.
[32] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual explanations from deep networks via gradient-based localization, in: 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 618–626. doi:10.1109/ICCV.2017.74.
[33] S. Spasov, L. Passamonti, A. Duggento, P. Liò, N. Toschi, A parameter-efficient deep learning approach to predict conversion from mild cognitive impairment to alzheimer's disease, NeuroImage 189 (2019) 276–287. doi:10.1016/j.neuroimage.2019.01.031.
[34] X. Jiang, W. Qian, L. Li, Quantitative CT analysis of different thoracolumbar vertebral bone densities in middle-aged and elderly populations, Chinese Journal of CT and MRI 20 (05) (2022) 179–182, in Chinese.

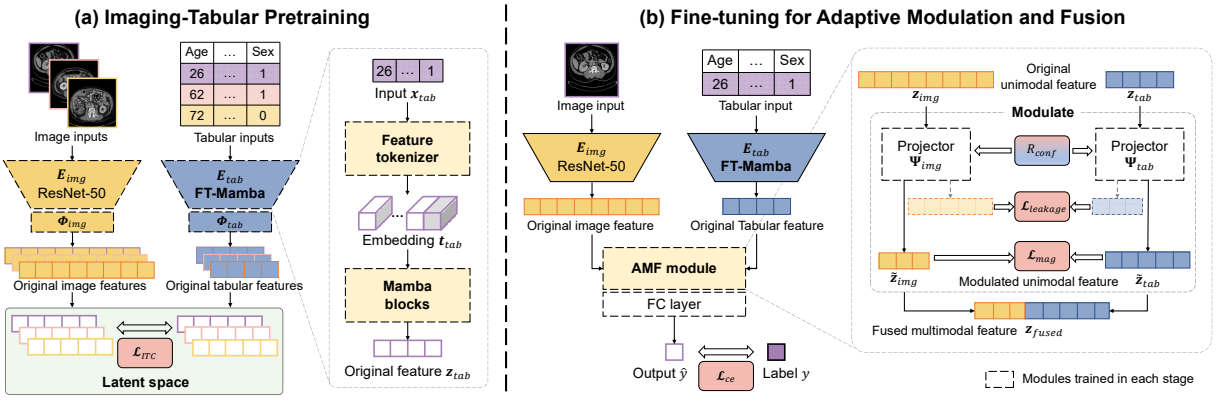


Figure 1: The overall framework of AMF-MedIT. (a) During the contrastive pretraining stage, unimodal encoders are trained without annotations to align image and tabular representations. We introduce FT-Mamba, a powerful tabular encoder for high-dimensional, noisy medical features. (b) In the modulation and fusion stage, we propose the AMF module, a novel modulation-based paradigm that harmonizes modality feature disparities and adaptively balances their contributions via prior-guided confidence ratio r_{conf}, along with the magnitude and leakage loss.

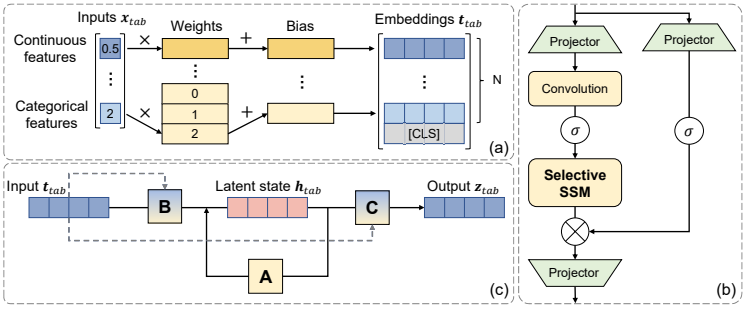


Figure 2: The key modules in FT-Mamba. (a) Feature tokenizer: categorical and numerical tabular features are separately tokenized into embeddings; (b) a Mamba block, and its core module (c) Selective SSM: it enables input-dependent parameterization of SSMs and dynamically adjusts them to focus on relevant input features.

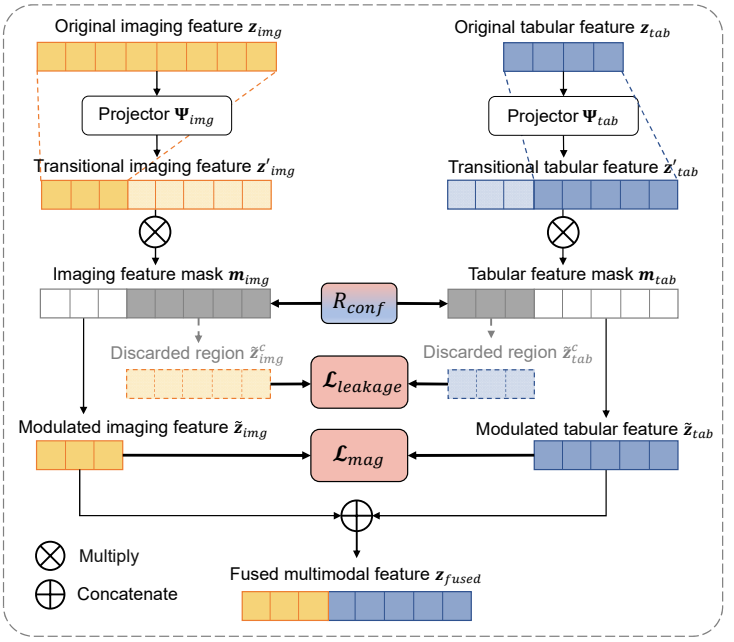
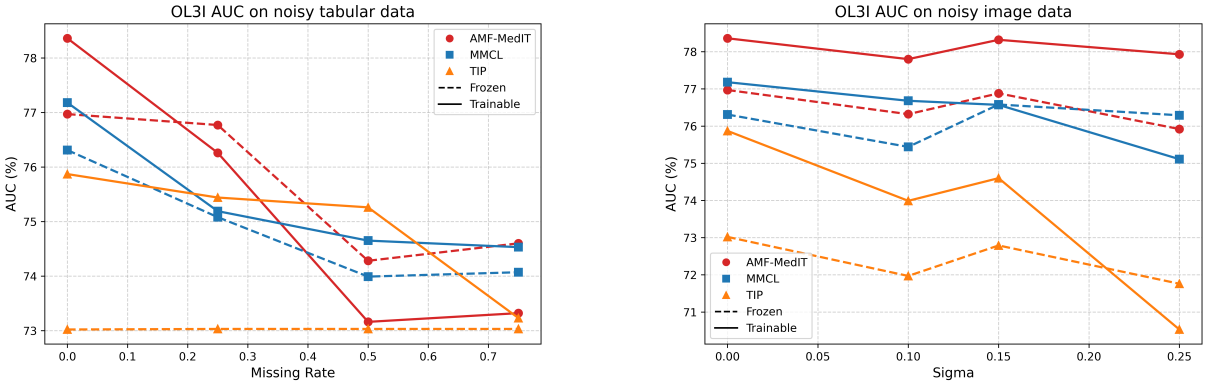


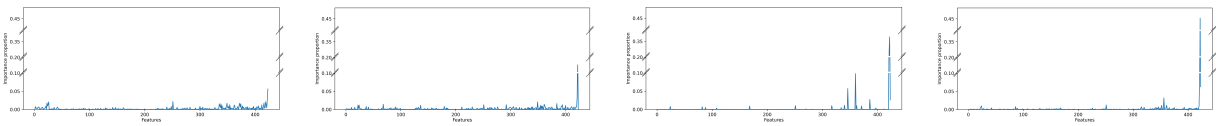
Figure 3: The structure of the AMF module. For each modality, the original features are mapped to the transitional features separately. Then feature masks implicitly partition them into retained region (solid lines) and discarded region (dashed lines). During training, the leakage loss compresses effective information into the retained region, and magnitude loss aligns their magnitude across modalities.



(a) Performance on missing tabular data and clean image data

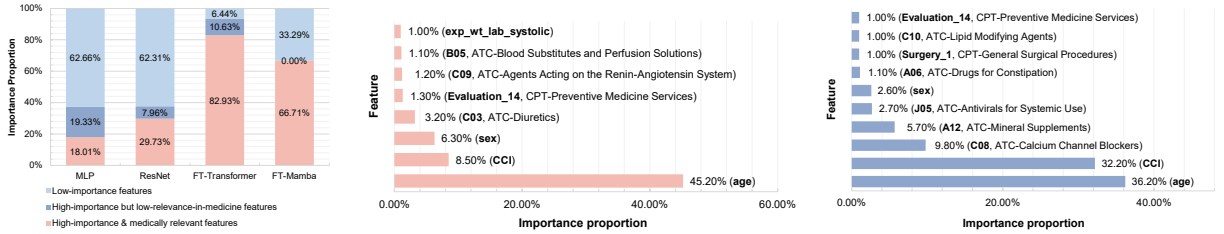
(b) Performance on noisy image data and clean tabular data

Figure 4: Performance comparison with SSL image-tabular methods on noisy data.



(a) MLP (b) ResNet (c) FT-Transformer (d) FT-Mamba

Figure 5: Importance distribution over 423 input features for unimodal tabular models.



(a) Importance proportions by category (b) High-importance features of FT-Mamba (c) High-importance features of FT-Transformer

Figure 6: IG analysis results for unimodal tabular models. (a): Feature importance proportions by category for the OL3I dataset. Features with an importance percentage less than 1% were classified as low-importance features (light blue). Features with an importance percentage of 1% or greater were classified as high-importance features. Based on [4] and established medical knowledge, we further divided the high-importance features into medically relevant features (pink) and low-relevance features (dark blue). (b) and (c): High-importance features and their importance proportion for FT-Mamba and FT-Transformer. (exp_wt_lab_systolic: exponentially weighted average of systolic blood pressure. CPT: Current Procedural Terminology codes, ATC: Anatomical Therapeutic Chemical codes.)

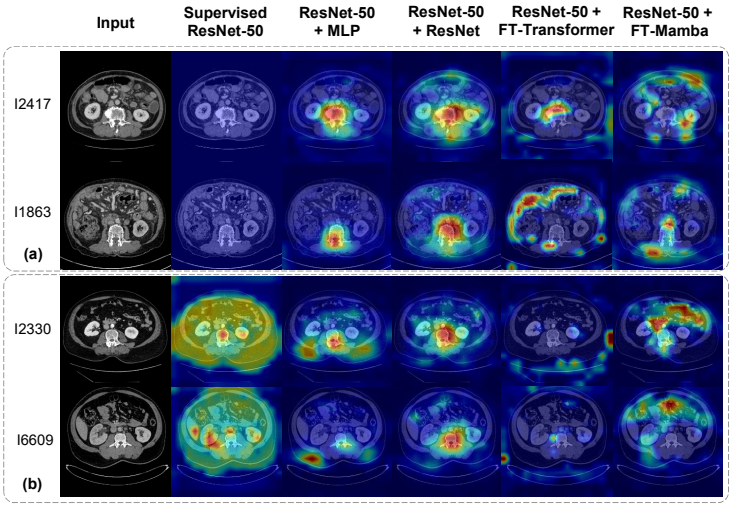


Figure 7: CAMs generated by supervised ResNet-50 and pretrained multimodal models for IHD prediction. Group (a) illustrates activation patterns in non-IHD cases, while group (b) are IHD-positive samples.